DOCTORAL THESIS

An initial validity argument for a new B2 CEFR-related baccalaureate listening test.

PRESENTED BY CAROLINE SHACKLETON

Universidad de Granada

DIRECTED BY DR. TONY HARRIS AND DR. JESUS GARCIA LABORDA

Programme: Lenguas, Textos y Contextos

June 2018

An initial validity argument for a new B2 CEFR-related baccalaureate listening test.

Caroline Shackleton

Editor: Universidad de Granada. Tesis Doctorales
Autor: Caroline Shackleton
ISBN: 978-84-9163-921-3
URI: http://hdl.handle.net/10481/52426

ACKNOWLEDGEMENTS

I would like to express my thanks to all the people who contributed to this study. Firstly, to my supervisors Tony Harris and Jesús García Laborda for their helpful advice, comments, suggestions and ongoing encouragement. To Tony and his colleagues in the department of Filologías Inglesa y Alemana for so generously allowing me access to their classes and especially to those students in the department who took part in the study.

I would also like to thank my colleagues and the students from the Centro de Lenguas Modernas who contributed to this study, and without whom the completion of this research project would not have been possible.

And last but not least a special thanks to Nathan Turner who not only contributed to this work but helped in ways too great to describe.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abbreviations

ACLES	= Asociación de centros de lenguas en la enseñanza superior
ALTE	= Association of Language Testers in Europe
AUA	= Assessment Use Argument
BFE	= Baccalaureate Final Exam
BOE	= Boletín Oficial Del Estado
CEFR	= Common European Framework of Reference
CLA	= Communicative Language Ability
CLIL	= Content and Language Integrated Learning
CLM	= Centro de Lenguas Modernas (Modern Language Center)
CoE	= Council of Europe
CTT	= Classical test theory
DA	= Dynamic assessment
DI	= Discrimination index
DR	= Discourse representation
EALTA	= European Association for Language Testing and Assessment
EAP	= English for academic purposes
EC	= European commission
ECD	= Evidence centred design
EHEA	= European Higher Education Area
ELF	= English as a lingua franca
EMI	= English as a medium of instruction
EOI	= Escuela Oficial de Idiomas
EU	= European Union
FL	= Foreign language
FV	= Facility value
G	= Gist
GEPT	= The General English Proficiency Test
IA	= Interpretative Argument
ICC	= Item characteristic curve
IELTS	= International English Language Testing System
FCE	= First Certificate in English exam
IPM	= Listening to infer (propositional) meaning
IRT	= Item response theory
IU	= Idea unit
KSAs	= Knowledge, skills and abilities
L	= Lexical recognition
L1	= First language
L2	= Second language
LOE	= Ley Orgánica de Educación
LOMCE	= Ley Orgánica para la Mejora de la Calidad Educativa
MALQ	= Metacognitive Awareness Listening Questionnaire
MCQ	= Multiple choice question
MECD	= Ministerio de Educación, Cultura y Deporte
MFRM	= Many-Facet Rasch measurement model
MISD	= Main ideas with supporting details
MM	= Multiple match question

MR = Meaning representation
MTT = Modern test theory
NF = Note form question
OECD = Organisation for Economic Co-operation and Development
OIB = Ordered item booklet
PAU = Prueba de Acceso a la Universidad
PCAR = Principle Components factor analysis of Rasch Residuals
PET = Prelimary English Test
PISA = Programme for International Student Assessment
PTE(A) = Pearson test of English (Academic)
QUAL = Qualitative
QUAN = Quantitative
RP = Response probability
SD = Standard deviation
SE = Standard error
SEM = Standard error of measurement
SIID = Specific information and Important details
SLA = Second language acquisition
TEFL = Teaching English as a Foreign Language
TOEFL (IBT) = Test of English as a Foreign Language (Internet-based test)
TOEIC = The Test of English for International Communication
TLU = Target language use domain
UGR = University of Granada
VA = Validity Argument

"A language test without validation research is like a police force without a court system, unfair and dangerous" (McNamara 2007, p.280).

Chapter 1. Introduction

It seems almost a commonplace nowadays to refer to the need for improved second-language acquisition skills in secondary school education systems in order to prepare students for the communication demands of an ever more rapidly globalised world (Rost, 2014). In a multi-cultural, multi-lingual environment such as the European Union, the potential difficulties for trade and cooperation have become increasingly apparent with the continued development of closer ties between its member states. In response to the communication challenges facing countries within its borders, the European Union provides the Common European Framework of Reference for languages (CEFR, Council of Europe, 2001) in an attempt to outline and standardise learners' communicative needs as an aid to language education policy (North, 2014).

Since its publication, the CEFR (CoE, 2001) has become the general framework for teaching and learning modern languages in Europe. Its aim is to promote transparency and coherence, and provide a common basis for language learning curricula, so that countries can be made comparable through the implementation of a shared conceptual framework. European education systems are invited to implement the CEFR in order to promote plurilingualism and allow for the mutual recognition of national exams (CoE, 2008). Despite claims that "the impact of the CEFR on testing far outweighs its impact on curriculum design and pedagogy" (Little, 2007, p. 648), the six-level scales (A1– C2), containing 'can-do' descriptors, have indeed found their way into many national language curricula (Moe, 2009, p. 131).

Within the EU, governments and education departments have been obliged to take the CEFR into account, and consequently new educational initiatives have been plentiful as policy makers attempt to incorporate competence-based language education into state systems (Lim, 2013). One of the main goals of the CEFR (CoE, 2001, p.1) is to provide a comprehensive description of "what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively". Consequently, assessment of these goals is a necessary component of the framework, and many European countries have been reforming their school-leaving examinations to reflect best practices in educational assessment. Such changes might be said to be particularly pressing in the teaching and assessment of English as a second language, due to its status as one of the main lingua francas within the EU, together with its spread as the current dominant language of international communication (Crystal, 2012). Indeed, this lingua franca status has led to the implementation of different national and international policies in order to improve L2 students' proficiency in the language.

In Spain, however, we are only just starting to see the influence of the CEFR. While improvement in quality of language learning is clearly now a major objective, results of the European Survey of Language Competence (European commission, 2012) which

measured competencies of obligatory secondary school pupils show Spanish language users of English as a second language to be lagging behind their European counterparts in terms of proficiency. It would therefore seem that efforts to improve language education are not at present having the desired effect. Indeed, this has been recognised in the recently revamped Spanish national curriculum *Ley Orgánica para la Mejora de la Calidad Educativa* (LOMCE) made law by the Ministry of Education, Culture & Sports (*Boletín Oficial Del Estado (BOE),* 2013); with a view to improving learning outcomes, its new syllabus for English has been elaborated using a competence-based curriculum. This curriculum, however, has met with resistance. There seems little interest on the part of the different national and regional governments in improving on the various school leaving tests currently in use—even though these old fashioned tests provide hardly any information about students' competences (Amengual Pizarro, 2005, 2006; García Laborda, 2010, 2012; Sanz Sainz, & Fernández Álvarez, 2005), and despite the fact that their replacement is a current requirement of Spanish law. Its future is thus somewhat uncertain. A clear sign of this lack of resolve is that while at present the LOMCE (further adapted by *BOE*s 2014, 2016) lays out the new proposals for educational reform in the Spanish baccalaureate teaching and assessment, the originally mandatory inclusion of both oral comprehension and production was postponed until 2018 and at present there are no visible indications of such changes to the test. Clearly further discussions are needed if progress is to be made in this area.

Nevertheless, there clearly exists widespread recognition throughout the higher education system that language ability, especially in English, is a key competence for the promotion of international mobility and the increase of job prospects for students (Dearden, 2015). The introduction of the European Higher Education Arena has meant that universities are necessarily following plurilingualism policies and numerous initiatives are now being implemented on degree courses in an attempt to help students develop foreign language competencies. In particular, there has been an increase in the offer of English as a medium of instruction (EMI) courses, as well as an increase in the provision of opportunities for participation in international programmes.

In addition to providing students with life-long transversal skills, those universities who do contribute to improving their students' language learning notice the welcome additional advantage that courses delivered in English also become more attractive to international students (Dearden, 2015; O'Dowd, 2015). Indeed, internationalisation has become a major policy objective in higher education (see committee of University rectors, CRUE, 2016) and there is currently a move towards more and more degree courses being taught in English. In 2014, the Spanish Ministry of Culture, Education and Sport published its 'Estrategia para la Internacionalización de las Universidades Españolas 2015-2020' (Strategy for the Internationalisation of Spanish Universities) and recommended that Spanish universities should be aiming to increase the number of degree and masters programmes taught in English or other languages. Such policies are also seen in obligatory state education, mainly in the form of Content and Language Integrated Learning (CLIL) in bilingual programs. In Andalusia, for example, CLIL has been introduced school-wide in all stages of compulsory education, particularly at primary level (Junta de Andalucia, 2005). However, in order to follow these new courses, students obviously need a certain level of language proficiency and this move towards subjects being taught in English is something that cannot be ignored by admission policies. Universities will typically have policies which admit students with sufficient language ability to participate successfully on their courses (Green, 2017, p.2). In fact, many degree courses already stipulate the required English language accreditation necessary in order to be considered for access. For example, in Navarra a CEFR B2 is necessary to follow degree programs in Primary teaching, an understandable requisite given the bi-lingual programs currently being implemented in schools which will require future teachers to impart the curriculum in English.

Nevertheless, despite such demands, students still do not leave upper secondary education with a CEFR-related qualification. Instead, they must turn to costly international exam providers in order to gain the necessary accreditation. Herein lies a glaring contradiction, for surely it is only fair practice that the state educational system provide students with the means and opportunity to achieve the educational requisites it demands of them? Indeed, it could be argued that rather than enabling second language

skills and mobility, the current hurried implementation of CEFR requirements without the necessary training at secondary school level only serves to further undermine students' educational possibilities. Seen in this light, the above-mentioned rejection by local authorities and teachers of the Spanish government's top-down implementation of CEFR requirements becomes all too apparent.

The lack of a listening component on the present *selectividad* test is particularly worrying given the importance of oral comprehension both as an essential component of communicative competence and as a contributory factor in successful language acquisition (Rubin, 1994; Vandergrift, 1999; Zhang, 2012). By including listening activities in the language learning classroom, comprehensible input is increased and, given that adults spend almost 50 per cent of their communication time listening (Miller, 2003), this important skill cannot be ignored. Good listening proficiency has been associated with academic success (Jeon, 2007); activities such as listening to teachers' explanations and classmates' questions are core academic activities. If a final secondary school qualification is to be used for university entrance, this is of obvious importance—especially for those courses which use EMI. Students will be required both to take lecture notes and to take part in question and answer sessions and they therefore need to be adequately prepared for success at these tasks. Indeed, lectures are probably the most important language event as regards the learning of subject matter content while at university (Lynch, 2011). Fulcher (1999) found that variance in EAP test scores was mostly due to language proficiency, and not specific subject-related knowledge. Listening is therefore vital in an academic setting, as well as for a student's later professional life (Vandergrift, 2007). General listening ability needs to be encouraged to equip students for success on the new courses now being initiated in Spanish universities. However, students are still not currently getting the necessary training to improve their listening skills in schools and as such are leaving the school state system inadequately prepared to be good listeners.

Due to the importance of the skill of listening for students' future performance, this thesis will argue that a listening component should be included in the final baccalaureate

test for school leavers, and furthermore that such a test should be a valid and reliable CEFR-related measure of students' proficiency levels. Tests in educational systems are fundamental for the establishment of a fair decision-making process, given that both funds and places on university courses are limited and that such scarce resources must be allocated fairly if we are to promote equal opportunities in a public university system. To this end, this thesis will both develop a prototype test and provide validity evidence to support the interpretation and use of scores in a school-leaving/university-entrance context.

Having introduced the motivation for the present study, I will now outline the structure of this thesis, which is organised into a further seven chapters. Chapter two provides a detailed statement of the current context and problem, and offers motivating reasons for the necessary development of a new B2 CEFR-related listening exam. Chapter three reviews the literature relevant to the construct of listening and validity theory as it relates to the development of such an exam. Chapter four provides a description of the test development process, detailing pre-pilot results, elaborating a framework for an evaluation of the test, and presenting my final research questions. Chapter five focuses on a discussion of the various methodologies used in the study. The results for each research question are then reported in chapter six, followed by a summary of these results in the form of the final validity argument in chapter seven. Finally, conclusions and future concerns are discussed in chapter eight.

Chapter 2. Statement of the problem

The last few years have witnessed substantial reform to the Spanish education system as a result of the LOMCE (BOE, 2013), which has attempted to bring Spain into line with other European education systems by providing a competence-based curriculum. One of the stated aims of this reform laid out in the BOE (2015) is to improve foreign language learning and proficiency in Spain, which would include an external assessment in English (the obligatory first foreign language (FL) in schools) as part of the university entrance exam provided by the Department of Education. This external assessment was scheduled to be introduced in 2017/18 and include a new oral (both receptive and productive, i.e., listening and speaking) component to the exam. However, no such reform has taken place and, moreover, the lack of clarity and direction by the MECD to describe either the form or content of the exam explains to a great extent the reluctance of regional governments to implement the reform and the consequent postponement of new assessment procedures for school leaving/university entrance (the Baccalaureate Final Exam - henceforth BFE).

It should be understood that any such assessment would form part of a significant decision for university entrance by 2018, and should consequently follow international guidelines for such high-stakes exams. And despite efforts by projects such as PAULEX and OPENPAU,[1] which focus mainly on speaking, no project seems to be in place at present to propose and develop the listening component of the new external assessment. In addition, previous attempts to introduce an oral section to the exam have also been repeatedly postponed. Initially the introduction of an oral section was programmed to come into force in 2012 (BOE, 2008), however this proposal was not put into practice. Then, Royal Decree 961/2012 of 22 June (BOE, 2012) outlined the introduction of an oral component by June 2014, which did not materialise either. Current guidelines for the reform of teaching and assessment for the BFE are derived from the LOMCE (2013) and reflected in the BOE (2015). This law stipulated that the inclusion of both speaking and listening would be postponed until 2018, yet it is clear that this proposal has still not materialised. In Andalusia, for example, the orientation given for 2017/18 by the regional government for the English section of the BFE is exactly the same as that used for the last 20 years and does not include either a speaking or a listening section.[2] Moreover, the apparent lack of direction, organisation and the expert knowledge required to instigate the necessary changes, in the the design and planning of the exam, appears to signal further delays.

In the following section, I look at language proficiency (specifically English language proficiency) in a European context and attempt to situate Spanish language-learners within it. I then examine the current Spanish university entrance exam, or PAU, as well as some of its main criticisms before highlighting the response to these criticisms as expressed in educational law. Finally, I evaluate recent proposals for a new BFE, and the extent to which the CEFR might be incorporated into such a high-stakes exam and which might then be used as a template for the listening component of the new BFE.

---

1 See http://www3.uah.es/proyecto_openpau/

2 See https-//www.juntadeandalucia.es/economiayconocimiento/sguit/examanes_anios_anteriores/selectividad/sel_Orientaciones_ingles.pdf

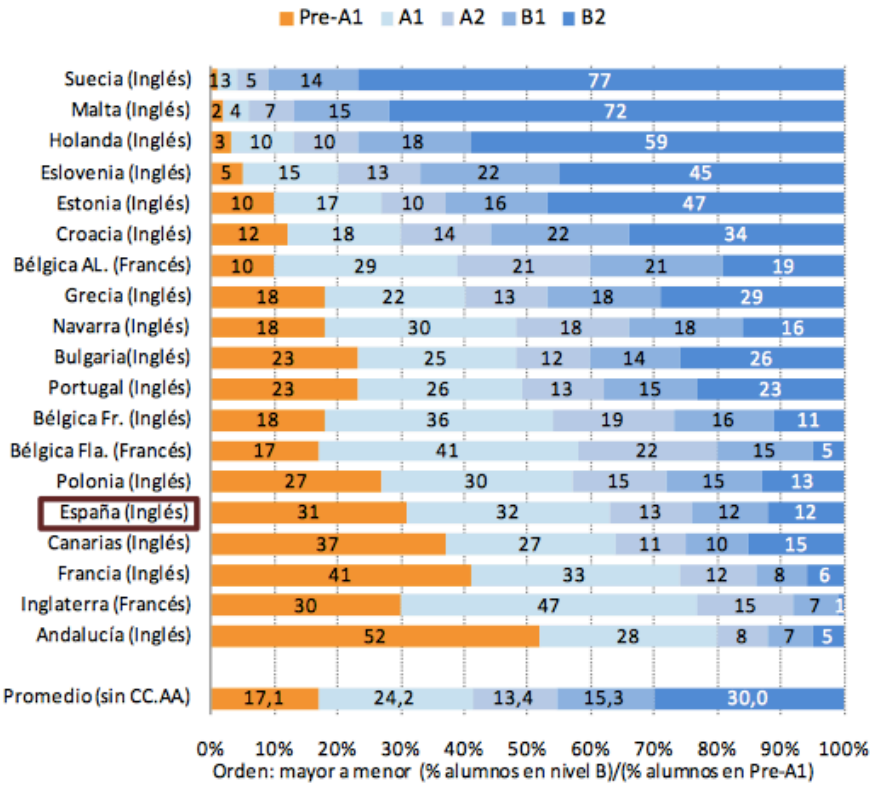
2.1 Language proficiency: Spain within a European context

In a globalised, mobile, free moving Europe, proficiency in a foreign language is of paramount importance (Rost, 2014). In Spain this need affects both internal economic growth and the educational and employment possibilities of its citizens. For example, the European Commission (2006) underscored the fact that Spain is one of the European Union (EU) countries which is least able to take advantage of export possibilities due to a lack of language proficiency. In 2013 it was reported that 39 million 16-29 year olds were either out of work or in education or training in Europe, and numbers were particularly high in Spain (OECD report, 2015). Furthermore, the link between academic achievement and economic growth (Hanushek, Ruhose & Woessmann, 2015), coupled with the idea that mastery of English (widely recognised as the global *lingua franca)* allows fluent speakers to communicate more freely in a globalised world, and consequently take up potential professional and employment possibilities, has not gone unnoticed by educational policy makers. Indeed, foreign language competence is the prime mover behind the EC's proposed benchmark objectives for 2020 for its member states, stipulating that at least 50% of students should have a CEFR B1 level or above in a foreign language by the time they are 15 years old (i.e., before they enter higher secondary school). Such proposals lead us to infer that a reasonable proficiency level for university entrance would be CEFR B2, a level beyond the current one. Moreover, as García Laborda and Martìn Monje (2013) note, it is debatable whether a CEFR B1 competence level has any real potential either in terms of employability or for pursuing studies in higher education (on this note, see also Green, 2008). The ideal scenario would, thus, appear to be a cohort of students leaving upper secondary school with at least a B2 level, as is the case in most other European countries (see for example, Deygers & Zeidler, 2015; Deygers, Zeidler, Vilcu, & Hamnes Carlsen , 2017; Lim, 2013).

Information collected on European secondary school students' foreign language proficiency in the European Survey of Language Competence has shown Spanish students to be lagging behind their European counterparts in this regard (European Commission, 2012). The Spanish version of the results (INEE, 2013) shows that among

the fourteen European countries surveyed, Spain holds 10th position, with the worst results shown to be in listening. The survey results (see Figure 1) show that only 28% of students reach a B1 level; the skill of listening is particularly worrying with only 24% reaching CEFR B1 level, a panorama which does not seem promising in terms of meeting the required standards set by the OCDE. In Andalusia the results are even worse, with only 12% of students reaching a CEFR B1 level or above in listening. The present system seems to be failing in terms of giving students the skills to be proficient in English, especially given that in Spain students begin studying a second language at a much earlier age than many of their European counterparts.

Figure 1. Distribution of CEFR levels achieved in oral comprehension in the European language survey (INEE, 2013, p.49)



Within the European arena member states have followed several different educational reform paths and changes to assessment systems with a view to address the need to improve foreign language learning. El Ministerio de Educación, Cultura y Deporte

(MECD), would be well-advised to look towards these reforms with a view to seeing what lessons could be learnt and subsequently applied to its own reforms. One thing is certain, the development and implementation of a national language assessment scheme is not an easy task and requires a number of years to complete. Eckes, Ellis, Kalnberzina, Pižorn, Springer, Szollás and Tsagari (2005) for example, note that the introduction of a centralised skills-based exam took 10 years to implement in the Baltic states. The other over-riding consideration is that, in accordance with European plurilingualism policies and so as to promote educational and professional mobility within the EU, any given national exam should be designed in such a way as to allow the linking of candidate performance to CEFR competence levels if it is to respect principles of transparency, comparability and coherence (CoE, 2008). An internationally valid certification is necessary to ensure that competence levels have the same meaning across Europe, which means that both theory-driven work at the construct level and data-driven empirical investigation are needed in order to make strong validity claims about an exam's relationship to the CEFR.

Amongst the various difficulties encountered in other countries, the lack of expertise in language test construction and validation has been cited as a major problem (see for example Eckes et al*.,* 2005, p.359), with several countries bringing in international expert consultants. Standardisation and validation work is essential, pilot studies are necessary to produce reliable tests that can be considered parallel in difficulty over administrations. This requires political willingness as well as investment and it has been noted that many European educational reforms have been hindered by political failings (Eckes et al*.,* 2005, p.375).

Many European countries, such as Sweden, Austria, Ireland and notably the Dutch CEFR construct project,[3] now attempt to align their educational systems to CEFR levels. For example, Slovenia (4th ranked in the European survey for listening) developed a new *Matura* to replace the university entrance exam over 20 years ago in 1995 (Eckes et al*.,*

[3] See http://www.research.lancs.ac.uk/portal/en/projects/the-dutch-cef-construct-project(4679872d-dd67-421f-b961-9ea6184d48ae).html

2005, p.367) and began work on relating the exam to the CEFR in 2008 (see Ilc & Stopar, 2015). This exam has certain parallels with the proposed new BFE exam in Spain in that a foreign language is included in the compulsory core subjects to be examined. The exam includes the four macro skills as well as language use and has two difficulty levels (basic and higher). France (ranked below Spain in the European Survey for listening) introduced reforms somewhat later, and foreign language study was not integrated into the general primary curriculum until 2002. The required exit level for lower secondary school is CEFR A2 (possibly explaining the poor results in the European survey), with CEFR B1 required for upper secondary. However, this assessment is not based on a national standardised exam but on internal teacher assessment, similar to the 60% of the final mark proposed by the Spanish reform (BOE, 2014). However, a special CEFR-related language certificate has been created for universities. This is perhaps similar to the situation in Spain, where language accreditation is the responsibility of universities, and begs the question as to whether or not national standardised exams should be introduced earlier in the curriculum.

Indeed, the implementation of the European Higher Education Area (EHEA) as part of the Bologna process in Spain in 2007 has brought about certain positive changes in language accreditation in universities. One consequence of Spain's recent integration into EHEA is that presently most universities offer foreign language accreditation exams in order to comply with new reforms requiring certification for graduation and beginning master's study. The ACLES association has played an important role in this process and has been instrumental in contributing to assessment literacy in university language centres. At present most universities require a CEFR B1 accreditation for graduation and master's study, the proficiency level that should be required at age 15 by 2020. Accreditation in a foreign language can also be granted by numerous international bodies apart from ACLES.[4] Consequently, the number of people doing international exams has increased substantially and, as the possibility of receiving a CEFR-related accreditation is not offered within the national education system, students are obliged to look elsewhere

[4]A list of accepted accreditation exams can be found in Annex 2 at http://www.acreditacion.crue.org/Documents/Interes/Modelo_acreditacion_ACLES.pdf.

and pay for the privilege, causing a subsequent outflow of money from the Spanish economy.[5]

Besides the economic results, such extra cost also makes for an unfair system that does not follow equal opportunities standards. In a recent impact study which interviewed teachers on the use of Cambridge exams in a primary context, Breeze and Roothooft (2014) reported that many teachers, especially those in state schools, were concerned about issues of fairness and equal opportunity resulting from the use of commercial exams paid for by parents in the public education system. In addition, there was a belief that external assessment should be provided by the Spanish state sector itself. I personally believe that students should be able to certify their language proficiency level as part of the free secondary educational system if fairness and equal opportunity is to be achieved.

All of the above would suggest a national standardised exam is therefore necessary. Indeed, it has been reported that such standardised exams have positive effects: for example, academic performance has been reported to be significantly better in countries that have external exit-exam systems (Hanushek & Woessmann, 2010). The PISA study (OECD, 2010) reported that 24 of the 34 OECD countries studied have an external standardised exam – with those countries that do tending to produce results some 16 points higher than those who do not. This is an aspect which will be discussed in more detail in chapter 3 in relation to exam impact. Austria is one country which has successfully implemented a compulsory CEFR-linked school-leaving exam.[6] To my mind, reasons for the success of this project include the following: international experts were brought on board; a centralised expert body was created to oversee the implementation of the exam; the exam was introduced in a piecemeal manner with a few pilot schools taking part in the initial stage; and teachers were included in the process from the beginning and given training in item writing, correction and so on. In Slovenia, mixed success has been reported (Pižorn & Nagy, 2009), with decision makers´ reluctance to understand the need for assessment cited as the main obstacle. Similarly, in

[5]However, since 2017 Granada University offers *CertAcles* exams to its students free of charge.
[6]For details see https://www.bifie.at/srdp.

Hungary initial progress was later thwarted by staffing issues and a lack of follow through procedures for quality control (Pižorn & Nagy, 2009).

In comparison with such initiatives elsewhere in Europe, Spain is clearly lagging behind. Results are currently still poor and if any real progress is to be made, changes to its education system and testing procedures need to be approached in a thorough, systematic and professional manner.

2.2 The present university entrance system in Spain

The Spanish educational system reflects the political organisation of the country. There are 17 separate regional governments, each with varying levels of legislative autonomy regarding educational policy. It is therefore essentially a decentralised education system. However, all regions must abide by national regulatory rulings ('Decretos Nacionales') enacted by the Ministry of Education, Culture and Sport (MECD). Local governments can then adapt these national regulations through regional legislation.

Baccalaureate is a non-compulsory stage in education from the ages of 16 to 18. At present a university entrance exam (PAU), often known as 'selectividad', must be taken at the end of upper secondary school studies. Each region has a 'Comisión de Selectividad', a group of experienced university lecturers responsible for producing the University entrance exam. The current English section of the PAU has had the same format for over twenty years (in fact Fernández Álvarez (2007) states that essentially the exam has been unchanged since 1984). The English assessment component has included a short text with comprehension questions, a few grammar exercises, and a short 120-200 word essay; a very traditional grammar-focused approach, which has little relevance to current communicative foreign language needs. Indeed, this exam fails to reflect the very communicative competences required by the current curriculum (Garcia Laborda, 2010; Fernandez Alvarez, 2007). I will now go on to discuss some of the various criticisms that have been levelled at the exam in this regard.

2.2.1 Criticisms of the present PAU

The English PAU exam has been criticised by a number of researchers throughout its existence. Yet, considering the high-stakes nature of the exam, these criticisms have not so far led to any changes being implemented by education authorities. In addition, as a regionalised exam under the control of independent bodies, studies of the PAU have tended to involve relatively small samples or focus on situations in individual Spanish regions.

The actual development of the exam does not follow international standards for the production of high-stakes exams. There are no test specifications, no piloting or statistical analysis of exam results takes place, and there is neither benchmarking of production tasks nor reliability checks. The persons responsible for writing the exam need not be experts in language testing and, as García Laborda (2012) has pointed out, the person in charge may be from any field of language, literature or linguistics; normally, a language testing expert is not a member of the commission (López Navas, 2012). The fact that non-experts are predominantly in charge also means that reliability and validity studies have been few and far between and have not shown positive results (e.g., Amengual Pizarro, 2003, 2006; Gila González, 1996; Herrera Soler, 1999, 2000-2001; Watts & García Carbonell, 1999, 2005). These studies have not been widely acknowledged and have largely been ignored by the competent administrations (Bueno & Luque, 2012). Indeed, the only statistical information made available is mainly concerned simply with the number of pass/fail students. Following López Navas (2012, 2015), there is a clear need to implement procedures that evaluate both the reliability and validity of the exam, which demands expert involvement in the test development and administration process.

Added to the fact that each autonomous region produces a different exam, no comparability studies are currently being undertaken and consequently, given that the exam purpose means that a student can use the results to enter any university in Spain, the issue of fairness becomes obvious (i.e., University admissions are not comparing like with like). The fact that non-experts are in charge of producing the exam and no rigorous

test development process is implemented also means that a lot of mistakes have been highlighted in exam content. Consequently, there are now calls for a national body of experts to oversee the creation and implementation of a new exam (e.g., Fernández Álvarez, 2007; López Navas, 2012, 2015).

Validity studies, have not only demonstrated that the exam is not a valid evaluation of the skills it purports to test (i.e., reading, writing and grammar), but have also have underlined the need for the inclusion of an oral section. The exam as it stands is not construct valid. For example, Sanz Sainz and Fernández Álvarez (2005) argue that, based on curriculum requirements, students should have reached a level of communicative competence which represents a B1 CEFR level by the time they take the PAU. They concluded that apart from the obvious lack of any oral component, for which students should have been prepared, the skills examined are not an appropriate test of the construct. The exam is not testing what it should; as Sanz Sainz and Fernández Álvarez (2005, p.2) put it "it lacks construct validity: the items do not present candidates with meaningful, purposeful activities; the test does not measure students´ communicative language ability". Similarly, González-Such, Jornet and Bakieva (2013) conclude that the present exam has no underlying construct definitions and cannot therefore be considered either fair or ethical.

Opinions of other stakeholders canvassed also show similarly negative perceptions. Both teacher and student opinions have highlighted negative views of the present system. For example, a small study by García Laborda, Bejarano and Simons (2012) gathered information on student perspectives of their secondary education in English. It was found that the present PAU very much influenced the content of classes during baccalaureate studies, with emphasis placed on the exam; as a result, students felt they made little progress in gaining communicative competence in English. The study showed that baccalaureate students lacked motivation, did not use their L2 in class for communication and consequently were lacking in confidence. In short, they were not satisfied with their English classes and many students believed that the language education they had received in high school had a very limited effect.

It is in this respect that the exam has received the most criticism, i.e., it produces negative impact (both at a micro and a macro level). Both speaking and listening are included as competences in the actual baccalaureate curriculum, and have been included in previous national curriculums, yet no provision has been made for an oral section in the university entrance exam.[7] This has led to numerous criticisms of teachers teaching to the test, ignoring this aspect in the English language classroom.

Indeed, the main criticism of the PAU is without a doubt the perceived negative influence the exam has on educational practice in Spain. Here, a number of researchers have raised concerns about the negative impact of the PAU in the English language classroom. It has been shown that the curriculum is in fact narrowed because teachers do indeed teach to the test. This is unsurprising and has been reported in a number of different contexts (e.g., Alderson & Wall, 1993; Cheng, 2008). Alderson and Wall's study confirmed that "the examination has had a demonstrable effect on the content of language lessons" (1993, p.126-127), a phenomenon also reported in studies in Spain criticising the PAU. Fernández Álvarez (2007) found that 75% of the teachers who completed a survey about the PAU felt pressured by their students to prepare them for this exam. Therefore, it is not surprising that teachers feel that they should teach using more traditional methods. Rubio and Tamayo Rodríguez (2012) put this down to the lack of an oral component in assessment and found that the evaluation criteria set by the law are neglected.[8] In a small study, Amengual Pizarro (2009) used results from a questionnaire to highlight the impact of the university entrance examination on the teaching of English. She found that although teachers were willing to include all aspects of the Spanish national curriculum in reality they tended to concentrate on the content of the PAU exam, as they were thinking about their students' marks. The teachers devoted less than a third of class time to developing skills not included in the exam. Teachers also reported that they would have changed their teaching methodology if they had not had to teach to the PAU exam (Amengual Pizarro, 2009, p. 586-590). It would therefore

[7]However, both Galicia and Catalonia have introduced a listening section to their exam.

[8]Although their study was about lower secondary.

certainly be interesting to research any positive washback in terms of students' communicative competence if an oral component was introduced into the exam. Similarly, Tragant, Miralpeix, Serrano, Pahissa, Navés, Gilabert and Serra (2014) surveyed high school teachers and found that though recognised as important by teachers, oral skills were not practised in class as much as teachers would have liked because the skill was not evaluated in the PAU. However, other reasons, such as large class sizes, lack of time, and the difficulty of getting students to speak English were also given. García Laborda and Fernández Álvarez (2011) again found that teachers do have an interest in developing speaking and listening components in the classroom, but current classes are mainly devoted to grammar and translation skills, as this type of task is included in the PAU exam. Teachers used mainly non-authentic exercises from textbooks and over 75% of the teachers devoted less than 10% of their class time to listening activities; in other words, the skill is very much being ignored in the classroom. Harris (2002) and Romero Garcia (2003) also found that, as there is no oral component in the exam these are less concentrated on in class. Summing up, García Laborda and Fernández Álvarez (2012) state that there is clear evidence in Spain for teachers teaching to the test, even if those same teachers feel that students are missing learning possibilities. Here, there is a clear case for the potential of well-directed washback in the classroom in improving the quality of language learning in Spain, where class time is shown to be dedicated to those aspects which appear on the exam (Amengual, 2010). García Laborda and Martín-Monje (2013) also point out the social impact that would be brought about by such positive washback, by providing citizens with the necessary skills to work in a globalised economy, that is, the effects would go beyond the classroom.[9]

In sum, the exam has not been valid either in terms of construct validity or construct coverage; as López Navas (2015) has put it "the PAU examination does not evaluate the Communicative Competence of students or reflect the established parameters in the baccalaureate curriculum for foreign languages". There is a clear demand for an oral

[9]They give the example of the demand for bilingual nurses in the European union.

section, both listening and speaking, to address the repeated criticisms and plan for positive test impact.

2.3 Educational reform in Spain.

The criticisms outlined above do presently appear to have been recognised, and in recent years attempts have been made by the government to rectify the situation through educational reform laws. The Ley Orgánica de Educación (LOE) was published in 2006 (BOE, 2006). However, according to Fernández Álvarez (2007), while this law recognised the need for cooperation between education departments and universities in the development and realisation of the PAU, it still did not address some of the main problems with the foreign language part of the exam. Specifically, the law did not refer to quality controls of the exam nor make any reference to linking the exam to the CEFR. Most disappointingly an oral component was not introduced.

The Royal Decree 1467/2007, of 2 November (BOE, 2007) established the way the baccalaureate was to be structured, minimum teaching requirements, and official documents for internal evaluation, to be followed by the PAU university entrance exam. This law referenced an oral component but still no reference was made to the CEFR. The law stipulated common objectives, content and evaluation criteria but did not refer to explicit competences, which was left to the different autonomous regions. As Bueno Alastuey and Luque Agulló (2012) have pointed out, each region has subsequently developed different curriculum content for competence in a foreign language. For example, Andalusia (*Boletín Oficial De la junta* (BOJA) 26/08/2008, no. 169, p. 98) stated that after completing baccalaureate, proficiency level should be CEFR B1. The guidelines are unclear and not standardised throughout the country and many contradictions in terms of the CEFR level expected at the end of baccalaureate study have been reported. Bueno and Luque (2012) and Couet and Arnaiz (2009) comparing the CEFR and the evaluation criteria of the 2007 Royal Decree did, however, propose that the latter was close to B1.

Royal Decree 1892/2008, of 14 November (BOE, 2008), regulated conditions for access to degree course study at Spanish public universities. The decree meant Spain now had legislation in place stipulating the need for an oral component in the foreign language part of the university entrance exam. Initially scheduled for the 2011-12 academic year, a lack of clear direction over just how the exam construct was to be defined and implemented meant the proposed reforms had to be postponed until 2014 (Royal Decree 961/2012, in BOE, 2012). The PAU was to continue only assessing written comprehension and expression.

The LOMCE was approved by the government in 2013 (La Ley Orgánica 8/2013, of 9 December, BOE, 3013) and was a modification of the LOE. This reform also introduced an external evaluation, designed by the government, of baccalaureate students as part of university entrance requirements. Under the reform the English PAU would disappear and be replaced by an achievement exam administered throughout Spain under the central supervision of the Ministry of Education. Baccalaureate students are given the opportunity to focus their studies on areas of interest, not by choosing individual subjects, but by following a particular route of specialisation. There are three such routes available, all of which include a modern foreign language as a core subject. External evaluation was set to begin in 2017 (though this would only be necessary for university entrance and not for graduation from high school) taking real effect as of 2018. To obtain this final qualification, students would have to pass all internal assessments and the final mark would be 60% internal assessment and 40% from this external centralised government exam. The LOMCE is a clear effort to bring Spain into line with other European countries, promoting plurilinguism as a priority, and recognising the need for competences in the oral skills in foreign languages. This law made passing reference to CEFR proficiency levels for the first time, stating that:

> The goal of foreign language teaching is the training of students in acceptable foreign language use beyond the standard phases of the education system and to this end it is organised into the following levels: basic, intermediate, and advanced. These levels will correspond to the levels A, B, and C of the

> Common European Framework of Reference for languages (CEFR), which are subdivided into the levels A1, A2, B1, B2, C1, C2. The structure and content of basic-level language teaching may be determined by the education authorities in question.
>
> (p. 97893, Translation of original text in Spanish by CS.)

Royal decree 412/2014, of 6 June 2014 (BOE, 2014) established the legislation for university entrance in Spain. As well as outlining the assessment criteria, it allows foreign students easier access to Spanish universities and gives universities the compensatory right to administer their own entrance exams. In addition, they now can create different entrance requirements for different degrees, consider non-academic factors when determining admissions, and offer conditional admission to foreign students. The new law outlines the introduction of an external exam which would be weighted at 40% of the final mark, the same as the present PAU system for university entrance. Only students who passed all parts of the internal evaluation would have been able to do the new 'reválida'/Baccalaureate Final Exam (BFE). The FL part of the exam would include both listening and speaking, skills which are prioritised in the national curriculum. However, the only indication of what would be included in this exam is that the obligatory first foreign language forms part of the general core subjects. These would be a compulsory part of the exam and it has been reported that there would be 350 MCQ or note form type questions designed by the MECD covering all the general core subjects.[10] The exam would be administered twice a year in situ in students' schools and students would have to obtain a minimum of 5 out of 10 to pass. Consequently, the English section of the external exam would only be worth a very small percentage of the total BFE mark and certainly would not provide a valid CEFR-related accreditation.

The new curriculum, as declared in Royal Decree 1105/2014 of 26th December (BOE, 2015), prioritises the improvement of foreign language education, where curriculum

[10]See http://www.elmundo.es/comunidad-valenciana/2015/04/12/5529572aca474170468b456d.html

content and evaluation criteria would be set by the Ministry of Education. The basic curriculum for first FL (English) is laid out in Annex 1, p. 422-447, which provides curriculum content and evaluation criteria for baccalaureate English oral comprehension in the first and second year courses. In terms of communicative linguistic competence, there is a focus on socio-linguistic, pragmatic, discourse, strategic and intercultural competences. As in the CEFR, the student is seen as a social agent. The curriculum therefore clearly defines foreign language competence in terms of a communicative competence construct as presented in the CEFR itself. However, while the curriculum does explicitly state that it follows the same action-orientated approach as that proposed by the CEFR, it provides no specific reference to the former's proficiency levels. Indeed, more detailed examination of the assessment standards for listening shows that they in fact resemble a mix of B1 – C1 competences (that is to say, one clear level has not been proposed). Furthermore, several of the competencies actually resemble CEFR spoken interaction descriptors and are not in fact descriptors specific to the CEFR listening descriptor scales. All this is highly confusing and my personal suggestion would therefore be that the assessment criteria for listening be changed to a) reflect CEFR descriptors for listening and b) target only one CEFR level in order to allow for strong claims of proficiency level in a given skill to be validly made. Nevertheless, Spain has at the very least now put in place a curriculum which goes some way to describing the linguistic competences students must possess in order to successfully integrate into University life under the new EHEA.

Since the initial proposals for the evaluation of English oral skills were outlined, further attempts have been made to give guidance on test development for individual autonomous regions and exam specifications can be found in the Orden ECD/1941 of 22 December 2016 established by BOE of the 23 December 2016 (BOE, 2016). This document lays out the specifications to be followed in order to develop the *reválida* for university entrance. For the foreign language part of the evaluation 60% of the possible marks should be allocated to the understanding of both oral and written texts (i.e., reading

and listening).[11] Also final marks are percentage based and no standard setting procedures are proposed, making it impossible to have defensible CEFR-related marks. There are five descriptors—two less than the original LOMCE (BOE, 2015, p.442)—for listening (p.13) which are not specifically taken from the CEFR listening descriptors and they still appear to contain a mix of B1 to C1 descriptors. For example, students are expected to understand irony and humor, skills which are not described in the CEFR descriptor scales and could be considered to be much higher than B2 skills, as well as 'follow extensive animated conversations between a number of interlocutors', a skill considered to be B2+ by the CEFR descriptor scales and C1 by the DIALANG self-assessment scales (see CEFR, 2001, p.234). It would seem then that the exams should include between 2 and 15 items, which should be a mix of open and semi-open (minimum 50%), and closed MCQ type items and last 90 minutes. Yet this part of the specification would include all the language use activities to be tested (reading, listening, writing and speaking). I would therefore argue that it would be difficult or even impossible to develop a valid listening exam based on these descriptors and very general test specifications, which do not allow much time to be dedicated to listening. In terms of university entrance, a separately administered CEFR-related proficiency exam would give universities as test users much more information.

In reference to CEFR levels, Lim (2013) states that, on average, school-leaving exam reform in Europe mostly requires a B2 CEFR level at the end of secondary school. In Spain, Madrid has set the goal for the majority of secondary school students to reach a CEFR B2 level in English before leaving school (Ashton, Salamoura & Diaz 2012). In bilingual schools in Madrid, students graduate at baccalaureate level with a CEFR C1 (G. Laborda, personal communication, December, 2015). There therefore seems to be much confusion about the expected CEFR level for school leaving and university entrance. As previously stated, a commonly held belief in Europe is that university entrance should be CEFR B2. Indeed, Sevilla-Pavón, Gimeno-Sanz, & García-Laborda (2017, p.3) state "we should demand that any student who passes the English PAU exam should be able to

[11] However, no explanation is given for this weighting.

carry out tasks in English in both an everyday context and in the sphere of their university studies" (Translation of original text in Spanish by CS.).

The BFE not will not only act as a proficiency exam but will also act as an achievement test for school leaving and should therefore reflect curriculum content. Here again, an examination of the new curriculum suggests CEFR B2 to be the appropriate level. Indeed, an examination of the textbooks used throughout upper secondary school shows that most students are using material aimed at B2 during their final year at secondary school.[12] Many other researchers argue that a CEFR B2 level of proficiency is most appropriate for either the labour market or university study (e.g., García Laborda & Martín Monje, 2013).

We have seen that despite a plethora of new laws, there is no real indication of either the content or the administration of the new exam. The LOMCE clearly states that a listening component would be introduced by 2017, which would become compulsory for university entrance by 2018. However, at present there is much opposition to this law and it is clear that such reforms will not be taking place in the near future. No external exam has yet been implemented and currently, the BOE (2018) stipulates the evaluation criteria for listening, which repeats those stated by BOE (2016), yet this evaluation is left up to each autonomous region to implement—an endeavour which still does not consider the implementation of an oral section for the English exam.

As previously stated, education reform is a laborious process and important changes cannot be implemented without the necessary planning and preparation. A review of research into the new listening section, however, shows such planning and preparation to be scarce. It seems that we are likely to see further delays unless steps are taken to address these issues.

---

[12] Upper secondary text books used in schools include: Living English (B1/B2), Burlington books; Top Marks, Burlington books; Viewpoints 2, Burlington books; Next Generation (B2), Cambridge; Out and about (B2), Cambridge. Advanced Contrast for Bachillerato 2.

2.4 Proposals for a new baccalaureate exam

Given the numerous criticisms of the PAU, a small number of researchers have suggested changes to the present system and furthermore have made proposals for how such changes should be made. However, this research is in no way exhaustive and has centred on four main areas. A small number of studies have looked into stakeholder opinions (mainly baccalaureate teachers). Most studies have centred on the introduction of a new speaking component, as well as how to adapt tests for computer delivery. Finally, only a handful of the studies which have suggested improvements to the present test have included a listening section. Here, I will briefly outline some of these studies and conclude that interest in a new listening section has been practically non-existent and certainly not enough to serve as a proposal for the introduction of a new exam. As García Laborda and Fernández Álvarez (2012) have pointed out, this research deficiency is mainly due to lack of expertise.

Perhaps the greatest research interest to date has been concerned with the introduction of the speaking section of the exam (e.g., see work by OPENPAU). Drawing both on a previous study about the writing section (Díez Bedmar, 2012) as well as advice given by experts (Amengual Pizarro & Méndez García, 2012), Suárez-Álvarez, González-Prieto, Fernández-Alonso, Gil and Muñiz (2014) propose that the speaking section should be based on a B1 CEFR level. They have developed test specifications, test tasks and evaluation criteria. Addressing the main issues surrounding the construct definition of a new speaking exam, Amengual Pizarro and Méndez García (2012) introduce an 'International Communicative' perspective to define a new speaking construct and argue that ideas about English as a lingua franca should be considered when defining the construct for speaking. Furthermore, their attempt to define the construct takes into account models of communicative competence, with evaluation criteria including pragmatic and strategic competence, as well as highlighting the importance of co-constructed meaning and negotiation. As such, it follows the CEFR model of language proficiency. I would personally agree that the construct of English as a lingua franca is an important consideration when talking about the Spanish context. Candidates will be

expected to use English in an international context and will be likely to communicate with other L2 English speakers. Indeed, the call for a revised construct definition placing pragmatic competence centre stage is not new (see for e.g., Canagarajah, 2006) and we certainly need to take into account communicative needs for English as an international language. To this end, it may well be that a completely new construct definition for English as a lingua franca should be developed (e.g., Harding, 2015).

Stakeholder opinions have been mainly concerned with teacher opinions of a new test format. A survey carried out by the CAMILLE group of the Polytechnic University of Valencia (Martínez, Sevilla & Gimeno, 2009) put forward possible tasks to be included in a new university entrance exam and asked teachers to rate them on preference. The most popular choices for the listening part of the exam were just one audio clip of two or three minutes followed by six multiple-choice items or one audio clip followed by eight True/False items. Similarly, Sevilla-Pavón et al. (2017) found that the highest proportion of teachers (21.63%) preferred the option of listening to a 2-3 minute video clip followed by six MCQ items. This was followed by the preference for a similar 2-3 minute video clip with true/false type items. Here, I would highlight the lack of assessment literacy held by teachers. Teachers are probably unaware that such measures of listening proficiency would be unlikely to provide either a valid or reliable measure of students listening ability. On a side note, teachers from both studies were also concerned about possible technical problems on the day of the exam, as well as having queries about the number and duration of sound files. They obviously felt uninformed about the new proposed exam and this highlights the need for detailed test specifications to be placed in the public domain along with an example exam. Certainly, if we are to follow international guidelines for test developers, we will be expected to make clear example exams available to both teachers and students (e.g., EALTA, 2006).

The use of computers and even tablets and mobile phones for test delivery has been another main line of investigation (Garcia Laborda, 2010, 2012; García Laborda & Gimeno Sanz, 2007; García Laborda & Martín Monje, 2013; Martín Monje, 2012). While mainly concerned with the delivery of a new speaking test, this area is something which

should also be taken into account when devising a new listening test. García Laborda, Gimeno Sanz and Martínez Sanz (2008) canvassed teacher opinion on a computerised oral exam in order to discover if it would be well received by teachers. The results were positive, although some teachers raised concerns about the availability of the necessary technology. García Laborda, (2010, p. 77) also points out that a computer-delivered exam would allow for the easy collection and subsequent analysis of data and therefore provide an impetus for change, especially in terms of teaching methodology. The advantages and disadvantages of using computers for test delivery are summarised by García Laborda, Magal Royo and Bárcena Madera (2015). They present a SWOT analysis and highlight the positive aspects of such a delivery method, such as the progressive use of technology in the classroom, the need to standardise language tests and the need to obtain objective data to take educational decisions. I would also note here that, given that the new curriculum includes digital literacies, the logical progression would be to use a computer delivery method for the external exam. Similarly, Gimeno Sanz and De Siqueira Rocha (2009) point out that computer tools provide far better and faster feedback than traditional procedures. García Laborda, Magal Royo and Bakieva (2010) also observed that students using computer-based language testing were more motivated towards the test, and Bueno Alastuey and Luque Agulló (2012) highlight that test anxiety would be reduced. The current OPENPAU project is working towards the integration of skills through a computer-based test approach as well as a paper-based one. This is certainly a major consideration of a new test; following García Laborda (2006), I would propose that, given the large numbers who will take the test, item types chosen should facilitate automatic correction regardless of whether candidate responses are computer- or paper-based (for listening, García Laborda specifically proposes MCQ and note form task types).[13]

While there have been some attempts at improving the present system through the outlining of prototypes for a new test, research for its actual development is extremely limited, especially for the listening construct. For example, the previously mentioned

[13] Ideas for the automation of online university entrance exams are expanded on in Magal-Royo and Laborda (2017).

study by Martínez Saez et al. (2009) proposed a model exam following preferences put forward by the teachers they had interviewed. The listening section included one audio clip followed by five MCQ items. Bueno Alastuey and Luque Agulló (2012) outlined the introduction of a listening test, including interactions, recorded material and academic presentations relevant to university life. Again, the suggestions they made were based on the aforementioned survey of teacher preferences; they concluded that the test should be CEFR B1 level using four-option multiple choice items delivered by computer. Besides a reliance on teacher opinions, no attempt has been made to either define the construct or develop test specifications. González-Such et al. (2013), citing construct validity as the most important feature of any test, have however claimed to define the construct. They put forward a proposal for a revision of the oral section of the exam, while admitting that the proposal is not exhaustive. They also recognise the need to include aspects of communicative competence in their construct definition (i.e., linguistic, pragmatic and socio-linguistic competence), and stress the importance of knowing exactly what it is that test developers wish to evaluate.

Other research projects have attempted to use teachers' ideas to implement the new tasks but, after a moderating process, this proved to be a far too traditional approach, with too many open questions to be considered possible for a computer-based exam. García Laborda, Bakieva, González Such and Sevilla Pavón (2010) suggest using two mini-clip tasks delivered by computer: one note form and one MCQ. García Laborda and Martín Monje (2013) have elaborated on this suggestion in greater detail, providing example items for listening comprehension (p.81). Here the emphasis is on the mode of delivery, in an attempt to speed up the correction process, rather than on the test content itself.

Possibly the most detailed study to date is that of Fernández Álvarez (2007), who redesigned and piloted items for a completely new English PAU exam for assessing the four skills. His proposed exam includes a 10-minute listening paper with three tasks based on the Andalusian baccalaureate curriculum (which he considered to be CEFR B1 level). López Navas (2012) argues that his study demonstrates that the curriculum is not the most important issue, but rather the way in which it is misrepresented in the construct

of the English PAU exam. Fernández Álvarez found that nearly 35% of baccalaureate teachers and 64% of students believed an oral comprehension mark should be a component of the overall English proficiency mark. However, when teachers were asked about the number and type of tasks they would prefer to see in this section of the test, they again showed a preference for just one task (45%), to be delivered as a short video followed by comprehension questions. Most teachers and students thought this part of the test should last about fifteen minutes. Fernández Álvarez's thesis goes on to design, pilot and statistically analyse a new BFE test. However, it should be noted that he makes no claim to have developed a perfect exam; rather, his aim was to show the authorities the way, by suggesting how changes could and should be made with the necessary backing. While certain parts of his study on oral comprehension may be criticised, it is certainly the most comprehensive study to date.

To conclude, it is clear from the above review that research proposals for a new listening part of the BFE are extremely scarce. It is precisely this gap which the present study intends to fill. A test development process needs to be begun well before it can be implemented and subsequently improved on as part of the ongoing test development cycle for university admission in Spain.

2.5 The CEFR

Since its publication in 2001, the Common European Framework is unquestionably the document with the greatest repercussions on the formation of recent language policies. As the major reference document for language education and assessment in Europe, it has brought with it a challenge to governments to change educational policies. The CEFR now frames language education policy and aims to provide common standards for levels of L2 proficiency in Europe. The framework consists of descriptive scales for six common reference levels of language proficiency (A1, A2, B1, B2, C1, C2) with both a horizontal and a vertical aspect referring to quality and quantity of language proficiency. It is not a linear scale; rather, language ability is represented as increasing in broad bands of proficiency (North, 2014). To date, the CEFR has had a diverse and

partial impact on different countries (Little, 2011), and Spain itself is only just beginning to make reference to the document. For example, in the LOMCE (2013) a specific reference is made to the CEFR levels (A, B and C) for the first time with respect to FL. We should take into account that it is increasingly the benchmark by which the foreign language qualifications of most European countries are judged, with many authors observing that assessment has in fact become its main use (e.g., Coste, 2007; Fulcher, 2008; Little 2007, 2011). Indeed, the recommendation on the use of the CEFR by the Council of Ministers includes the call for countries to:

> ensure that all tests, examinations and assessment procedures leading to officially recognised language qualifications take full account of the relevant aspects of language use and language competences as set out in the CEFR, that they are conducted in accordance with internationally recognised principles of good practice and quality management, and that the procedures to relate these tests and examinations to the common reference levels (A1-C2) of the CEFR are carried out in a reliable and transparent manner.
>
> (Council of Europe, 2008, p.4)

Despite the widespread acceptance of CEFR proficiency levels, especially in the context of assessment, it has also received numerous criticisms. Fulcher (2004, 2008) and Hulstijn (2007) criticise the CEFR on the grounds that its descriptive scales are empirically derived on the basis of teacher judgements, are therefore atheoretical, and have no basis in second language acquisition (SLA) research, with language testing bodies being forced to link their tests to CEFR proficiency levels. Furthermore, equal attention is not paid to all the macro skills (Alderson, 2007; Davidson & Fulcher, 2007; North, 2014; Weir, 2005a). Similarly, it is argued that the CEFR lacks the theoretical rigour needed to build tests and that it is not comprehensive enough to be sufficient as a tool for language test development (Alderson, Figueras, Kuijper, Nold, Takala & Tardieu, 2004, 2006; Weir, 2005a). It is vague and there are problems with wording (Alderson, 2007; Alderson et al., 2004, 2006; Fulcher, 2004, 2008; Weir, 2005a). Indeed, Simons

and Colpaert (2014) recently reported the results of a survey about the CEFR, and concluded that people found the terminology vague and mostly wanted more fine-tuning. Jones and Seville (2009, p.51) argue that there are fears that the CEFR is being used as "an instrument of centralisation and harmonisation". Kaftanjieva (2009) argues that linking procedures lack quality and that consequently qualifications across Europe cannot be compared. Similarly, North acknowledges that "unfortunately, in many contexts a CEFR level (e.g., 'B1') continues to be plucked out of the air without an assessment of the realism of the objective or a consideration of the investment that would be necessary to achieve it" (2007, p.25). Consequently, Alderson (2007) has called for a regulatory body to oversee the validity of tests claiming to be at a certain CEFR level. At present, the Council of Europe plays no role in the monitoring of the levels used in individual countries and of how accurately they correspond to the CEFR (Coste, 2007; Goullier, 2007).

However, the CEFR was not developed as a prescriptive tool but 'is purely descriptive' (CoE, 2008). The CEFR is not a complete blueprint for language test development, but instead intended to provide guidelines as a point of departure. This has been stated to be its main advantage: it can be used and adapted in different contexts, yet it is not prescriptive nor is it intended to be the only document to be used when implementing changes in educational policies. Indeed, Davidson and Fulcher (2007) argue that although it does not provide enough detail to build test specifications, it can be adapted to local needs and the fact that it is underspecified is therefore beneficial. They argue that it should be viewed as heuristic in nature and that it gives the test developer a starting point. It should be used selectively in each context of use. A similar response is given by North, Martyniuk and Panthier (2010), who stress that the CEFR is descriptive and was never meant to be normative. It is both language and context neutral and should therefore be taken and fleshed out to make it relevant to each specific context. In short, we are encouraged to add to the CEFR in our own assessment contexts. North (2014) goes on to argue that the CEFR descriptor scales are to be used for profiling rather than levelling; as such the macro skills should be evaluated separately as learners can have very different profiles across language skills.

We can conclude that using the CEFR in isolation is not enough and that, as well as taking into account our specific assessment context, other documents provided by the Council of Europe should also be consulted. Goullier points out that "consistency is more achievable when the CEFR is used not as an isolated document, but as part of an overall approach incorporating other language policy instruments developed by the Council of Europe" (2007, p.18). Other documents have been created to complement the CEFR in the development of educational and assessment policies. The Manual for Relating Language Examinations to the Common European Framework of Reference for Languages (CoE, 2009) gives guidance on familiarisation, specification, standardisation and empirical validation when linking tests to the CEFR and comes with a range of reference supplements intended to assist test developers. The Manual for language test development and examining (ALTE, 2011) "aims to provide a coherent guide to test development in general which will be useful in developing tests for a range of purposes, ... this manual is for anyone interested in developing and using language tests which can relate to the CEFR" (Milanovic, in ALTE, 2011, p.8). Specifically regarding the skill of listening, there exists further documentation for the specification and description of tasks. The Dutch grid for listening allows for an exhaustive analysis of listening test tasks when relating tasks to the CEFR. These documents will be discussed in more detail later in this thesis, as they will be drawn upon in my research design and methodology.

2.6 Summary

The arena in Spain is ripe for change; with the necessary legislative reforms already in place, steps now need to be taken to make sure these changes are implemented correctly. After years of academic criticism, there is no doubt that such changes are necessary. A new test relevant to the context of school leaving and university entrance and which clearly evaluates the competences outlined in the new curriculum needs to be developed to create a positive impact on teaching, learning and the wider social sphere. The Spanish educational authorities should now be working towards providing valid and internationally recognised qualifications that will equip students both for undergraduate study and full access to the European job market. If Spain wishes to be involved in the

changes currently taking place in the EU, then there can be no more delays; a significant investment in resources will need to be made. Economically, this investment may even prove beneficial to the Spanish economy. As López Navas states:

> Taking into account the importance of languages and the aim of the EU to standardise qualifications across Europe, Spain cannot afford to delay the implementation of changes for much longer if it wants to compete with other countries and also to offer students the possibility to study or work abroad.
>
> (López Navas, p.178, 2012)

This necessary communicative competence is now embedded in the national curriculum; all that is needed is for Spain to implement a CEFR-based assessment system similar to those of its European counterparts. By doing so not only will it become a more important player in the European arena but it will also be satisfying its own curricular objectives. However, while no strong construct definition exists, the interpretation of PAU scores is meaningless, and as such it is impossible to transfer scores to any external benchmark such as the CEFR (García Laborda & Martín Monje, 2013). A CEFR-related test needs to be developed which not only has a properly-stated construct but also covers the four macro skills: reading, writing, speaking and listening. I have argued that these skills should be tested at a B2 CEFR level. While there is still some discussion in this regard, there are a number of compelling reasons why this should be so. These include the fact that B2 descriptors most resemble current required curriculum content and the fact that B2 is the required university entrance level in most other European countries as it is thought to be the most appropriate level for academic study and work insertion.

In terms of the listening section, research is needed which will build on both the current curriculum and the CEFR. A clear construct definition needs to be developed in line with the CEFR and the national curriculum, and research-based evidence needs to be presented to the many stakeholders involved to justify the interpretation and uses of the test scores. These stakeholders need to be convinced that the reforms are the right ones

for the context. García Laborda and Martín Monje (2013, p.69) state "the new diploma may serve to overcome the deficiencies of the current exam if rigorous studies are undertaken and research-based decisions are reached".

The high stakes nature of such decisions means test developers have both an ethical and professional duty to ensure that the test is a valid measure of language proficiency. Such an endeavour requires expert knowledge and arguably the creation of a national expert body will be necessary for the test's correct design and administration. I would agree with López Navas (2015), who states that "a regulatory assessment body for foreign languages and a specialised group of examiners are fundamental for the successful development and implementation of future tests".

The law already states that the test will be developed and administered by a national central body. However, such a project cannot be rolled out over night. As the OECD has warned:

> Governments increasingly look at how to achieve ambitious reforms in education to improve results. But such changes are not easy to make: education change takes time, options for improvement may not be evident, groups with vested interests may hamper reforms, and politicians may face conflicting priorities or lack evidence on what can work best within the context.
>
> (OECD, 2015b, p.3).

I would argue that a high stakes test such as this must in fact be well-planned and researched and may well need to be rolled out gradually as in the Austrian Matura, where implementation of the test was only piloted on a small number of participating schools in the first few years. I would also argue that the test be developed as a separate exam. If students are to be provided with a valid CEFR accreditation, the four macro-skills must be evaluated and accredited. Results cannot simply be reported as part of an overall score of 10 for upper secondary education. Given the provision already stated by law that universities can decide on their own admissions criteria and different degree courses will

have different language proficiency demands (for example, EMI and CLIL content), the test could (at least at first) be voluntary.

As such, this thesis hopes to provide a timely contribution by presenting a rigorous validity study of a new listening test to be included in the English BFE. Any new assessment procedures should be adequately described to the many stakeholders involved (García Laborda & Martín Monje, 2013), and in a well-timed fashion. Test development is cyclical and it may well be that stakeholders call for changes in any proposed project. Such an endeavour will become a much easier task if strong validity evidence is presented to support and justify the interpretation and uses of the test scores.

Having outlined the current situation in Spain and the particular lack of any serious attempt at developing a valid listening component, the aim of this thesis is to fill that gap by developing a CEFR B2 listening section for a new BFE based on a clear construct definition. The initial test specifications themselves will draw largely on substantive theory and a large part of this thesis concerns actual test development. A validity argument approach will be adopted with the aim of providing evidence to justify the interpretation and uses of test scores. To this end, I will now examine the current literature concerning the construct of listening and validity theory and thereby outline a sound theoretical basis on which to build the present project.

Chapter 3. Literature Review

3.1 Introduction

A valid CEFR-related B2 listening test needs to be built upon a sound theoretical framework. The ALTE handbook states that a testing project must start with an explicit model of language use and competence (ALTE, 2011). In other words, we need a definition of language proficiency from which a test construct can be developed. It is therefore important that investigation and support for any theory employed be rigorous and consensus-based. Canagarajah has defined language proficiency as "the ability to use the English language effectively for specific purposes, functions, and discourses in specific communities" (2006, p. 235). The necessity of context-specific models, focusing on a construct of 'language ability in person in use', has also been increasingly echoed by others (see for example Chalhoub-Deville, 2003; O'Sullivan, Weir & Saville, 2002). The on-going debates over language proficiency models mean that test developers must use a 'pick and mix' approach in the application of theoretical models to testing practice (Fulcher, 2003); they must draw on those models applicable to the specific use demands

of a test. The CEFR (2001), similar to Fulcher's philosophy, invites us to draw on the parts which are relevant to our context. Furthermore, a test must be demonstrably useful, and it must allow for useful inferences from the results to be made (Bachman & Palmer, 1996). The test construct must therefore convincingly reflect the *Target Language Use* (TLU) domain as authentically as possible. The construct requirements for a specific needs test (e.g., air-traffic controllers) would obviously not be the same as that of a general English proficiency test for school-leavers. Once the construct has been well-defined, detailed test specifications can be drawn up to provide the blueprint for test development; that is to say, the way in which the construct is to be operationalised in the test (Alderson, Clapham & Wall, 1995). The test must then be a valid representation of the test specifications, and here evidence that this is so must be provided. Test score interpretations must be valid for the purpose of use, and therefore, as Fulcher and Davidson (2012, p.1) state "language testing is intimately concerned with validation theory".

The aims of the literature review section are twofold. Firstly, I will discuss the issues surrounding the definition of language proficiency with specific reference to the CEFR and present a definition of the listening construct, referring to the main scholars in the field in order to develop a clearly-defined model of language proficiency for the proposed test. Secondly, I will outline current thinking on language test validation in order to establish a coherent framework within which to evaluate the proposed BFE B2 listening test.

3.2 Defining language constructs.

The theoretical underpinnings of a test require that before test development can begin, a theoretical stance on the nature of language ability must first be taken. The theory-defined construct is a central issue in language testing (Chapelle, 2012). As Alderson, Clapham, and Wall (1995, p.16-17) remind us, a theory about language is an "abstract belief about what language is, what language proficiency consists of, what language learning involves, and what learners do with language.…Every test is an operationalisation of some beliefs about language."

As part of the continual debate over language constructs, a number of scholars have presented seminal work outlining theoretical models for communicative competence (Bachman, 1990; Bachman & Palmer, 1996, 2010; Canale 1983; Canale & Swain 1980; Celce-Murcia, Dornyei &Thurrell, 1995). Bachman (2007, p. 44-5) has provided a historical overview of approaches to defining the language proficiency construct in which different theoretical perspectives depend on different assumptions. At present, the prevailing theories mostly define language proficiency from either a task-based (e.g., Norris, 2002) or a competence-based (e.g., Bachman, 2002) perspective; more recently, the two have been seen to converge into an interactionalist perspective.[14]

One such interactionalist perspective is the socio-cognitive one, whereby an individual uses their knowledge and abilities in a given context (e.g., Chapelle, 1998). Another views the construct neither simply as an ability within an individual nor as the context of situation; rather it is seen as being jointly co-constructed by drawing on the linguistic and strategic knowledge in TLU situations involving language use. Here, language use is viewed as dynamic rather than static (e.g., Chalhoub-Deville, 2003). Such an approach has been called a 'language ability in person in use' construct or 'moderate interactionalist' (Bachman, 2007). Strong interactionalist approaches draw on socio-cultural theories and have mainly been discussed in terms of interactive speaking, where

[14] See Bachman (2007) for a detailed discussion of the three approaches.

the notion of co-constructed conversation has led to a definition of 'interactional competence' (see for example Walsh, 2011). A recent development in this approach is that of dynamic assessment (DA) (e.g., Poehner, 2008), which separates learning and assessment; following Vygotsky's idea of the *zone of proximal development* tasks in DA provoke change. Vygotsky argued that observations about a learner can only reveal some of their abilities and tell us nothing about nascent emerging abilities. It is argued that DA can provide scaffolding through intervention, such as mediation, prompts or leading questions, thus supporting learner development and allowing students to perform beyond their current capabilities. As such, DA has been mainly applied to classroom assessment as it shares some features of formative assessment (Rea-Dickins & Poehner, 2011), rather than proficiency testing. However, ideas about assessment to support and promote learning are certainly becoming more prominent (e.g., Hamp-Lyons & Tavares, 2011); as such DA may well be suitable for introduction at an earlier point in the Spanish education system.

The Communicative Language Ability model (CLA) (Bachman, 1990; Bachman & Palmer, 1996, 2010) has been extremely influential in language testing; it is 'an ability to use language communicatively' framework (Bachman, 1990, p.81). This model views language competence as consisting of language knowledge, organisational knowledge (grammatical knowledge and textual knowledge), pragmatic knowledge (functional and socio-linguistic knowledge) and strategic competence. With reference to discussions surrounding language constructs, Bachman (2007) argues that for practical purposes test developers need to develop local theories or 'operational models', with an *Assessment Use Argument* (AUA) guiding the design and development of a specific language assessment. Bachman and Palmer (2010, p.33) remind us that in language assessment we are interested in language ability, which consists of language knowledge and strategic competence. Language users interact with language use tasks, i.e., with the characteristics of a particular situation. Here, 'listening' or 'oral comprehension' would be considered one specific language use activity, an interactive activity which consists of constructing meaning and discourse by using language ability in the context of the situation.

Similarly, Fulcher and Davidson see a language model as an abstract, theoretical description about language knowledge and language use from which an assessment framework may be developed. In this regard "frameworks are selections of skills and abilities from a model that are relevant to a specific assessment context" (Fulcher and Davidson, 2007, p. 36). It is precisely such a model of language proficiency as outlined in the CEFR which will need to be taken into account in the present study.

3.2.1 The CEFR and language proficiency testing

The CEFR is essentially a task-based approach, where students have to show that they can carry out some kind of task represented by 'can do' statements and in order to carry out such tasks candidates must however have the relevant knowledge, skills and abilities (KSAs). The CEFR model of language proficiency includes a domain definition as well a description of abilities or traits and is therefore in effect an interactionalist approach, based on a post-positivist perspective such as Chalhoub-Deville's (2003) 'language ability-in person-in-context'. It is divided into a descriptive scheme, emphasising the complexity of foreign language learning, and a series of proficiency scales for oral and written reception and production as well as spoken interaction and mediation. The proficiency scales for each of these traits have both a horizontal and a vertical element (representing quantity and quality of language use), where the language user as a social agent increases in proficiency level as contexts become more complicated and require more complex language skills. Learner competence is described within a socio-linguistic model of communicative competence which includes linguistic, strategic, pragmatic and sociolinguistic competences.[15]

The first dimension deals with language activities with respect to contexts of use outlined in the CEFR (CoE, 2001, p.44-56). Quantity can be defined as the number of domains, functions, notions, situations and locations that the learner can deal with. The second dimension is general competencies (such as general knowledge) and

[15] This is very similar to the CLA model (Bachman, 1990; Bachman & Palmer, 1996).

communicative competence (linguistic, socio-linguistic, pragmatic and strategic) and refers to quality: the degree to which language is used effectively and precisely, as well as how efficiently communication takes place.

Communicative language competence then, can be considered as comprising several components, each consisting of relevant KSAs described in detail in the CEFR:

1. Linguistic. The range and quality of knowledge about a language system, as well as the storage, retrieval and cognitive organisation of such knowledge (i.e., the extent to which this knowledge is readily accessible). It includes knowledge about lexical, phonological and syntactical aspects of a language system.

2. Socio-linguistic. The sociocultural conditions of language use, such as rules of politeness, affecting all communication between participants of different cultural groups.

3. Pragmatic. The functional use of linguistic resources; the ability to communicate successfully in a meaningful and coherent way within a particular socio-cultural group.

A language user not only needs to understand literal meanings but also to be able to decipher sociolinguistic and pragmatic implications. For Bachman and Palmer (1996) language use is considered to be interaction between a language user and his or her particular context. Here, ‘pragmatic knowledge’ refers to the way in which users relate language production to their communicative intent within the specific features of a language-use environment. In other words, meaningful language is only ever produced in a socially-mediated, communicative context.

Language interaction and cultural environments play an important role in successful communication and the CEFR (CoE, 2001, p.123 – p.129) describes pragmatic competences as comprising:

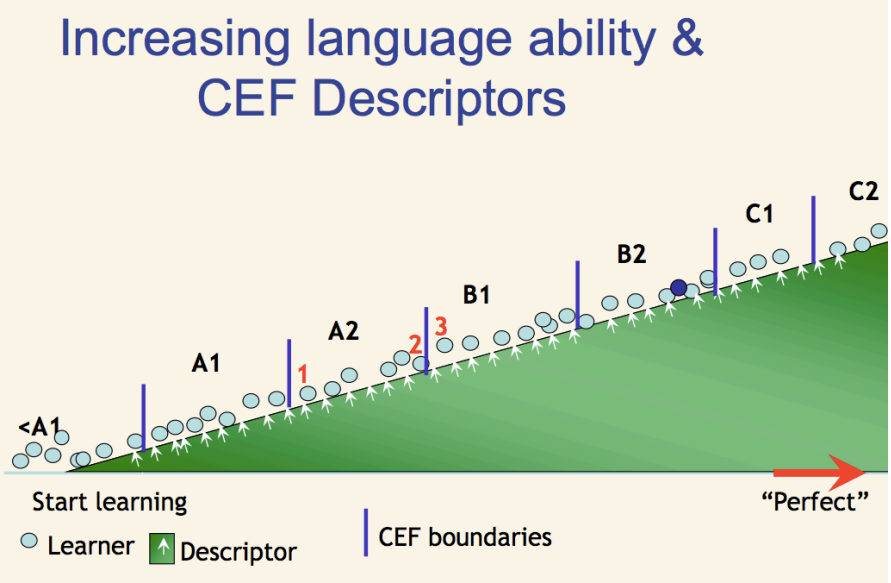
1. Discourse competence.
2. Functional competence

These competences are governed by strategic competence, which will be further discussed in section 3.3.3 with regard to the listening construct. The CEFR uses an action-orientated approach which is summarised here:

> Language use, embracing language learning, comprises the actions performed by persons who as individuals and as social agents develop a range of ***competences***, both ***general*** and in particular ***communicative language competences***. They draw on the competences at their disposal in various ***contexts*** under various ***conditions*** and under various ***constraints*** to engage in ***language activities*** involving ***language processes*** to produce and/or receive ***texts*** in relation to ***themes*** in specific ***domains***, activating those ***strategies*** which seem most appropriate for carrying out the ***tasks*** to be accomplished. The monitoring of these actions by the participants leads to the reinforcement or modification of their competences.
>
> (CoE, 2001, p.9; emphasis in original)

The calibrated scales across the six proficiency levels (A1, A2, B1, B2, C1 and C2) provide criterion reference bands representing how far up the language-learning ladder a student has reached. Each skill is therefore represented on a unidimensional scale and is presented graphically in Figure 2. This figure clearly shows that learners can move up the unidimensional scale; as they have more ability, they are able to perform more difficult tasks until they cross a boundary which takes them from one proficiency level to another. Learners can have mixed profiles and be at different proficiency levels in different skills.

Figure 2. Graphical representation of a CEFR unidimensional proficiency scale (De Jong, 2014).



However, while the CEFR provides common reference points for language testing, it does not provide guidance for test development. A valid claim for CEFR linkage must be supported by evidence; test developers must show that their test is representative of, for example, listening proficiency at CEFR B2 level. In response to this need, a manual for relating exams to the CEFR was produced (CoE, 2009), along with various reference supplements. This linking process, which will be discussed in greater detail in the methodology section, consists of four inter-related activities:

1. Familiarisation. Members of any linking panel must be familiar with the content of the CEFR and its scales.
2. Specification. This should include a detailed description of the test and its relationship to CEFR categories. The aim of this stage is to build a linking claim about the content relevance of the exam to the CEFR. Specification forms for mapping the test to the CEFR are provided by the Council of Europe (2009). For listening, the recommended form is the Dutch CEFR construct grid originally developed by Alderson et al. (2004, 2006) in an attempt to aid test developers specify proficiency tests.
3. Standardisation. Standard setting techniques are outlined in the manual.

4. Empirical validation. Both internal and external validation projects are required. Indeed, Alderson (2012) points out that if a test is not valid or reliable it is meaningless to link it to the CEFR and we still need to produce evidence of traditional validities.

A guideline has also been produced to help testing bodies design CEFR-related tests (ALTE, 2011). This document advises the use of a validity argument approach, where test validation is built into the whole process of test design and implementation. Throughout the test development cycle validity evidence must be collected to support the validity argument of a test. Such a test development cycle for a new test is recommended by Green and Spoettl (2011) as shown in Figure 3.

Figure 3. Test development cycle (Taken from Green & Spoettl, 2011)

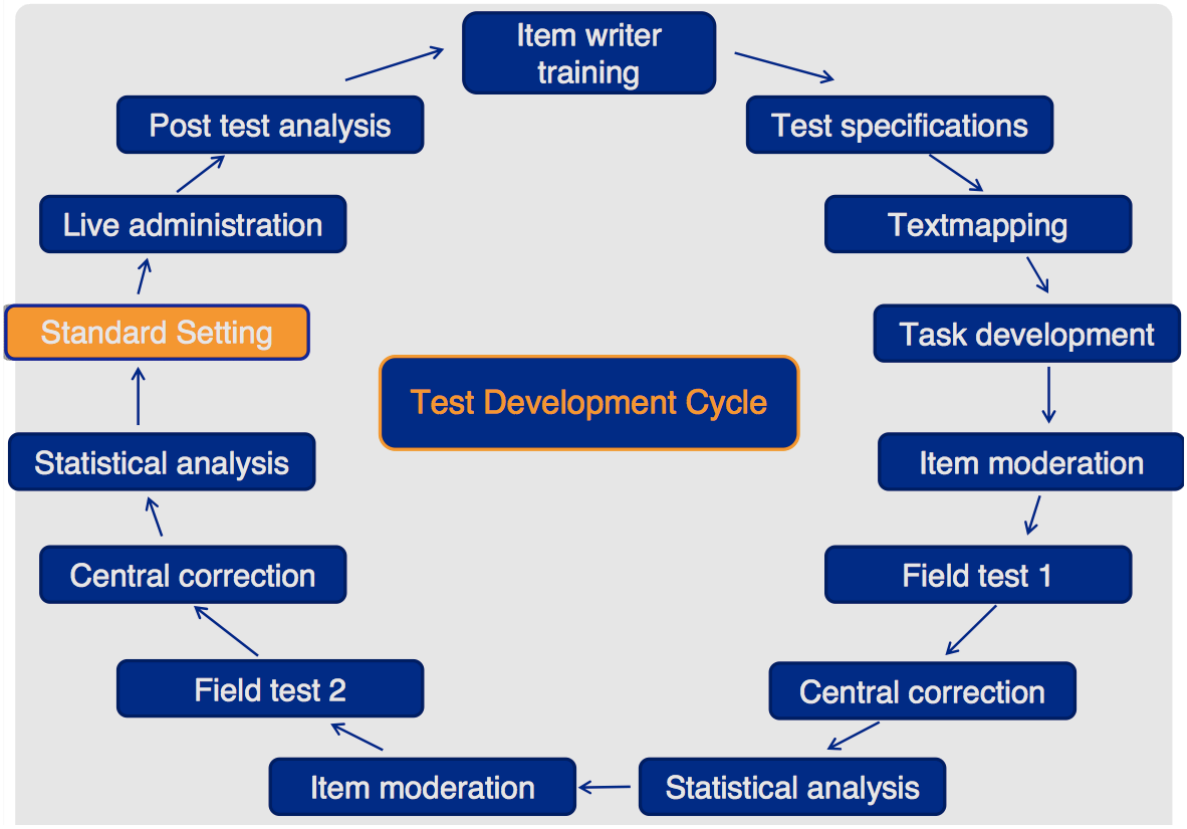


Figure 4. Chain of reasoning in a validity argument (ALTE, 2011, p. 15)

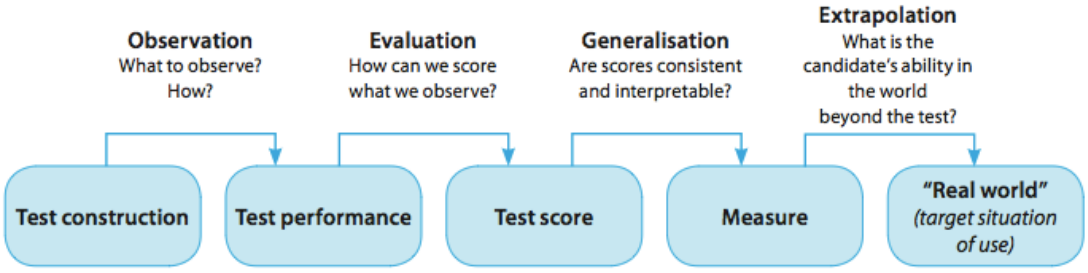


Figure 4 shows the validity argument approach recommended by ALTE. It does not follow Bachman (2005) and Bachman and Palmer (2010) as it begins with the test construct, rather than decisions based on test scores. Instead, it links performance on test tasks to an inference about test takers' language ability beyond the test and the argument ends with the 'extrapolation inference'. This inference would link the measure to a CEFR level using the 'can do' statements as a guide. Details of how to build such a validity argument are given in the appendix (ALTE, 2011, p.56 – p.81). The world beyond the test, however, is not specified; it is a general claim linking to the CEFR. Tests, however, are normally developed with a purpose, a proposed use or decision. In Kane's (2013, p.35) words "it is hard to imagine validating a test as such without having some idea of the proposed interpretation and use".[16] We could therefore argue that the ALTE model directs our interest to the interpretation of the score, that is to say, 'can the score be interpreted as a CEFR proficiency level?'. For the present study, however, the purpose of the test and the decisions based on the score interpretation are known. Test purpose is paramount and the final link should therefore be included in the *Interpretative Argument* (IA).[17]

As previously stated, the CEFR is not meant to be prescriptive. Despite its criticisms it has become an important external standard of reference, enabling reported scores to provide a meaningful, user-friendly description of what a typical test-taker can do at any

[16] Kane, unlike Bachman (2005) and Bachman and Palmer (2010), gives both score interpretation and score uses equal weight and typically a score use will rely on the relevance of score interpretation.
[17] See section 3.6

given level (Alderson et al., 2006). Following Davidson and Fulcher (2007), these descriptions are felt to be heuristic in nature. The framework gives us a design pattern, a valuable starting point and a useful tool for test development:

> The 'Can Do' statements offer guidance to educators so that they can recognise and talk about ability levels. We can use them as a guide for test development but should not feel that adopting them means the work of defining ability levels for the test has been completed.
>
> (ALTE, 2011, p.13).

We need to adapt the CEFR descriptor levels to the context of use of the test. This is particularly relevant for the present study as the test not only acts as a CEFR proficiency test, but will specifically be used as a university entrance test. Also, it should act as an achievement test for school leaving and should therefore cover the content of the national curriculum. In this sense, the test could be called a '*CEFR Plus'* test—that is to say, it is a measure of CEFR proficiency as well as another context-specific construct. As such, the descriptors included in the national curriculum should also be included in the construct definition. Such demands are by no means new; for example, De Jong (2014) has shown how new descriptors could be incorporated into the different CEFR levels using expert judgements and Rasch modelling and Díez-Bedmar (2017) used learner corpora to develop more elaborate written production descriptors at the CEFR B1 level.[18]

In sum, current language proficiency theory can be seen to recognise the complexity of language use contexts. Furthermore, it is fairly clear that while the CEFR model provides a general description of language proficiency, this in itself is not sufficient for the practice of test development. Context of use must be taken into account

[18] However, as has been previously stated, the evaluable standards presented in the Spanish curriculum for listening (BOE, 2018) may need to be adapted to be more representative of one CEFR level (B2); as they stand it would be difficult to incorporate all the descriptors in a CEFR B2 listening test specification. This is especially true as two of the listening descriptors reference spoken interaction.

and clear validity evidence must be provided throughout the test development process. Within such a framework, it is furthermore clear that the learner will need to make use of multiple abilities and strategies in order to comprehend real world communicative situations. Each language use situation must therefore be evaluated separately, as learners can have mixed ability profiles. Consequently, in order to successfully evaluate proficiency in listening ability, it is essential we start from a clear description of this theoretical construct and define what we understand the skill of listening to include.

3.3 The listening construct.

The construct of L2 listening is indeed difficult to define, with past description making use of a potpourri of research taken from a variety of fields, including second language acquisition and psychology as well as recent discoveries on mind-brain function from areas such as cognition and neurology (Rost, 2011). Part of the problem of describing the listening construct is the complexity of different processes and factors involved in L2 oral comprehension, making it almost impossible to provide a global, comprehensive definition (Aryadoust, 2013; Batty, 2015; Bloomfield et al., 2011; Wagner, 2002, 2004, 2013a).

Some definitions of listening ability have centred on a sub-skills approach and provide numerous taxonomies and lists of such skills (e.g., Buck & Tatsuoka, 1998; Munby, 1978; Richards, 1983; Weir, 1993). These lists have however been criticised as being essentially hypothetical in nature and having little empirical investigation of their veracity (Buck, 2001; Field, 2008a),[19] and although more recently some researchers have started to provide limited evidence for such sub-skills (e.g., Aryadoust, 2013; Goh & Aryadoust, 2015; Song, 2008), this line of research is still very much in its infancy. Research has also shown that different learners arrive at the correct answer in different ways (e.g., Buck, 1991, 1994; Shackleton, 2014) and use a number of skills and strategies

[19] They have, however, been seen as useful for syllabus design and teaching (Field, 2008a).

at the same time to solve test items (e.g., Goh, 2002; Vandergrift, 2003). Sub-skills have also been found difficult to operationalise; there is little consensus about the sub-skills necessary to answer a test item, or indeed what sub-skills an item actually tests (Alderson, 2000; Field, 2008a; Taylor & Garenpeyah, 2011). As Buck commented on his own study:

> Listening comprehension is a very individual and personal process – an active inferential process of constructing an interpretation which seems reasonable in the light of the listeners' own assessment of the situation, the listeners' background knowledge and the purpose for listening.
>
> Buck (1991, p.86)

Such a conglomeration of contextually-based processes makes the separation of sub-skills a practical impossibility; furthermore, lists of sub-skills tell us neither the relative importance of each skill nor how they should be sampled for test construction (Buck 2001).

A number of other models of listening comprehension have been provided (e.g., Anderson 2009; Bejar, Douglas, Jamieson, Nissan, & Turner, 2000; Buck, 2001; Field 2008a; Flowerdew & Miller, 2005) which generally describe listening as a cognitive process and normally include both cognitive and meta-cognitive strategy use—a definition which resonates with that presented by the CEFR. These cognitive processes take place online within the given contextual constraints of the audio input to be processed and, in the context of language testing, the task which needs to be carried out. As in any model of communicative language ability (CLA), such as that provided by Bachman and Palmer (1996, 2010), both ability and context of use must be taken into account. The listening process and contextual features of a language-use task will be examined in greater detail below.

3.3.1 The listening process

One important listening model in the context of second language assessment is that provided by Buck (2001, p.104),[20] which mirrors CLA views about language proficiency. Here, the listener uses both linguistic and non-linguistic knowledge in order to perform a task within given contextual parameters. Buck (2001) goes on to describe listening as an inferential, interactive process which uses bottom-up (speech perception and word recognition) and top-down processes (applying non-linguistic knowledge, schema, frames and background and topical knowledge) in a parallel form in order to decode a message and build meaning.[21] Similarly, Vandergrift (2003) defines the listening process as neither bottom-up nor top-down but an interactive, interpretive process in which listeners use both prior knowledge and linguistic knowledge in order to understand the message. Knowledge of the language, familiarity with the topic and the purpose for listening will dictate the degree of usage of bottom-up and top-down processes. These complex cognitive processes take place online while the listener constantly updates mental representations, using both cognitive and meta-cognitive strategies, in order to extract coherence and relevance from any given audio (Rost, 2011).

Listening then is very different from reading; speech is often unplanned and temporary, it requires the ability to perceive and segment the incoming stream of language and integrate information in real time; the listener cannot refer back to the text. It is often a controlled conscious activity and therefore taxing on the working memory (Baddeley, 2003). Furthermore, "lexical units are not necessarily as clearly marked as in written text; this lack of clarity of spoken language makes word segmentation an extremely difficult task for L2 listeners" (Staehr, 2009, p.582).

---

[20] See Flowerdew and Miller (2005) for a similar model which also includes affective factors.

[21] However, Wagner (2002), using different item types and exploratory factor analysis, did not support Buck´s model. Instead, he found the items loaded on explicitly and implicitly stated information (similar to Buck's (2001) listening default model).

Buck (2001) goes on to argue that as well as having a 'competence' perspective, the listening construct can also be defined in terms of TLU domain tasks, but this is more difficult if the TLU domain is broad. He concludes that a 'default listening construct' can be used to assess the skills unique to listening (Buck 2001, p.114). Here the listener must:

1) Process extended samples of realistic spoken language automatically and in real time.

2) Understand the linguistic information that is unequivocally included in the text.

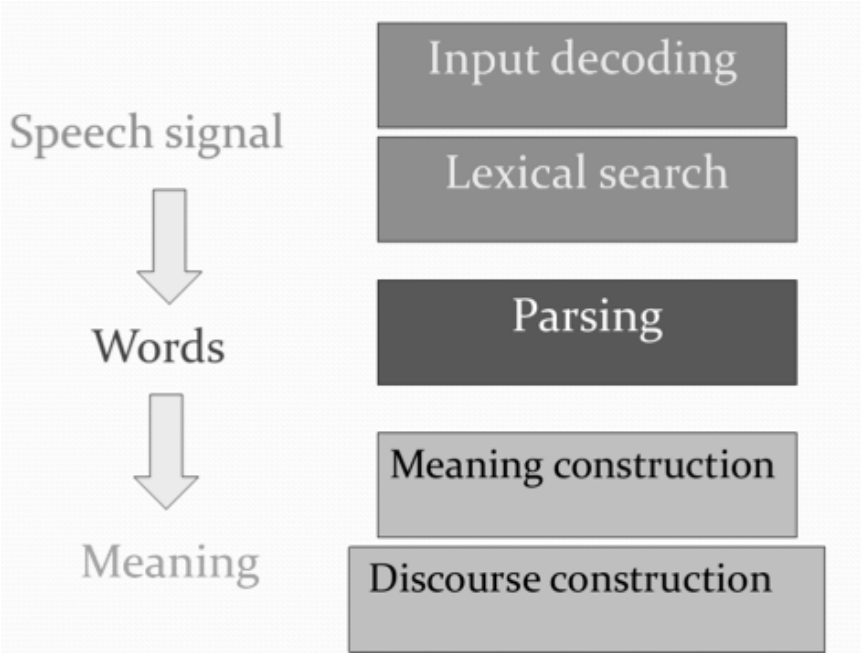
3) Make whatever inferences are unambiguously implicated by the content of the passage.

Vandergrift and Baker (2015) claim that Buck's (2001) listening default construct "is sufficiently flexible and broad to fit most contexts and to allow listeners to demonstrate their comprehension ability in real-life listening contexts" (p. 392).

An effective listener is therefore someone who is able to carry out parallel processing through the activation of different skills and knowledge (Vandergrift, 2007). This notion that "listening is primarily a cognitive activity" (Rost, 2011 p.57) is widely accepted and has been described as a multi-componential process. External sound waves need to be converted into auditory perception. Bottom-up linguistic processing is necessary to decode the sounds and intonation units, recognise words and activate lexical knowledge associated with these words, segments of speech need to be parsed in order to build syntactic representations. Top-down semantic processing permits the listener to link linguistic information with schemata (world knowledge and personal experience). Pragmatic processing includes both the integration of contextual clues and a comparison of speakers' meaning against listeners' expectations. Whilst processing a given auditory input, listeners draw on their linguistic, socio-linguistic and pragmatic knowledge.

Drawing on research into L1 listening, Field (2008a, 2012a, 2013a, 2013b) presents just such a process approach. He describes the process in five stages: aural input is (i) decoded, (ii) parsed, (iii) propositional meaning is established, (iv) a mental model is built, (v) a situation or discourse model is finally constructed (see Figure 5).

Figure 5. Field's (2008a/2013a) representation of the listening process.



This model essentially divides the listening process into lower-order processes (stages one to three) and higher-order processes (stages four and five). While this process involves the interaction of bottom-up and top-down processes, a linear approach is implied. This approach resembles that of Anderson's (2009) model,[22] which has already been successfully applied to research into L2 strategy use (e.g., Goh, 2000; Vandergrift, 2003). Field's approach has subsequently been incorporated into Weir's (2005b) socio-cognitive framework for listening test validation under the heading 'cognitive validity',

[22] This model breaks down the language comprehension process into three stages: perception, parsing and utilization.

which replaces the previous category of 'theory-based validity' (Taylor, 2013, p.28).[23] It has, however, been argued that cognitive validity is in fact the same as construct validity (see Harding, 2014a). Here, it is argued that cognitive validity (using a description of the listening process) is a description of the construct, the latent trait of exactly what kinds of knowledge, skills and abilities are necessary for successful listening. The listening process represented by Field's model will therefore be described in more detail here in order to present a description of the listening construct.

Lower-level linguistic processing involves the first three levels (input decoding, lexical search and parsing), they occur as part of input decoding or perception of the continuous speech. At this level, the listener must recognise the incoming sounds and essentially word segmentation skills must be relied upon drawing on linguistic knowledge (Vandergrift & Goh, 2012). Here auditory, phonetic and phonological mechanisms such as stress and intonation patterns have a role to play. This phoneme and word recognition stage has been found to pose major problems for low-level listeners (e.g., Goh, 2000) and becomes increasingly more automatic with the acquisition of phonological knowledge (Vandergrift & Goh, 2012). Recently, academics have proposed that students should be trained in such processes in the classroom and offer exercises which could be used for this purpose (e.g., Field 2008a, Vandergrift & Goh, 2012).

Once the perceived phonetic representation has been retained in the working memory, it is parsed to construct meaning. Listeners have to determine word boundaries and identify both content and function words in connected speech based on their lexical knowledge (Field, 2013a). The listener needs to identify the word and activate their vocabulary knowledge in order to recognise and recall the meaning of words. The listener tries to match sounds to lexicon stored in the memory using lemmas and lexemes, which give information about the properties and morpho-phonological form of a word. Here, it has been found that listeners are more successful at identifying content words rather than function words (e.g., Brown, 2008; Field, 2008b; Hulstijn, 2011; Shang, 2008;

[23] An earlier model (Taylor & Geranpayeh, 2011) includes separate categories for goal-setting, inference and monitoring comprehension.

VanPatten, 2004). Vocabulary therefore plays an important role in listening development and a number of studies have found correlations between vocabulary knowledge and listening proficiency (e.g., Bonk, 2000; Mathews & Cheng, 2015, Mercartty, 2000; Staehr, 2009). Indeed, it has been argued that lexical knowledge can compensate for uncertainties at the phoneme level (Field, 2008c). Similarly, Aryadoust (2015) points out that good lexico-grammatical knowledge helps the listener both directly and indirectly as it facilitates multi-tasking, that is, the listener does not have to rely on mental translation mechanisms.

During parsing, the utterance is segmented following syntactic structures and semantic clues in order to create a mental representation and give meaning (Vandergrift & Goh, 2012). However, even at this level, decoding is tentative and the listener is constantly forming and revising hypotheses (Field, 2008d). Inference therefore has a role to play at every level of the listening process and such bottom-up processes are continually being informed by top-down processes. Similarly, Buck (2001, p.148) points out that "inferencing is involved at all levels of language processing, even where information is explicitly stated."

Once parsed, the segmented utterances and words are transformed into a mental representation of the combined meaning of the words and a proposition or idea unit is formed using syntactic knowledge and group intonation boundaries. According to Buck (2001) such a proposition could simply be an adjective with a noun and "since storing a large number of propositions in memory is a tremendous burden, we make mental models of the content" (Buck, 2001, p. 28). These mental models are continuously updated and revised as more input is decoded, or as monitoring calls for the construction of a different mental model of the discourse.

During higher-level linguistic processing (meaning construction and discourse construction), the listener tries to understand groups of words by organising them into "familiar clusters corresponding to frequently encountered chunks of language" (Field, 2008a, p.113). Here a listener who has had more exposure to large amounts of language

input will learn that certain patterns and categories in the target language are more possible than others (Hulstijn, 2003), making processing easier, faster, and more accurate. The mental representation held in the memory is not a replica of the actual words in the text but is a representation of those words (Vandergrift & Goh, 2012).

The listener then starts to construct meaning by relating propositions to their own prior knowledge of the world or schemata, as the decoded information alone is not sufficient to give the complete meaning of the input (Field, 2013a). This type of processing draws heavily on the context and the listener draws on knowledge sources such as pragmatic knowledge and discourse knowledge stored in the long-term memory (Vandergrift & Goh, 2012). The listener now has the ability to construct meaning, taking what has been understood and interpreting it in relation both to what has previously been said and to speakers' attitudes and intentions in combination with prior world knowledge or schemata (Field, 2013a). Prior knowledge sources therefore also include pragmatic knowledge and socio-linguistic knowledge as well as linguistic knowledge. Rost (2011) argues that in order to test listening we must replicate real-life, communicative situations and that an utterance must be interpreted in its specific context. That is, the meaning is shaped by the context and the listener constructs the meaning through interpretation. Indeed, as Buck puts it "meaning is not something in the text that the listener has to extract, but is constructed by the listener in an active process of inferencing and hypothesis building" (2001, p.29).

During discourse construction, information is linked together and main and minor points are identified enabling the listener to report a line of argument (Field, 2013a). According to Field (2008a, p.119) "the more ideas there are in a short space of time and the more intricate the links between the ideas, the greater the demands made upon the listener". Discourse construction is therefore a semantic representation of the inter-related propositions or idea units, including a pragmatic interpretation of the input. Listeners also have to rely on the co-text and apply contextual and semantic knowledge to the propositions or use inference to decipher meaning that has not been explicitly stated. During discourse construction, the listener uses processes of selection to assess the

relevance of the information, integration to add and integrate information and develop the discourse, self-monitoring to check that new information is consistent with what has previously been said, and structure building to prioritise and organise information according to its importance and relevance (Field, 2013a). In this way, the listener is able to re-construct the macro-structure of a given audio input.

We can see then that the listening process cannot be broken down into a linear succession of cognitive processes. It is an interactive process and successful listeners must use both lower and higher levels of processing. A good test should target all levels of listening in order to provide a complete picture of a candidate's listening proficiency, although lower ability test takers may focus much of their working memory on word level decoding and so not be able to generate wider meaning and therefore in practice should be asked only about factual information (Field, 2013b, 2017).

3.3.2 Purpose for listening

In the same way that context plays an enormous role in comprehension, it is arguable that the purpose for listening affects how a listener approaches a given task, as we listen in different ways depending on the information we wish to extract and act upon (Vandergrift & Goh, 2012). In a testing situation, an audio file is accompanied by a set of comprehension questions, which the test-taker has to answer by extracting information from the audio file. This is important, as noted by Field (2017), who states "in teaching or in testing, the only way we can establish if 'comprehension' has taken place is to ask some kind of question". Indeed, *response* or *utilisation* form an important part of some listening models (e.g., Anderson, 2009; Bejar et al., 2000).

Vandergrift and Goh (2012) also highlight the fact that different texts will automatically require more or less of a discourse model. They give the example of a safety message on an airplane compared to a song, the former requires the extraction of specific factual information and the latter is open to a much wider range of pragmatic interpretations. This will affect the degree to which listeners use one process more than

another and a test item requiring specific information, such as a price, may well engage the listener in more lower-level bottom-up processing. Drawing on Urquhart and Weir's (1998) account of reading skills, Field (2008a, p.66) proposes a tentative model of listening types, which are determined by listeners' goals. Different types of listening will arise depending on the purpose for listening and a competent listener can select the type of listening appropriate to the input and task (Field, 2008a). The amount of information which needs to be extracted will therefore govern the amount of spoken input which is processed. A clear distinction is made here between local and global understanding. In this study I will argue that communicative purpose is paramount for a CEFR-based test and that in order to have construct coverage different types of listening must be evaluated if we do not want to represent the trait in an overly narrow fashion.

Different types of listening are discussed in Vandergrift and Goh (2012) in the context of teaching the listening skill. They state that "listening tasks should also offer opportunities to develop core skills … The skills used to achieve comprehension are mainly influenced by the purpose for listening." (p. 168). These core skills are presented in Table 1, and teachers are encouraged to develop them in the classroom. In order to have construct coverage in a testing situation, as many of these skills as possible should be included on a test. I would also argue that by doing so we would be adhering to principles of assessment for learning (e.g., Hamp-Lyons & Tavares, 2011) by focusing on communication goals and consequently positive washback would be supported.[24] In this way, teachers would thus be further encouraged to include skills and strategy training as part of classroom practice. For example, listening purpose can lead to 'selective' listening, whereby some aspects of the input scaffold understanding of a text (Graham & Macaro, 2008). Indeed, studies have found that 'key content words' can perform this function (e.g., Brown, 2008; Field 2008b).

[24] Indeed Green (2015) points out that in this context if teachers believe that all parts of any curriculum could appear on the test we remove the possibility of teaching to the test.

Table 1. Core Skills for Listening Comprehension (taken from Vandergrift & Goh, 2012 p.169)

Listen for Details
Understand and identify specific information in a text: for example, key words, numbers, and names.
Listen for Global Understanding
Understand the general idea in a text: for example, the theme, the topic, and the overall view of the speaker.
Listen for Main Ideas
Understand the key points or propositions in a text: for example, points in support of an argument, or parts of an explanation.
Listen and Infer
Demonstrate understanding by filling in information that is omitted, unclear, or ambiguous, and make connections with prior knowledge by "listening between the lines": for example, using visual clues to gauge the speaker's feelings.
Listen and Predict
Anticipate what the speaker is going to say before and during listening: for example, use knowledge of the context of an interaction to draw a conclusion about the speaker's intention before he/she expresses it.
Listen Selectively
Pay attention to particular parts of a message and skim over or ignore other parts in order to achieve a specific listening goal or, for example, when experiencing informational overload, listen for a part of the text to get the specific information that is needed.

Strategic competence is also clearly an important component of any model of CLA, is included in the CEFR proficiency scales, and is therefore important to the construct definition. The listening process involves both cognitive and meta-cognitive strategy use, which will be triggered in response to listening purpose. Vandergrift and Goh (2012) argue that skills are automated but strategies are controlled and require effort and are activated in order to compensate for difficulties in understanding. There is in fact much debate about strategy use and I will now go on to present this literature in an attempt to define the strategies relevant to the present study.

3.3.3 Strategy use in listening

Much recent research has focused on strategy use in order to propose models of listening comprehension and cognition, with varying degrees of success and agreement. In this section, I will attempt to clarify a definition of the strategies necessary to the present study and will briefly examine some of the more salient findings. Any understanding of the listening process must also take into account the differences between L1 and L2 listeners. L1 listeners are experts and have highly automated decoding routines which are accurate, rapid and effortless; it is an automated process (Buck, 2001). When bottom–up processing is accurate and automatic, it frees working memory capacity and thus allows the listener to build complex meaning representations. The L2 listener, especially at lower ability levels, will not have the same highly automated routines as L1 listeners; the process will be controlled and will necessarily use a range of compensatory strategies to aid comprehension. Listeners have to draw on L2 linguistic knowledge and adapt their L1 processes to the understanding of the L2 language (Field, 2008a).

As previously stated, Vandergrift and Goh (2012) make a distinction between skills and strategies. We have seen that listening ability has been described as a cognitive process which involves both lower and higher level cognitive processing skills. Field (2008a) restricts the term processes to the cognitive operations which underlie all listening, whether in L1 or L2, and the term strategies to compensatory techniques that are used to fill gaps in word recognition or in understanding. Similarly the CEFR states:

> Skills that are an inevitable part of the process of understanding or articulating the spoken and written word (e.g. chunking a stream of sound in order to decode it into a string of words carrying propositional meaning) are treated as lower-level skills, in relation to the appropriate communicative process.
>
> (CoE, 2001, p.57)

That is to say, skills are related to the listening process described in the previous section and "communication strategies can be seen as the application of the metacognitive principles: *Pre-planning*, *Execution*, *Monitoring*, and *Repair Action* to the different kinds of communicative activity" (CoE, 2001, p.57).

Any validity study must distinguish between construct-relevant and construct-irrelevant strategies (Cohen, 2012, 2013). However, terminology applied to strategy use has been debated (see Cohen, 2005, 2007a, 2011, for discussion on disagreements between experts), these debates have been referred to as 'terminological fuzziness' (Field, 2008c). Cohen (1998) defines strategies as conscious acts which are accessible for description, implying an element of selection. Here, processes are considered to be unconscious and automatic. Following this view, some strategies may become processes as they become automated (see also Saville-Troike, 2005). In any study of the L2 listening process therefore, we should see a continuum in which many processes will not yet have been automated, depending on the L2 listener's proficiency level.

This is a definition is endorsed by many (e.g., Cohen 2011; Goh, 2002; Oxford, 2011). Similarly, Macaro (2006, p.328) argues that strategies are conscious mental activities used to attain a goal within a learning situation and are transferable to other situations or tasks.[25] They are therefore useful for providing construct validity evidence for one particular task or test version with the possibility of making generalisations about a specific test taker to other similar TLU domains. This definition resembles the description given for strategies by the CEFR itself: "A strategy is any organised, purposeful and regulated line of action chosen by an individual to carry out a task... with which he or she is confronted" (CoE, 2001, p.9).

The CEFR (CoE, 2001, p.72) goes on to give a description of receptive strategies that mirrors interactive process views of listening and includes identifying the context, activating appropriate schemata and building meaning using linguistic and non-linguistic

[25] He was, however, referring to language-learning strategies.

cues, inference and hypothesis testing. In the action-orientated approach, appropriate strategies for the given task must be activated (CoE, 2001, p.9). This lends further support to the argument that task-specific behaviours are construct-relevant as an interactionalist view of language ability in person in use (Chalhoub-Deville, 2003).

The present study will follow this definition and consider the skill of listening to be a language use activity which is the performance on specific tasks (Bachman & Palmer, 2010). This model includes all the processes and skills outlined above, which include the inferential processes needed to build mental representations of an input text and language use task, as well as meta-cognitive strategies used to carry out a communicative task. According to Bachman and Palmer's (1996) model of CLA, meta-cognitive strategies are mental activities which perform an executive management function and control cognitive strategies. This model of strategic competence has been validated by Phakiti (2008) using structural equation modelling, and by Zhang, Goh and Kunnan (2014) using questionnaires and test performance data, although both these studies were based on reading and not listening. Here, it was observed that meta-cognitive strategy use impacts the cognitive strategy use, which has a direct effect on successful reading performance. Similarly, studies have shown the importance of meta-cognitive strategies for listening.[26] Vandergrift, Goh, Mareschal, and Tafaghodtari (2006) using the Metacognitive Awareness Listening Questionnaire (MALQ),[27] found that approximately 13 % of variance in listening achievement could be explained by the use of meta-cognitive strategies. This finding was supported in Vandergrift and Baker's (2015) study of learner variables, which also demonstrated a link between successful listening and metacognition. Similarly, Goh and Hu (2014) discovered that meta-cognitive awareness had a positive influence on listening performance, accounting for 22% of the variance. Indeed, meta-cognitive instruction has been shown to benefit listeners, leading to increased proficiency (e.g., Vandergrift & Tafaghodtari 2010). In a review of research

[26] Also see Oxford (2017) who outlines metacognitive, meta-affective and meta-sociocultural strategies under the broad heading of 'metastrategies', (though she is discussing language learning strategies).
[27] This is a validated questionnaire, which has been further validated by Ehrich and Henderson (2018).

into listening strategies, Macaro, Graham and Vanderplank (2007) identified the following meta-cognitive strategies important for the listening process:

1. Making predictions about the content before listening. Buck (2001, p.104) calls this process 'assessing the situation', a part of the test-taking process felt to be important because the items not only provide the candidate with a purpose for listening but also a context from which to activate schemata and generate hypotheses (Shohamy & Inbar, 1991). Such predictions are believed to reduce the cognitive load because the number of possible propositions becomes more limited (Graham & Macaro, 2008). Here, listening selectively has a role to play, as by listening for certain words and phrases, predictions can be confirmed or rejected.

2. Monitoring and evaluating comprehension. Successful listeners continuously check and update their comprehension (see for example Goh, 2002).

3. Inference. The use of linguistic and non-linguistic clues to infer meaning and compensate for lack of knowledge. As previously stated, inference can occur at any level of language processing and is "at the core of language processing" (Buck, 2001, p.147).

The listening process is therefore governed by meta-cognitive strategy use and in terms of having a valid listening test, the knowledge, skills and strategies used to solve test items should be construct relevant and reflect the above description of the listening construct. Cohen (1998, 2007b, 2011, 2013) makes a distinction between test-management strategies (for example, timing, instructions, reading questions first) and test-wiseness strategies (for example, knowledge of the world).[28] The latter are not necessarily determined by proficiency; test takers answer questions without using linguistic knowledge and cognitive processes related to the construct (Cohen, 2012, 2013). These strategies thus threaten the validity of the test and it is therefore arguable

[28] See Cohen (2013) for a description of such construct irrelevant strategies.

that investigation concerning strategy use is imperative in any validity study of a listening test.

3.3.3.1 Previous studies involving listening strategies

Listening strategy use research has mainly been undertaken in the context of SLA, where studies have shown that strategy use varies depending on language proficiency and task demands (e.g., Graham, Santos & Vanderplank, 2008, 2010; Vandergrift, 1997, 2003). However, caution is recommended over claims as to which strategies lower- and higher-ability listeners may use, as the different studies carried out make use of different proficiency measures and are therefore not comparable (Macaro et al., 2007).[29]

Some studies have shown that lower level candidates had difficulties answering items based on global questions requiring top-down processing (e.g., Hansen & Jensen, 1994; Osada, 2001; Shohamy & Inbar,1991).[30] For example, Shohamy and Inbar (1991) found that high-level learners were much better able to synthesise information, draw conclusions and make inferences. Similarly, Graham et al. (2008) found that lower abilities did not evaluate comprehension using contextual knowledge. Yet, Tsui and Fullilove (1998) found that lower-level listeners focused on background knowledge as they had problems with decoding, which led them to impose their own incorrect understanding of what was heard. Their study concluded that bottom-up processing is more important than top-down for successful listening. Similarly, other research has shown that more proficient listeners use more meta-cognitive strategies (e.g., Goh, 2000; Nguyen 2008; Vandergrift, 1997).

[29] In this context studies have found that strategy training in the classroom has positive effects on learners listening ability (e.g., Graham & Macaro, 2008; Goh & Taib, 2006). Indeed, in the context of SLA there is a widely held belief that by teaching listening strategies we are actually teaching learners how to listen (e.g., Seigel, 2011).

[30] This supports Field's (2008a/2013a) model.

Recent evidence suggests that successful listeners use a range of strategies which are selected and used in response to task demands (e.g., Graham & Macaro, 2008; Vandergrift, 2003, 2007). Indeed, Chamot (2005) argues that previous research has "confirmed that the good language learners are skilled at matching strategies to the task they were working on, whereas less successful language learners apparently do not have the metacognitive knowledge about task requirements needed to select appropriate strategies" (p. 116).

Nevertheless, listening processes and strategies remain a somewhat unexplored field of study as a method of test validation and most studies to date have taken place in non-test situations. Indeed, Cohen (2007b) notes that strategy data, particularly in validation research, are not usually collected in actual high-stakes testing situations. It may be that the strategies actually used in responding to tests in high-stakes settings differ from those identified under research conditions. The following presents the main studies which have been carried out on listening strategy use for the purpose of language test validation.

Buck (1991,1994) carried out studies to discover how test takers arrive at answers on a listening test using retrospective interviews and verbal reports. He found that the participants employed a selection of skills and strategies and concluded that each combination of listener, text and question resulted in very different individual processes to answer items correctly. His results therefore support the view that task specific behaviour should be included as part of a test construct.

Yi'an (1998) investigated an MCQ test of listening using retrospective reports and found that both linguistic and non-linguistic knowledge was used to answer the questions. It was found that non-linguistic knowledge used to compensate for lack of linguistic knowledge often resulted in an incorrect answer. It was also found that MCQ items allowed for uniformed guessing. Wu (1998), using verbal reports, investigated an MCQ listening test in a Chinese context and supported these findings showing that the MCQ format allowed for guessing which is informed by information other than that in the audio file. Yang (2000), using verbal reports and expert judgement, found that 48% to 64% of

the items from the listening and reading subtests of the old TOEFL Practice Test B could have been answered by using test-wiseness techniques by using cues such as absurd options, similar options, and opposite options. This serves as an important warning to item writers and highlights the well-documented difficulties in writing good MCQ items with plausible distractors (Haladyna, 2004).

In an attempt to identify the sub-skills and strategies used to answer both an MCQ and table completion listening test, Barta (2010) presented the results as a taxonomy consistent with theoretical models of CLA for listening, such as that provided by Buck (2001). However, it was also reported that the test did not elicit pragmatic knowledge, which was explained by the fact that B1/B2 level tasks are explicit in nature and the function or illocutionary force of discourse from simple domains is difficult to measure. Furthermore, guessing was also evidenced on the table completion task.

However, in my own study investigating the cognitive validity of a CEFR B1-related test (Shackleton, 2014), verbal reports showed that B1-level listeners did use pragmatic knowledge when answering test items. Here, following Buck (2001), coding of verbal reports included the category of 'cognitive environment'. This category included top down processing and strategy use such as pragmatic and sociolinguistic knowledge, drawing on co-text and context of situation by using prior knowledge and inference. The respondents used both linguistic and non-linguistic clues to solve test items and the data matched a process view of listening. In this study, the highest scoring participant directly reported a text representation of what had been heard, showing a more automated ability and consequently a better ability to perform the CEFR B1 'can dos' on which the test was based. Similarly, the lowest scoring participant gradually built meaning from smaller units and often reverted to a top-down approach to fill in the gaps. This participant was sometimes unable to build meaning and answered items incorrectly. As such, scores on the test reflected the ability to perform the CEFR 'can do' descriptors. It was seen that a certain amount of correct decoding had to take place before enriched understanding could be reached using the 'cognitive environment'. Likewise, no test-wiseness strategies were reported and all instances of complete random guessing led to the choice of an incorrect

answer. In sum, the study provided good construct validity evidence for the test under investigation.

Nguyen (2008) investigated the construct validity of the listening section of the TOEFL IBT and IELTS tests in a Vietnamese context. Using verbal protocol and retrospective questionnaire methodology, it was found that both tests elicited similar strategy use and minimal test taking strategies were used, giving evidence of construct validity. It was also found that the relationship between strategy use and test scores was similar on both tests and effective strategy use had a high correlation with test scores. However, it should be pointed out that most of the evidence presented is taken from the strategy questionnaire data rather than the verbal reports. Similarly, Cohen (2007a) cites the example of Douglas and Hegelhiemer's (2005) validity study of the TOEFL IBT listening test, which uses the *Morae* software package to analyse verbal report data with follow-up interviews –though he does point out that the sheer extent of the data collected can be complex and difficult to interpret. The results showed that both strategy use and sources of knowledge were similar to those used in the non-test situation, thereby supporting the construct validity of the test.

In a cognitive validity study of the lecture part of the IELTS listening paper, Field (2012b) added the category of test specific behaviour [31]– which he considered to be construct- irrelevant – to his categories of L1 processes and compensatory strategies. Using retrospective verbal reports to compare the test situation to the non-test situation of note-taking, he found that similar strategies were not used in both conditions, which led him to question the test's construct validity.

Results from Badger and Yan's (2012) think-aloud study showed both experts and L2 Chinese speakers of English used similar strategies to complete test tasks on the IELTS listening test, thereby providing evidence of construct validity. Both groups, however, used test-taking strategies or task-specific problem solving strategies. If, however, we are

[31] Field describes these behaviours as an ability to exploit the written information in the items.

modelling the skill on expert behaviour, as proposed by Field (2012a), the fact that experts use problem-solving strategies just like non-experts do means we would need to question whether or not all of Field's (2012b) test specific behaviours are construct-irrelevant. In a validity study, it is precisely here that terminological differences become fundamental and we need to be completely clear about just what elements we consider to be construct irrelevant. In both Field's study (2012b) and Badger and Yan's study (2012), the TLU situation is specifically described as listening to an academic lecture. A more general CEFR-based study with a wider construct definition, however, contains a much broader range of possible TLU communicative situations to be sampled. Besides, because task-specific behaviour is based on purpose for listening, it is construct-relevant.

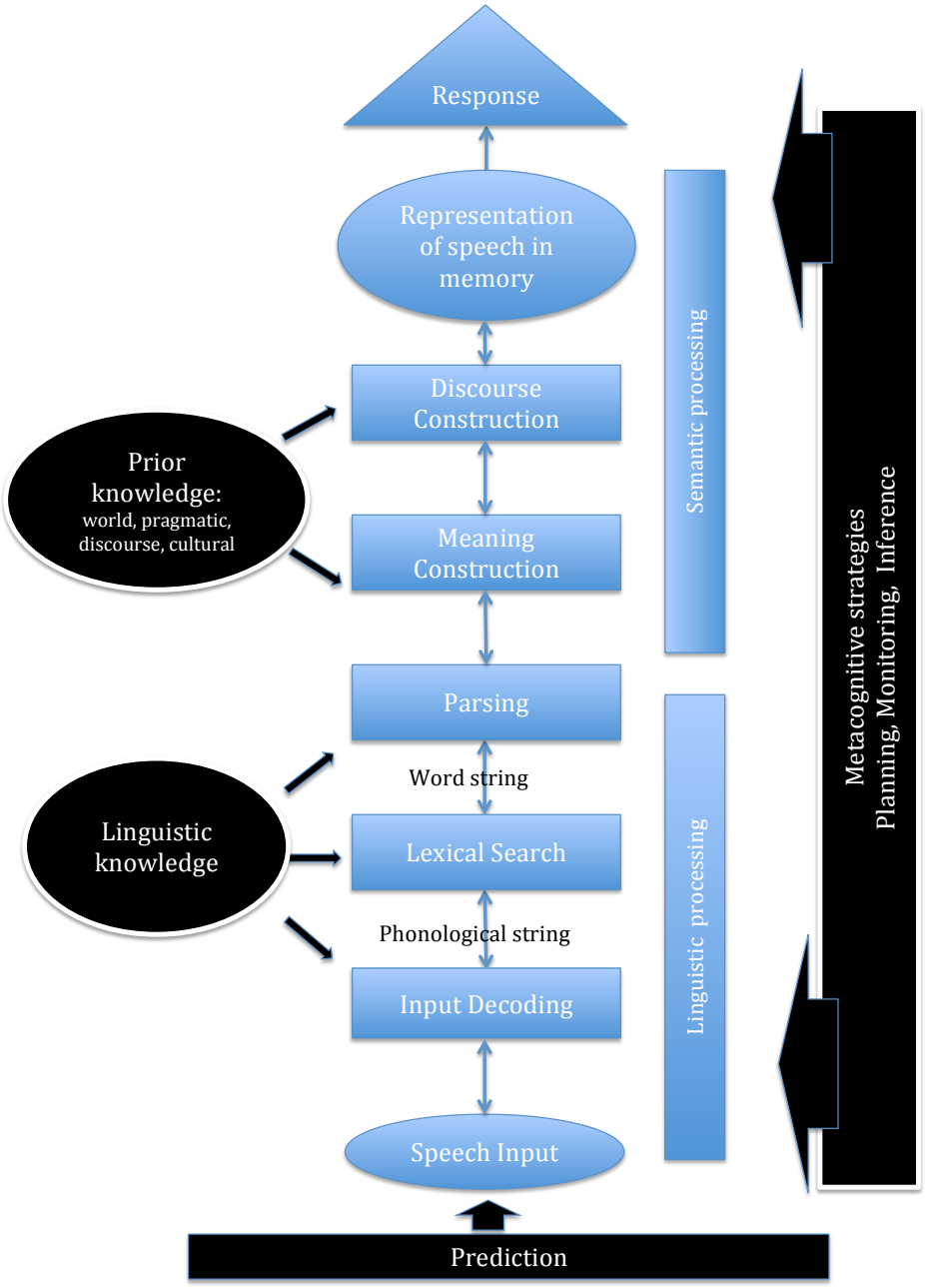
3.3.4 Summary and presentation of the proposed BFE CEFR B2 listening ability model

Listening is therefore a complex interactive process which reflects the ability to understand authentic discourse in context. Higher-level processes involve the listener relying on previous knowledge of the topic, the co-text, plus pragmatic and socio-linguistic knowledge (Field, 2008a; Rost, 2011). If a test is to be shown to be valid it must engage the correct cognitive processes. As a CEFR-related test, the notion of listening ability in this study is firmly grounded in a communicative competence model of language ability. The KSAs outlined above will be adhered to and the model draws largely on Field's listening process model. What is more, a communicative competence model of listening ability should include meta-cognitive strategies.[32] In Chapelle's (1998) interactionalist view, meta-cognitive strategies are responsible for mediating between trait and context, task specific behaviours are context relevant. Similarly, Field (2008a) argues that listeners need to develop 'real world strategies', the strategy which is relevant to the situation needs to be applied, and here distinctions between cognitive and meta-cognitive strategies become ambiguous. The purpose for listening should dictate the choice of strategy use and it is not the same to be listening for a piece of local specific information as it is to listen for global meaning. In terms of strategies, the CEFR

[32] In this respect, my model is similar to that presented by Taylor & Geranpayeh (2011).

specifically states in the 'IDENTIFYING CUES AND INFERRING (Spoken & Written)' scale (p.72) that a B2 student "can use a variety of strategies to achieve comprehension, including listening for main points; checking comprehension by using contextual clues."

Figure 6. Proposed model of listening ability for BFE CEFR B2 listening test (based on Field, 2008a, 2013a)



Following the substantive theory outlined above, Figure 6 shows a representation of the ability model to be used in the present study.

3.4 Language use listening test task characteristics

An interactionalist approach to the construct definition must take into account the language use task. Here, a balance must be struck between our listening ability model and the context of use. Only by introducing interactionally authentic tasks can it be demonstrated that similar cognitive processes are reproduced in the test situation as those in the TLU domain. The authenticity of a test task is important since that authenticity provides a critical link between the test results and the desired outcome of scoring interpretations—the target language use situation (Bachman & Palmer, 1996). Thus, more authentic test tasks lead to more valid score interpretations. The tasks in a test represent the content of that test, and content relevance and representativeness provide important types of evidence in support of a validation argument (Bachman, 2002). For example, in Weir's (2005b) socio-cognitive approach, 'cognitive validity' can be compared to what Bachman and Palmer (1996) have termed as 'interactional authenticity'. Similarly, according to Weir (2005b), a test must have 'context validity' relating to the performance conditions of the social arena in which an activity is undertaken. To this end, Bachman and Palmer (1996, 2010) provide a framework of test task characteristics in which test tasks must represent the larger TLU domain and a test must have 'situational authenticity'. Likewise, Kane's argument-based approach requires that test content is representative and relevant to the domain being tested in order to provide evidence for the extrapolation inference. Language processing does not take place in a vacuum and the communicative purpose for listening has a critical role to play (Vandergrift & Goh, 2009). Certainly, if we want our test to relate to the CEFR we must consider the social dimension and recognise that language use takes place for a communicative purpose (Vandergrift & Goh, 2012).

The context of use should therefore govern the appropriateness of the test task, both in terms of linguistic or content demands of the audio file to be processed and the features of the task setting (Weir, 2005b). These are both aspects of a test which would normally be found in test specifications for item writers. According to Weir (2005b), evidence for 'context validity' should be a priori, a part of the test development process. Yet reported studies usually restrict content analysis to a posteriori surveys of expert judgement (e.g., Pardo-Ballester, 2010). In a discussion of the context validity of the Cambridge main suite tests, Elliot and Wilson (2013) simply provide lists of corresponding content found in these proficiency tests. Kane (2006), however, argues that such content-related evidence from the test developers themselves could have 'a confirmationist bias'.

Contextual features of the test tasks play an important role in determining task difficulty and such decisions are therefore not only important for detailed test development guidelines but also for comparability between test administrations. Test content for a CEFR-based test is governed by a description of domain and ability in the form of performance descriptors which provide the test developer with a description of communicative listening activities to be expected at each proficiency level and so should be adhered to at the task development stage. These descriptions are, however, extremely general and not enough to provide a basis for test development. Alderson et al. (2004, 2006) provide much more detail for a CEFR-based test in the form of a grid, which is included as a document for linking tests to the CEFR (CoE, 2009). The test task represents the context of use and decisions need to be made about contextual parameters related to both the audio input texts and the task in order to provide detailed test specifications outlining the test tasks which resemble the relevant TLU domain. Weir's (2005b) socio-cognitive framework divides contextual parameters into: 'task setting' (e.g., response mode and channel), 'linguistic demands' (e.g., grammatical and lexical resources) and 'speakers' (e.g., accent and speech rate), as well as 'administrative setting'.[33]

[33] See Elliot and Wilson (2013, p.152-241) for detailed discussion in relation to the Cambridge Main Suite exams.

3.4.1 Characteristics of input passage

Yanawanga (2012, p.61), building on Bachman and Palmer's (1996, 2010) task characteristics framework, provides an exhaustive framework of contextual parameters for L2 listening tests. Here, I will examine the debates about those variables which I feel are pertinent to decisions which need to be made in the present study in order to develop my own CEFR-related test specifications based on a sound theoretical description. These variables relate to input passage characteristics and test task conditions.[34]

3.4.1.1 Authenticity

Having previously mentioned general concerns about the authenticity of test tasks (situational and interactional), another important debate for listening tests is the authenticity of the sound-file to be processed. The degree of authenticity of the oral input of test tasks would normally be included in test specifications for communicative language tests (see Alderson et al*.,* 2004, 2006). Many language tests rely on scripted, revised and edited sound files produced in a studio by actors. Consequently, there are many calls in the literature for a move towards more authentic input texts for both teaching and assessment purposes (e.g., Field, 2008a, 2013a; Gilmore 2007, 2011; Vandergrift & Goh, 2012; Wagner, 2013a), as these polished texts are very different to unscripted spontaneous spoken discourse that occurs in most real world communicative situations. Wagner (2013a) argues for the incorporation of all the linguistic characteristics of unplanned spoken discourse in order to better represent the TLU situation. In reality, there exists a continuum of aurality (Shohamy & Inbar, 1991; Vandergrift & Goh, 2012), from a planned talk to a spontaneous conversation. A lecture, for example, would often be guided by a powerpoint presentation or the lecturer's notes or handouts, and an audio-guide would probably be scripted in order to convey the important information. The TLU situation would therefore need to guide the types of spoken discourse we wish to test. In this respect, Shohamy and Inbar (1991, p.37) conclude that listening tests should reflect

[34] See also Bloomfield *et al*. (2011) for a detailed review which also includes listener characteristics.

the range of genres. Buck (2001) argues that the better the task replicates the TLU domain, the better it will inform us about performance in that domain, adding that when the TLU situation is explicitly stated it may be better to define our construct in terms of test tasks. It is here that the CEFR descriptors give test developers a basis from which to draw authentic input texts. Indeed, Taylor (2013) acknowledges the contribution of the descriptors as regards listening context and input. Similarly, O'Sullivan (2008) reported that by including CEFR descriptors in the test specifications, a 'substantial leap forward' was made in the professionalisation of the assessment practices of the City and Guilds test. The 'can dos' specifically describe the types of communicative listening activity to be expected at each proficiency level and should be adhered to at the task development stage. In terms of CEFR B2 listening proficiency, we are given a range of contexts and domains (see CEFR p.48 and 49), which should therefore be well sampled and included in a relevant assessment tool if we do not want construct under-representation.

An input text's characteristics are of the utmost importance for a listening test as we must aim to test those aspects which are unique to listening (Buck, 2001). These include the ability to process online, connected speech, which is very different from the written word. Here, we may find all the linguistic characteristics of typical spoken language: hesitations, false starts, stress and intonation patterns, redundancy, lack of complete sentences and speaker overlap in dialogues as well as other phonological characteristics such as assimilation, vowel reduction and Sandi-variations such as weak forms and elision. Spoken discourse does not follow the same rules as written discourse and can include grammatical mistakes, shorter idea units, and ellipses; also, as it is unplanned it is less logically organised (Wagner, 2013a). A scripted text, in contrast, would lack many of these characteristics (Field, 2008a, 2017). Indeed, in a comparison of a description of the same event, Vandergrift and Goh (2012, p.152) showed that a scripted text lacked many of the features of natural language contained in an unscripted text. Field (2013b) argues that scripted and even semi-authentic recordings bear little resemblance to natural language. He points out that actors mark commas and full stops, there are no hesitations or false starts and voices do not overlap. Also, test developers sometimes put in scripted distractors – making the recording much more informationally dense than a natural piece

of speech would be, which can place too great a strain on the working memory (Field, 2013a).

There is, in fact, dispute over the extent to which scripted texts simplify comprehension and research suggests that such scripted audios are not, in fact, easier to understand. Similarly, the CEFR (CoE, 2001 p.165) states that "syntactic over-simplification of authentic texts, however, may actually have the effect of increasing the level of difficulty (because of the elimination of redundancies, clues to meaning etc.)" This concurs with earlier findings that aural texts incorporating unscripted dialogue were easier to understand (Shohamy & Inbar, 1991). Yet Read (2002) found scripted monologues were easier than unscripted discussion of same content. Brindley and Slatyer (2002) did not support either of these studies, drawing no significant conclusions. A study by Yanagawa (2012) which employed a common item-common person design using Rasch analysis looked at task difficulty when input included more or less Sandi variation. It was found that there was no significant difference in difficulty between the two test forms. It was even shown that some of the items were in fact easier when the input text was presented in a more natural form. Similar results were reported by both Kostin (2004) and Brunfaut and Révész (2013), although both these studies used scripted input texts performed by actors. Similarly, authentic spoken discourse uses stress and intonation patterns to give clues to meaning, for example words which represent core meaning get stressed and intonation can indicate clausal boundaries (Buck, 2001).

Papageorgiou, Stevens and Goodwin (2012) compared test performances on both monologues and dialogues with identical content and vocabulary using a Rasch analysis of test results and a content analysis. Their findings partially support the idea that dialogues are easier to understand. However, it should be noted here that both test forms were scripted and other variables such as speed of delivery may not have been representative of the TLU. Indeed, faster speech is thought to be more difficult to understand (e.g., Buck, 2001, p.38). The CEFR also recognises speech rate as a factor which effects task difficulty. However, research has shown that there is not necessarily any correlation between the difficulty of a text and the speed of speech delivery. Derwing

and Munro (2001) found that participants preferred the natural rates over slowed down ones. Similarly, Brunfaut and Révész (2013) found that speech rate had no significant effect on task difficulty. However, Zhao (1997) found that by allowing listeners to adjust speech rate they performed better when they chose to slow down the audio input. Contrasting results have been explained by the fact that it is difficult to study speech rate separately from other speech features. Brindley and Slatyer (2002), for example, concluded that it is extremely difficult to isolate speech rate because in a listening task there are a variety of complex factors at play. Rubin (1994) states that it is difficult to compare evidence from different studies due to varying rates of normal speech; the rate for native speakers of English is anywhere between 165 to 180 words per minute (wpm). Yet conversations and lectures will vary in speed and again we could argue that if authentic texts are used in a listening test they should be delivered at the rate which would be found in the TLU domain. I would agree with the conclusions of Chapelle, Enright and Jamieson (2008) who experimented with speech rates but concluded that they would not use speed of delivery to manipulate task difficulty on the TOEFL test as in order to make a difference in difficulty speech rates would have to be adjusted so much as to make them sound unnatural. It was also pointed out that speech rate is not constant and varies within a conversation or lecture, often as a clue to meaning. Furthermore, one explanation for the differing results of these studies could be because the biggest factor contributing to real life slower speech is, in fact, more pauses, rather than slower articulation (Field, 2017).

In a 10 month longitudinal classroom based study, Gilmore (2011) compared the development of communicative competence in students learning from input with authentic spoken discourse with students using pedagogically designed input; he found that learners who were given authentic input outperformed those who were not. Indeed, the fact that learners are not exposed to authentic discourse in the classroom has been suggested as one of the reasons why L2 listeners have such difficulty in understanding authentic texts (e.g., Flowerdew & Miller, 2005; Gilmore, 2007; Wagner, 2013a). Similarly, Wagner and Toth (2014) compared performance on scripted and unscripted Spanish listening texts and found that the scripted texts were significantly easier. The

contradictory results, along with completely different methodologies for creating the authentic versus scripted audios means we cannot draw definite conclusions.

Furthermore, I would argue that task difficulty is not the issue here, it is an issue of enhancing construct representation and positive washback. Indeed, it has been argued that the complexity of the sound-file to be processed is not the main indicator of task difficulty, but rather the demands of the task (Field, 2008a). Field argues that the same sound-file could easily be presented to listeners at multiple proficiency levels if they were asked to extract different information. The suggestion seems to be that as long as the input file is authentic it should necessarily be representative of the TLU domain, and it is the task demands (i.e., the information which needs to be extracted) which would need to be representative of the kind of operations necessary at each CEFR proficiency level. Certainly, if we are to represent our TLU, a range of input types is recommended which would include monologues, dialogues and, for a university entrance test, lecture type presentations.

The use of authentic input in a test could lead to the use of authentic texts in the classroom, thus developing learners' ability to understand real world connected speech, which is after all the goal of listening instruction. For this reason at present there is a great interest in the SLA literature in using authentic materials in the classroom in order to teach and test listening skills (e.g., Field, 2008a; Rost, 2011, Vandergrift & Goh, 2012). As Field states:

> A switch from scripted to unscripted has to take place at some point, and may, in fact, prove to be more of a shock when a teacher postpones exposure to authentic speech until later on. It may then prove more not less difficult for learners to adjust, since they will have constructed well-practiced listening routines for dealing with scripted and/or graded material, which may have become entrenched.
>
> Field (2008a, p.281)

A move needs to be made by test developers to address this issue and plan for positive washback on language learning and teaching. It has also been pointed out that authentic materials are intrinsically interesting, motivating and can be found in many TLU domains (Vandergrift & Goh, 2012).

Similarly, Field (2013a, p.143) states that "if a test is to adequately predict how test takers will perform in normal circumstances, it is clearly desirable that the spoken input should closely resemble that of real-life conversational or broadcast sources". Field concludes that the test development process for the Cambridge main suite exams threatens construct validity because of its over-reliance on the written transcript, recommending that Cambridge test developers should have a long-term goal of moving towards more authentic sources. This is an observation which is repeated by Weir (2013) and Taylor and Garenpeyah (2013) in their recommendations for the future direction of the Cambridge main suite exams. Similarly, Salisbury (2005) describes how IELTS item writers look for suitable written texts from articles, journals and magazines to be used as a basis for script design. Most of the item writers devise items before writing the script, raising the question of whether listening is really being tested. As Vandergrift and Goh (2012) have pointed out "far too often listeners are expected to be able to understand texts that are meant to be read" (p.167).

The idea of using authentic material is not new, however, even though test material is often informed by authentic input, unscripted spoken audios are not normally used for assessment. As Wagner highlights;

> A review of the spoken texts used in the listening section of some of the high stakes English proficiency tests (i.e., the IELTS, TOEFL, and Pearson Test of English (PTE)) suggests that virtually all of the texts are indeed scripted, written, and read aloud.
>
> (Wagner, 2013a, p. 7-8)

The fact that most listening tests do not use authentic sound files can be put down to the fact that it is argued that it is often difficult and inefficient to create comprehension questions from such texts (e.g., Buck, 2001). However, I would argue from experience that nowadays this is no longer the case due to the vast resources that can be found on the internet. Furthermore, samples of natural, non-adapted, connected speech can be easily collected by using pre-prepared interviewer prompts (see Green, 2017).

In sum, oral input in a test which does not contain the typical attributes of real life listening would mean that the construct is under-represented in the test. Consequently, following the general call for the use of authentic input texts in listening tests, the present study will only use authentic texts sourced either from the Internet or constructed in response to prompts in order to collect samples of non-adapted, continuous speech.

3.4.1.2 Channel

The choice of mode of presentation of input will rest on a number of factors, not least the practical considerations of test delivery. Obviously, video input will be much easier and more efficiently delivered in a computer-delivered test where each candidate has their own screen and headphones. The particular TLU domain in question will also influence the choice we make: as we wish to have both situational and interactional authenticity, we need to decide whether or not visual information would be available in the TLU (Wagner, 2010). For example, 'understanding a radio documentary' would only require an audio channel in order to be authentic and using video here would introduce construct-irrelevant variance. The call to include video in listening tests comes mainly from SLA researchers, who highlight the important non-verbal information listeners receive such as appearance, gesture and body language. Also, using videos in tests would help to replicate what happens in the language classroom (Wagner, 2013a). Nevertheless, there is still much debate concerning the use of videos in the language testing field. For example, Buck (2001, p.172) argues that visuals should not be included in the L2 listening construct because "we are usually interested in the test-takers' language ability,

rather than the ability to understand subtle visual information", which could well disadvantage those test takers who are unable to utilise non-verbal information. Investigation into the use of video in language tests has not been conclusive, and following Buck, there exists a strong belief that using video input actually contaminates the construct of L2 listening comprehension. Li (2013a) attempted to address the issue of construct definition and test authenticity using a validity argument approach, and found that the supposition that video listening tests increase authenticity is not completely supported because of issues with 'interactional authenticity'. Consequently, Li calls for more research.

Gruba (1993) found that a video version of a simulated lecture did not improve test performance when compared with the audio-only version. This view is supported by Batty (2015), whose detailed study using multi-faceted Rasch modeling found no interactions either between format (audio or video) and text-type (monologues, conversations and lectures), or between format and proficiency level. In a comparative study of a text presented in both audio and video format, Coniam (2001) actually found that the participants understood more of the text when only the audio was presented. Furthermore, 80% of the participants did not believe that the video input had enhanced comprehension, with nearly a third reporting that they did not look at the screen and the majority stating they preferred audio input only. Coniam concluded that high stakes listening comprehension tests should only use audio input. In a study on the comparative effects of audio-only input, video input, and audio input with still pictures, Suvorov (2009) found that scores on the video task were significantly lower than on the other two task inputs. It was also found that those listeners who stated a preference for audio-only text performed better on this task type. Vandergrift and Goh (2012) suggest that this could be because listening ability is related to learner style. Similarly, Wagner (2007) videoed learners doing video listening tasks and found that they paid attention to the video 69% of the time on average, with some learners watching as little as 15% of the time. In a follow-up study, Wagner (2010) showed that the participants only paid attention 48% of the time on average and this, again, seems to point to differing listening styles. What is more, time spent viewing the video correlated negatively with score.

Wagner (2010) argues that this could be explained by the fact that higher proficiency listeners focus only on the audio whilst lower proficiency listeners look for contextual clues in the video to fill in gaps in understanding. Wagner (2013b) later found that test takers who received audio-visual input scored higher than those who received audio only input. Li (2013b) used verbal report and semi-structured interview methodology to investigate strategy use on a listening test presented as audio only compared to video. Different strategies were reported for the two input channels, with test takers using the visual input to build, refine and confirm hypotheses. It was also found that some of the visuals confused or hindered understanding.

Nevertheless, it has, been found that content visuals can have a positive impact on listening tests. Ockey (2007) found that still images helped listeners, but met with differing opinions as to the usefulness of video input. Similarly, in a study of context and content visuals, Ginther (2002), found that not only did test-takers prefer tasks with content visuals, but that the still images used could provide content information and help listeners by complementing the audio. Yet the context visuals had a negative effect on listening comprehension and it could be argued here that they give clues which are construct irrelevant. A recent study by Suvorov (2015) triangulating results from eye-tracking technology, verbal reports and test scores, found little difference when either content and context videos were presented to the test takers.

A related opinion is that some contextual information would be available in most real-world listening situations which could help the listener anticipate what the speaker might say next and so allow them to activate appropriate schemata (e.g., Wagner, 2013a). Here, it is argued that a context visual on a paper-based listening task could be helpful before listening for activating top-down processing strategies to compensate for inadequate linguistic knowledge. However, they are less helpful during listening because these visuals require processing in addition to the audio, thereby consuming additional attentional resources and limiting the amount of working memory capacity available to the listener to attend to the audio (Vandergrift & Goh, 2012).

Test-taker interaction with video is therefore complex and it would seem then that there is a body of research which supports the view that video input may well prove problematic in listening tests. Taylor and Geranpayeh (2011), for example, conclude that while academic lectures may indeed provide visual clues such as a powerpoint presentation or facial expressions and gestures, their inclusion in listening tests also increases the cognitive load for the test taker. An extended video-listening construct may be more authentic yet still bring with it a certain amount of construct irrelevant variance. As Li (2013a) concludes, the inclusion of such tasks would require a well-defined construct of video-listening for TLU domains involving visuals – including necessary visual literacy skills. Similarly, Batty (2015, p.18) calls for "the definition of a new construct of visual listening comprehension", which may be appropriate in some testing situations. Again, we are reminded of the primacy of the TLU domain. The CEFR provides separate descriptors for non-participatory transactional listening, which do not include visual literacy skills.[35] The present study will propose that there is ample reason therefore to present the test tasks in an audio-only channel, with a description of the context of use along with a still image context visual to help the test taker imagine the context of the situation.

3.4.1.3 Lingua Franca and accent

The use of English as a lingua franca has been the topic of much recent discussion and debate (see for example, Canagarajah, 2006; Jenkins, 2007; Jenkins & Leung, 2013; Taylor, 2006). The notion of the 'native speaker' found in many assessment scales is being increasingly questioned in a globalised world where English is used as a lingua franca. Indeed, it is worth noting here that the new CEFR companion volume (2017) removes any reference to native speakers from the CEFR descriptor scales. However, use of L2 speakers does not figure in any of the previously discussed task characteristic frameworks, although accent is included as a general category (see for example, Bejar et al., 2000; Weir, 2005b). Studies have also shown that stakeholders have questioned the

[35] Though descriptors are also provided for interaction (CEFR. p. 74-79) and watching TV and film (CEFR. P.71).

relevance of international standardised tests which only use native speaker input. For example, Nagao, Tadaki, Takeda and Wicking (2012) found that teachers questioned the relevance of the Cambridge PET listening test content in a Japanese context because world English varieties were not represented. In the context of the present study, I therefore strongly believe that a discussion of this aspect is necessary, as Spanish students will most definitely be using English as a lingua franca.

English as a lingua franca (ELF) is defined as "communication in English between speakers with different first languages" Seidlhofer (2005, p.339). The fact that most English language tests ignore this important communicative context has been widely criticised (e.g., Jenkins & Leung, 2013). Harding (2014b) argues that the core of communicative competence in ELF relies on language users' adaptability, that is, on an ability to move between different language varieties. Learners need to be able to "tolerate and comprehend different varieties of English: different accents, different syntactic forms and different discourse styles" (Harding, 2012). Harding (2014b) proposes two solutions to this problem: 'drop-in' ELF competences within the present framework of CLA, and completely new purpose-built ELF frameworks, which he considers to be superior but difficult to implement.

In sum, when talking about language use we must include context as an important part of the measure and this depends on the purpose of the test as the social context dictates the language variety which is used. The context of language use is expanding due to increased globalisation and L2 users will find themselves in a range of situations which may include not only many different native speaker varieties of English but also non-native speaker varieties. This issue has been addressed by the call for localised context specific tests (Canagarajah, 2006), with the fragmentation of the testing industry predicted as tests become more localised (O'Sullivan 2011). The argument here is that it is impossible to create a test of 'universal proficiency'. It is clearly a test developer's responsibility to ensure tasks adequately represent the TLU situation in question. As such, I would argue that in the context of the present study, ELF has a place in the construct definition.

In terms of listening test input then, accent is an issue which relates to the previous discussion. Learners should be able to adapt to situations which present different varieties of accents. Most international tests include the standardised accents from major native-speaker varieties (British English, American/Canadian English and Australian English) to "reflect varieties of English that enable (test takers) to function in the widest range of international contexts" (Taylor, 2006, p.57). Indeed, it is felt that an unfamiliar accent will affect the difficulty of the task, a belief reflected in the CEFR listening descriptors.[36]

Yet research to date regarding this premise has been varied. In an attempt to provide evidence for a multidialectal listening test, Ockey and French (2014) found that both familiarity with accent and strength of accent had an effect on comprehension. In a different study, Major, Fitzmaurice, Bunta, and Balasubramanian (2002) examined the effects of non-native accents for the new TOEFL test on listening comprehension. They produced TOEFL test tasks using L1 Chinese, Japanese, Spanish and American English speakers. Test takers with these native languages were then asked to complete the tasks. While the results showed that Spanish speakers performed better on tasks where the speaker was also L1 Spanish, this same advantage was not seen to extend to Chinese and Japanese test takers. However, the study did acknowledge the fact that task difficulty was not controlled across the tasks and so these results should be interpreted with caution. Harding (2008) used questionnaires, interviews and focus groups to investigate test-takers' perceptions of an academic listening test which included speakers with different accents (Australian English, Mandarin Chinese, Japanese and Bengali). He found that perceptions of task difficulty related to accent differed depending on the proficiency level of the test-taker, and in some instances a shared L1 distracted listeners. In a more recent study focusing on the Australian context, he found that using speakers with Japanese and Chinese accents did not cause major problems for the listeners and argues that L2 accents should be incorporated into academic listening tests if we want to have good construct coverage (Harding, 2011).

[36] For example, 'Standard dialect' is mentioned in some of the B2 listening descriptors (CEFR, p.66-67).

It may be necessary to have multidialectal listening skills to communicate successfully in English speaking contexts. If we want our test to represent key features of the TLU domain, the fact that a range of English accents are used in an international context should be reflected in assessment procedures. Indeed, Yanagawa (2012) cites lack of variety of English accents and lack of L2 speakers as two possible features which could threaten the validity of the JNCTL test. In the context of the present study, most universities in Spain expect students to study some of their subjects in English following the concepts of EMI, and these classes are normally given by an L2 English speaker. I would also point out here that if we use authentic listening audio files we are likely to find an abundance of both L1 and L2 accents, and would not simply have the accents of actors, often chosen in the British SLA context for their Standard Southern British accent. Again, our TLU domain must be represented in the test and I would therefore argue that for the present study a range of standard dialects as well as some representation of L2 accent should be included in the new BFE listening test. Here, it is useful to provide an example item taken from the beginning of the audio file, thereby giving the test taker some time to adjust to the speaker's voice.

3.4.1.4 Linguistic complexity

Other passage characteristics which are argued to have an effect on task difficulty include the linguistic complexity of spoken discourse. As previously stated, lexical complexity has been shown to correlate highly with L2 listening ability. Listeners need to recognise the words in the audio file and are probably more likely to recognise high frequency vocabulary. Using multiple regression analysis, Mecartty (2000) found that lexical knowledge was a significant predictor of listening performance. Bonk (2000) used passages with different lexical familiarity and, using a dictation test, found that in general, higher scores were obtained on the passages where subjects were more familiar with the vocabulary. However, some of his subjects could understand a passage even when their lexical knowledge was less than 75% of the text and others did not perform well even when they had 100% lexical knowledge of the passage. In a study of TOEFL dialogue items, Nissan, DeVincenzi, and Tang (1996) found that the occurrence of

infrequent words was a predictor of item difficulty. Yet Yanagawa and Green (2008) found the opposite, with infrequent words making dialogue items easier. Brunfaut and Révész (2013) carried out a detailed study which included high frequency words, formulaic expressions and lexical density, and found evidence to support the idea that more frequent words and expressions are easier to understand – especially for the parts of text which contained the necessary information. Here, the participants themselves identified lexical complexity as the most important factor affecting text difficulty. Similarly, Staehr (2009) found that 51% of listening variance could be explained by L2 vocabulary and Mathews and Cheng (2015), in a study using the IELTS listening test, found that recognition of words from speech from the third thousand-frequency level could predict 52% of the variance observed in the listening comprehension scores. Vandergrift and Baker (2015) have presented similar results regarding vocabulary knowledge. Their study was innovative as they used an oral vocabulary knowledge test. The results led them to conclude that "listeners need to attain a certain level of vocabulary knowledge before they can efficiently transfer L1 skills to L2 listening tasks" (p.407). Similarly, Cheng and Mathews (2018) found a strong correlation between productive phonological vocabulary knowledge and listening ability and this knowledge explained 51% of the variance in listening test scores. In reading, the vocabulary knowledge necessary to understand a text has been reported to be 98% (Hu & Nation, 2000) and Staehr (2009) reports that the same figure should be used for listening; however in a more recent study it was reported that the figure could be reduced to just 90% for listening (van Zeeland & Schmitt, 2013).

Kostin (2004) carried out a study similar to that of Nissan et al. (1996) in an attempt to inform item writers about factors which could help with the difficulty levelling of listening items. The results showed that, of the variables analysed, predictors of item difficulty included lexis, idioms, negations, syntax and content. Here, I would point out that isolating vocabulary frequency as a predictor of item difficulty is not an easy task and it is more likely that a combination of textual features are working together to make an item easier or more difficult. However, many testing bodies see lexical frequency as an indicator of task difficulty and include this aspect of contextual features in their test

specifications. For example the British Council's APTIS test includes a section in their task specifications for lexical frequency bands to be used with each CEFR level. With regard to the linguistic parameters of both the APTIS reading and listening task input, O'Sullivan (2015, p.48) states that "lexical profiles are provided for all input texts (including instructions and prompts) and are based on the *Compleat Lexical Tutor* (www. Lextutor.ca)". Similarly, in my own experience as an item writer for an international proficiency test provider based in the USA, specific guidelines were given that only vocabulary which pertained to the CEFR level as stated by the Cambridge *English Vocabulary Profile*[37] project could be included in the listening input text scripts for that level. However, I feel that it should be highlighted here that tools developed to be used with reading texts may not be the most appropriate for a listening text.

Information or propositional density is another variable which has been studied as a measure of task difficulty. Indeed, Field (2013a) argues that many exam boards tend to include in their texts scripted information to be used to create distractors. He argues that this information is often much more dense than would normally be found in spoken discourse and as such presents a type of 'cognitive overload' on the listener, whose short term memory is unable to deal with the input. Buck and Tatsuoka (1998) reported a positive association between task difficulty and the proportion of content words which surrounded the necessary information. Similarly, Rupp, Garcia, and Jamieson (2001) concluded that, along with other variables related to the audio input, information density was a predictor of item difficulty. Here again, a number of variables were studied and by using linear regression the researchers were able to show that the text input itself, together with an interaction with the text, can explain item difficulty. It is also argued that a large percentage of content words can have a negative effect on the listener, as they carry more information (Bloomfield et al., 2011). Such lexical density was found to be a predictor of task difficulty in the study by Brunfaut and Révész (2013).

[37] See www.englishprofile.org

A related variable which has been investigated is the explicitness of the information in a text. It has been found that more abstract texts, that is texts in which the listener has to make more use of inference skills, are more difficult (e.g., Freedle & Kostin, 1999; Kostin, 2004; Nissan et al., 1996). However, the evidence is not conclusive and Brunfaut and Révész (2013), for example, did not find this to be the case.

Conversely, studies have shown that grammatical knowledge does not affect task difficulty. Mecartty (2000) found that syntactic simplification does not make a text easier, possibly because the text becomes less coherent and clues to meaning and redundancies are not present (CEFR, p.165). However, cohesive links were not found to affect task difficulty in the Nissan et el (1996) study, but causal content and referential cohesion were reported as indicators of difficulty in both Revesz and Brunfaut's (2013) and Brunfaut and Révész's (2015) studies.

Several studies have used detailed statistical techniques to examine the effect of linguistic features on task difficulty. In a study of the listening section of the GEPT test, Liao (2009) provided evidence that lexico-grammatical knowledge is a significant predictor of listening ability. Aryadoust (2011a) applied the fusion model to the IELTS listening test: he found that linguistic features of the items, the ability to paraphrase, the ability to understand specific information and the ability to integrate listening and reading in the short term memory had an influence on test scores. That is to say, difficulty depended on an interaction between task and audio file. Similarly, Huff (2003) studied the reading and listening sections of the TOEFL test and reported that both characteristics of the text and the interactions between item stems and the oral passages accounted for 48% of the variance in listening item difficulty. Another study by Sawaki, Kim, and Gentile (2009), which applied the fusion model on the TOEFL iBT, found listening ability was influenced by ability to understand general information, details, text structure, and the intention of speaker, as well as the ability to link ideas. Lee and Sawaki (2009) found similar results using latent class analysis, diagnostic models and the fusion model.

As may be seen from the studies reviewed, the results are conflicting and no definite conclusions can be drawn. It may be that the division of lexical and grammatical knowledge is an impossibility, as listeners may well combine both syntactic and semantic cues in interpreting the sentence. As Brunfaut and Révész (2015, p.160) state "a one-to-one relationship between each individual task variable and task difficulty is unlikely; manipulations of individual task characteristics would be expected to bring about changes in other characteristics". Likewise, Jensen, Hansen, Green, and Akey (1997) found that no text-related variables had any effect on item difficulty. Rather it was the items themselves which made a task easier or more difficult. Jensen et al. (1997) showed that there was a relationship between task difficulty and lexical overlap between words in the text and the items. This highlights the fact that items should not contain the exact words as they appear in the text. It would seem then, that as previously stated, item difficulty depends on the demands of the task, an interaction between text and task, and not isolated variables concerning the audio input.

3.4.2 Characteristics of test task

As well as considering characteristics of the audio file, decisions need to be made about the test tasks. These decisions should then be outlined in the test specifications, and in effect be a description of the final test format.

3.4.2.1 Number of plays

Wagner (2013a) states that a superficial analysis of the TLU domain tells us that listeners would not normally have the opportunity to listen twice to any given spoken input, yet in most situations listeners would be able to interact with speakers and ask for repetition or explanation. There is therefore debate about the number of plays of an audio file during a listening test. Theoretical arguments include the authenticity of the TLU situation, although in most situations we only listen once it is argued that technological advances have meant we can listen to online materials as many times as we want (Field, 2015). It would follow that audio files containing normal conversation would not

normally be repeated, at least not word for word, but radio podcasts and other material sourced from the internet can be repeated many times. Here, we must also think about the practicality of delivering the test. Double play means a test is twice as long, while single play would allow for twice as many items making the test more reliable. It is further argued that, in a testing situation, listeners may have other problems to deal with, such as the absence of visual support, bad sound quality or background noise, or impromptu occurrences such as someone coughing or traffic noise which could take place on the day of the exam (Field 2015; Geranpayeh & Taylor, 2008). Geranpayeh and Taylor (2008) go on to argue that the Cambridge main suite listening tests are played twice due to tradition— this is what is expected by stakeholders. It is also argued that because a listener is not familiar with a speaker's voice, a second play allows time for listeners to adjust to the speech and that test taker anxiety is therefore reduced (Field, 2015).

There are other arguments for and against double play based on empirical investigation. Much of this investigation has been concerned with the effects of double play on item difficulty and discrimination. Research shows that as the number of plays increases, so does test taker performance. Chang and Read (2006) found that repetition and provision of background knowledge had a significant effect on the final listening test scores, especially for higher proficiency participants. Field (2009, cited in Field, 2015) found that middle ability participants with IELTS scores between 5 and 6.5 made the most gains during the second listening, as opposed to either the lower or higher scoring participants (although the sample used in the study was very small). Sakai (2009) found that the second listening did lead to more precise comprehension, once again for the higher proficiency participants in particular. Vandergrift and Goh (2012) argue that this is likely to be because more meta-cognitive strategies are used on the second listening in terms of reflecting, planning and selecting. Conversely however, it has also been found that repeating input has no effects on item difficulty (Brindley & Slatyer, 2002).

More recently, there have been concerns about the effects of double play on the test construct, that is to say an exploration of whether test takers listen differently on the second play. In Field's (2015) study of an IELTS listening test, normally played only

once, the test was played twice to the participants. Field found a general increase in test scores, but makes the caveat that results varied between individuals. His results showed that 61.65 % had an increased score, 27.4% had no change in score and a small minority had a decreased score on second play. Many participants did not answer items which had not been answered on the first listening and they did not change incorrect answers. In terms of item discrimination all the participants did better regardless of proficiency level, and in particular participants did better on the second play for constructed response type items – a finding also shared by Boroughs (2003). Using verbal report and semi-structured interview methodology, Field (2015) found differences in cognitive behavior during the second listening. These included: more familiarity with input, reduced anxiety, and a tendency to use lower level processes on the first listening but higher level processes on the second. This latter finding was also reported by Buck (1991): seemingly, listeners used more lower-order processes on the first listening and more higher-order processes on the second. Field (2015) argues that therefore listeners use more authentic processes during the second listening and concludes that, consequently, double play is better.

3.4.2.2 Item preview

Another discussion about task characteristics has centred around question preview, that is to say whether the test taker should be able to see the question before listening or not. Buck (1991), amongst others, argues that the items provide contextual information and help listeners to know what they are supposed to be listening for and so act as a motivating factor. Furthermore, as previously stated, it is important for a CEFR-related test to provide a purpose for listening. Conversely, however, it has also been argued that by providing the test takers with the items before the input text we are giving the test taker contextual clues which would not normally be found in the TLU domain, thereby decreasing the authenticity of the test (Hughes, 2003).

Empirical studies investigating this aspect have reported varying results. Berne (1995) found that participants who previewed the MCQ questions scored higher than

those who did not. This finding was supported by Elkhafifi (2005). Similarly, Chang and Read (2006) found that MCQ question preview helped higher ability listeners, however they found that this was not the case for lower ability listeners. A follow-up study, Chang and Read (2008), found that question preview helped higher-level listeners because of their choice of strategy use. It was argued that question preview reveals content clues and encourages prediction and selective listening. Yanagawa and Green (2008), in a study of TOEIC listening scores, used an ANOVA analysis to look at MCQ question preview under three conditions: (i) a full preview of question and options, (ii) a preview of options only, and (iii) a preview of question stems only; they found that while (i) and (iii) yielded similar results, (ii) led to a significantly lower score. This is probably unsurprising, as without the stem the listener cannot contextualise the required information. Also, it seems that by having access to the options the participants were not given an advantage over simply seeing the stem. On the other hand, Badger and Yan (2012) concluded in their study of strategy use by IELTS candidates that the information in the items leads to the increased use of test-wiseness strategies, and consequently recommended that questions should not be previewed. Field (2015), however, argues that such conditions can only feasibly work for short recordings which are followed by one item only, due to working memory constraints.

For note form (NF) type items, one important study is that of Sherman (1997), who investigated this item type under three conditions: (i) listening twice with no item preview, (ii) item preview followed by listening twice, and (iii) listening once followed by an item preview before the second listening. Sherman found that when items were presented between first and second play scores were significantly higher than other forms of item presentation. She argued that the reason for this was that during the first play the test-takers listened in a more natural way. Similarly, Field (2015) argues that such an item presentation format has a number of advantages, such as eliciting more global processes. However, he does point out that such a test format lends itself to computer delivered tests and would be difficult to implement in paper-based delivery formats. In both the above mentioned studies the test takers believed that question preview would be

helpful, even though this was not found to be the case. It could therefore also be argued that question preview reduces test-taking anxiety.

The above cited studies were all carried out with audio only input tests. A recent study by Koyama, Sun and Ockey (2016) looked at the effects of item preview on a video MCQ listening test. They initially found that the amount of item preview did not affect test scores, though further analysis using a one-way ANOVA with amount of item preview as the independent variable did show that question preview led to higher scores. However, as in Yanagawa and Green (2008), the question only and question with options condition showed no significant resulting difference in test scores. Unlike the studies by Sherman (1997), and Chang and Read (2006), high and low proficiency students benefitted in similar ways. The researchers suggested that this could be because different item types benefit different students, and recommend more research in this area. However, Wagner's (2013b) study of a video listening test showed that question preview did not significantly affect test scores.

The differing findings reported above mean that no general conclusions can be drawn. I would again argue that the important question here is not one of item difficulty but rather one of construct validity. As previously argued, listening purpose is paramount and, following the CEFR, the test taker should be able to activate those strategies relevant to the task. Indeed, (Weir, 2005a, p.289) argues that "having a clear purpose in completing a task will facilitate *goal-setting* and *monitoring*, two key meta-cognitive strategies in language processing". I, like Buck (1991), would argue that the items provide the context and the purpose for listening and should therefore be previewed.

3.4.2.3 Response format

In order to determine whether or not a test taker has understood a given audio input, we must provide (normally written) comprehension items. It is an indirect test of the skill, and we can only make inferences about test takers' ability based on their answers to these comprehension questions. For a listening test, we have a number of item types to choose

from, both constructed response and selected response (for an exhaustive list of possible task types, see Vandergrift and Goh, 2012, p.172). Some typical item types can be disregarded from the off. For example, 'True/False' type items present the test taker with a 50% chance of getting the item correct, the obvious problem being their openness to guessing, which thus negatively affects test reliability.[38] For such items to be valid, very large numbers would need to be included and so they are not considered practically appropriate (Alderson et al., 1995; Hughes, 2003). Likewise, the test method of giving a third 'not given' option, as seen in many reading tests, would not be appropriate for a listening test because listeners tend to listen for what is said rather than what is not said, and this question format would therefore not be a representative test of the skill. In a listening test, construct irrelevant variance can be introduced by the response format. Written questions and options on an MCQ item involve reading, and open format responses such as NF involve both reading and writing. Indeed, Field (2013a, p.131) draws attention to verbal report data from his 2012 IELTS study, arguing that gap-fill type items are more complex than real world listening because test takers must read, write and listen at the same time.

Selected response type items, such as MCQ or MM, have the obvious benefits of being objective and therefore more reliable: they are easily scored and results are readily available for statistical analysis. The type of response format chosen has been shown to have an effect on item difficulty. In'nami and Koizumi (2009) present a meta-analysis of MCQ type items versus open-ended questions and conclude that there is evidence to show that MCQ formats are easier. It could be argued that this is because it is more difficult to generate accurate information than recall it. Also, investigation into listening strategy use has shown that MCQ type items allow for uniformed guessing (e.g., Barta, 2010; Buck & Tatsuoka, 1998; Wu, 1998; Yi'an, 1998). However, it has also been argued that selected response type items are in fact the most efficient at measuring cognitive abilities and that criticisms of these item types can be mainly put down to badly written items (Downing 2006). Another consideration for MCQ type items is the number of

[38] Elliot and Wilson (2013) argue that because of the problems with this item type this is an aspect which should be further researched in terms of the true/false items presented in the Cambridge Main Suite exams.

options to be presented to the test takers. Here, the jury is still out and there are differing opinions as to whether three or four options should be used (e.g., Boroughs, 2003; Lee & Winke, 2013; Rodriguez, 2005). Obviously, badly written items will have an effect on test reliability in either case and attempts should be made for all options to be as plausible as possible.

Furthermore, it has been shown that long question stems or options affect listening item difficulty (Jensen et al, 1997). This study found that lexical overlap between words in the text and the response reduced task difficulty. Aryadoust (2013) found that the ability to understand and paraphrase the written stem caused extra difficulty to the listener and an inability to do so would lead to an incorrect answer even if the input audio had been understood. We do not want test takers to spend too much time reading item stems and options as they may miss subsequent items. Listening test items should therefore be kept purposely short and easily understood in order to reduce the amount of reading and construct irrelevant variance, and great care should be taken during the item writing process that clues which could help listeners answer correctly without understanding the text are not provided in the written questions. Questions should be presented in easily understandable language, perhaps one level below that of the targeted proficiency level (Green, 2017).

It has also been argued that certain response formats lend themselves to different listening types when following a cognitive processing view of listening ability. Field (2011) notes, for example, how a NF task targeting specific local information necessarily requires successful decoding. A certain amount of hypothesis testing takes place during parsing, and the distractors presented by an MCQ item could arguably produce this forming-and-testing hypothesis. Field does however qualify this, noting that the hypothesis originates in the written items and not the input text itself. He argues that as well as minimising the reading load, multiple match (MM) type items work well for testing global understanding because the items do not have to be presented in the same order as the input text (Field, 2013b).

In sum, it is generally advised that a test should include multiple tasks in order to cover a range of language and provide wider evidence of test takers abilities and a variety of response modes should be used to lessen the chance of construct irrelevant variance (Alderson et al., 1995, p.44-5). Tentative conclusions can be made from the above review and we should aim to minimise construct irrelevant variance introduced by reading and writing. Listening test items should be presented at a level below the proficiency level of the test and should be kept as precise as possible. Certain item types lend themselves to the testing of certain listening skills. For example, MM type items are suitable for the testing of Gist/Main ideas and NF type items can be used to test local specific information (SI) and search listening.[39] This is something which will be taken into account when drawing up the test specifications for the present study.

3.4.3 Summary

The fact that an interactionalist approach rests on interactions between context and use makes the two very difficult to separate. Weir (2005b, 2010) confirms that his framework's elements are presented separately for descriptive purposes only. Bachman (2007, p.55) also highlights the problem of solving the issue of just how abilities and contexts interact and the degree to which they mutually affect each other. Some aspects of context validity are addressed by Field (2013a) in his discussion of cognitive validity (e.g., purpose for listening and test method), again showing the indivisible nature of the two validities. The symbiotic nature of context and ability as part of construct validity should therefore be emphasised.

The present study presents an ability model based on a cognitive processing view of listening comprehension. The test tasks should represent the context of use and the test should allow for the collection of evidence about the ability of test-takers in the context of interest. By using tasks and contexts which can be specifically related to our TLU

[39] However, Elliot and Wilson (2013) highlight the difficulties in constructing the key for these types of items and state that the Cambridge Main Suite exams tend to use nouns as the key because there exist fewer possible paraphrases.

domain, the accuracy of extrapolation from task to TLU will be greater (Bachman & Palmer, 2010). The present study has a TLU relevant to school leaving and university entrance and here we need to include tasks which do not require any specific prior knowledge, as we do not wish to disadvantage any of the test takers. As Rost (2014) argues, listening involves schematic transfer and our cultural and general prior knowledge plays an important role in L2 listening. Topics used for the present study should therefore reflect those topics used in the baccalaureate classroom and be relevant to university life without being too culturally or content specific. The TLU domain needs to be well sampled and include a number of different tasks each with different listening purposes in order to sample a range of types of listening and enable good construct coverage.

In terms of the items, we should try to make the task as genuine as possible. For Vandergrift and Goh (2012) authentic listening is that which reflects the purpose, skills and outcomes of real-life listening. Here we need to model the skill on an expert listener (Field, 2012a), and consider just what an expert listener would need to take away from the text. According to Field, items should never be developed from a transcript; rather, the acoustic phonetic signal should be placed centre stage in order to take account of the "relative salience of ideas" (Field, 2013a p.150). The *text-mapping process* described later in Chapter 4 allows the test developer to do this.

There is a general consensus that the test should be practical and easy to administer, which is a strong argument for a computer delivery mode. However, while much has been made of the argument for moving towards a computerised delivery mode (Chapelle & Voss, 2016; García Laborda, 2007), I would argue that before any mode of delivery is decided, it is essential that a new exam construct be defined. There is no issue with the test being initially implemented on paper and the platform being gradually moved towards methods of computer delivery. As López Navas (2012) comments, any plans to computerise the test are simply concerned with the delivery method; it is the test construct which should first be addressed. As long as the current construct is thoroughly revised so that the English PAU overcomes its most significant drawbacks, the process

could eventually lead to a more natural transition of the exam format, provided that the teaching community and the financial situation allows for this to happen (López Navas, 2012).

The present study will develop test specifications and test tasks to be delivered as a paper-based test using a scannable answer sheet to enable fast accurate scoring via an optical reader, but which could easily be transferred to a computer-based delivery mode in the future. Once the test has been developed by operationalising the construct through the test specifications, it further needs to be evaluated within a sound theoretical framework for evaluating test validity. In order to provide such a framework, therefore, we must next take into account current literature in the field on validity theory.

3.5 Contemporary Validity Theory

Current validity theory draws on Samuel Messick's (1989) progressive matrix for framing test validation (see Figure 7). For Messick, validity is seen not as an innate property of the test but rather as a unitary, multi-faceted concept which should be based on multiple sources of evidence in order to substantiate the inferences made about candidate performance and justify the resulting decisions taken on the basis of test scores. He states:

> Validity is not a property of test scores and other modes of assessment as such, but rather of the meaning of the test scores. Hence, what is to be validated is not the test or observation device per se but rather the inferences derived from test scores or other indicators—inferences about score meaning or interpretation and about the implications for action that the interpretation entails.
>
> (Messick (1996, p. 245).

Figure 7. Messick's progressive matrix for validity

Source of justification	**Test inferences**	**Test use**
Evidence	Construct validity	Construct validity + Relevance/utility
Consequences	Construct validity + Value implications	Construct validity + Value implications + Relevance/utility + Social consequences

(Adapted from Messick, 1989, p. 20)

This is a definition which has shaped subsequent notions of language test validity and forms the basis of current language test validation work. It can be seen that construct validity is present in every cell in Messick's matrix and is central to any test validation endeavour. Indeed, test validation has become very much construct-driven (McNamara,

2006), a practice applauded by many. Alderson and Banerjee (2002), for example, have stressed that test validation should not only be concerned with the psychometric analysis of tests but should be based on theories of language proficiency informed by knowledge and experience from other fields.

This integral role of the construct in any validity claim means that our construct needs to be extremely well defined. However, despite consensus on the need to define our constructs, the language constructs themselves are continually being debated and a strong theory of language proficiency is still lacking (Chapelle, Enright & Jamieson, 2010). Discussing this issue, Bachman (2007) outlines three approaches to the definition of language constructs: ability-focused, task-focused, and interaction-focused. He concludes that all three perspectives should be included in validity research and both qualitative and quantitative methodologies should be employed to strengthen any claims made about a test. A uniquely competency-based approach might well include construct irrelevant factors (Chapelle et al., 2008), while a uniquely tasked-based approach which tried to use tasks eliciting the same performance as a real-world situation might suffer from construct under-representation. Constructivist views on defining the test construct see the construct as a trait that exists within an individual, (e.g., listening proficiency), and attempts are made to model such traits. For example, many models for communicative competence exist which include all the knowledge and skills necessary for having the ability to perform in a particular domain. It has been argued that such a view is difficult to operationalise (e.g., Kane, 2004) and that by limiting validation research to the construct other validity evidence pertinent to test use could be ignored due to the fact that no attempt is made to expand the conceptualisation of validity (Aryadoust, 2013). As a result of these shortcomings, recently test developers have begun to adopt an argument-based approach to test validation, which allows the combination of both approaches by providing a framework to collect multiple types of validity evidence in support of the inferences and decisions to be made on the basis of test scores.

Indeed, while Messick's model provides a sound theoretical basis upon which to carry out the validation process, we still require a clear conceptual framework within

which inferences can be investigated in order to make use of any evidence we collect (McNamara, 2006). In practice, Messick's model has been noted as being too abstract (Xi, 2008) and therefore difficult to operationalise (Bachman, 2005; Davies & Elder, 2005; Kane, 2012). Various efforts have therefore been made to provide operationalisation frameworks that offer guidance on how to integrate validation into the test development process (Alderson & Banerjee, 2002), often with a view to address the social consequences of tests (e.g., Bachman, 1990, 2005; Bachman & Palmer, 1996, 2010; Kane, 2002, 2012; Weir, 2005b). McNamara (2006, p.48) has argued that this 'consequential aspect' of validity is one of Messick's most important influences on language testing, claiming that "Messick's [work] remains the most comprehensive conceptualisation of the validation process available to date".

Certainly, the high-stakes nature and serious consequences of many language tests has prompted much debate about ethical and fair practice issues. As a consequence of a call for greater professionalism (e.g., Davies, 1997, 2008, 2010, 2012) a series of codes and standards for good professional conduct now guide the work of test developers (EALTA, 2006; ILTA, 2000, 2007). For example, it is argued that the EALTA codes could be used to provide a framework for test validation (Alderson, 2010, p.63) and a few studies have followed these guidelines to undertake the test validation process (e.g., Pižorn & Moe, 2012; De Jong & Zheng, 2011). However, these codes are also open to criticism; because they need to be agreed on by everyone within the testing community, they are necessarily vague (Fulcher & Davidson, 2007).

As well as codes of ethical professional practices, there has also been a call for the implementation of test fairness frameworks. Kunnun (2008) defines fairness in terms of the use of fair content, test method, and scores, as well as test administration. Similarly, Xi (2010) concludes that fairness is part of test validity. Yet in contrast, Davies (2010) believes that absolute fairness in testing is an impossibility and that because social values change, arguments about such values will be open-ended and discussion between all stakeholders remains an ongoing necessity. Likewise, Fulcher and Davidson (2007) argue for effect-driven, democratic testing which welcomes collaboration among all

stakeholders. It should, however, be highlighted here that different stakeholders' views may well conflict (Bachman, 2005; Hamp-Lyons, 1997, 2000).

Despite advancements in the areas of ethics, fairness, transparency and accountability, McNamara and Roever (2006) have argued that social and political consequences have not as yet been fully addressed and that the test construct itself embodies social values. This view of the social responsibility of test developers (Shohamy, 2001; McNamara, 2006) places a much larger accountability on test developers, making them responsible for all social consequences and giving test developers an obligation to ensure their tests are not misused. In an attempt to address this issue, McNamara and Ryan (2011) question definitions of fairness, using the term 'justice' to describe social values implicit in both test constructs and the social uses of tests. Their notion of 'fairness' may be seen as corresponding to the first row of Messick's matrix, while 'justice' would equate to the broader social implications, requiring evidence relating to wider educational, social and political policies. Such concerns have generated much debate about the extent of test developer accountabilities (Davies, 2008, 2012; Hamp-Lyons, 1997, 2000; Kane, 2012; Shohamy, 2001, 2007).

To summarise, the process of validation currently involves logical thinking about test design and uses, as well as an examination of the empirical evidence from test trials and administrations (McNamara, 2006) in which validation is guided by clear theoretical frameworks. Here, both ethical judgements and a clear understanding of test consequences are necessary in order to provide fair and valid tests to the many stakeholders involved. Test developers must always plan for positive impact (Messick, 1996) and must provide evaluative evidence using both quantitative and qualitative methods to support the inferences to be made from test scores. The argument-based approach to validation provides such a conceptual framework and its recent acceptance as the recognised way to provide evidence of test validity is highlighted by the fact that the new *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) calls for a 'validity argument' supporting the

appropriateness of the inferences to be made on the basis of the assessment results. It is a framework which allows for combining quantitative and qualitative methods, because they can be used to support different yet interconnected inferential links (Xi, 2008). It is this framework which will guide the present research and which will be outlined in more detail below.

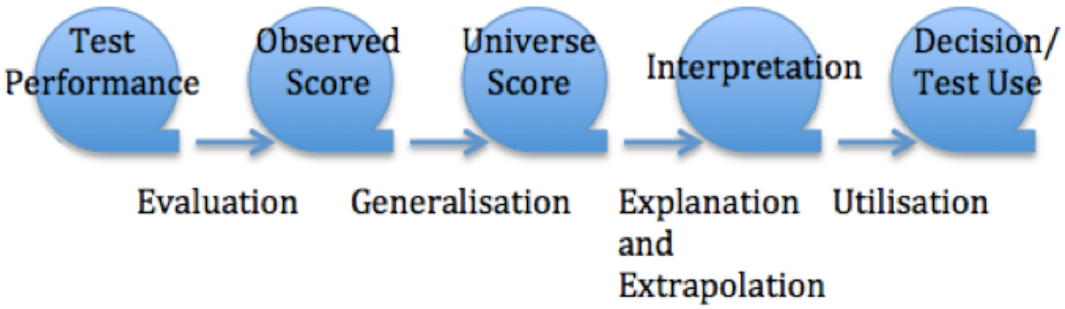
3.5.1 An argument-based approach to validity

Validity and validation is central to any new test development project which attempts to build a suitable assessment instrument from which scores can be used to make relevant decisions. While there have been numerous attempts to provide frameworks for test validation purposes, modern views propose an argument-based approach (e.g., Kane, 1992, 2001, 2002, 2004, 2012, 2013; Kane, Crooks & Cohen, 1999). In language testing the approach has been adopted by a number of authors, such as the well quoted validity study of the new TOEFL iBT by Chapelle, Enright and Jamieson (2008). An important contribution to the discussion is also provided by Bachman (2005) and Bachman and Palmer (2010) who, building on these concepts, provide an '*assessment use argument'* (AUA), which specifies conceptual links between interpretations, decisions and consequences of test use. By following an argument-based approach, we can incorporate the test development stage of the test development cycle and then logically collect different kinds of evidence to support the use of any given assessment. Indeed, Bachman (1990, p.55) asserts that "the single most important consideration in both the development of language tests and the interpretation of their results is the purpose or purposes which the particular tests are intended to serve". Ultimately, the purpose of the test is the driving force behind the types of evidence to be collected (Fulcher & Davidson, 2012; Fulcher & Owen, 2016), as we must validate the interpretation and use of the scores, rather than the scores themselves. Indeed, validity is not a quality of a test; a test cannot be validated, rather we need to demonstrate the meaning of test scores and justify their uses (Chapelle, 2012). As Messick (1996) points out:

> validity is not a property of test scores and other modes of assessment as such, but rather of the meaning of the test scores. Hence, what is to be validated is not the test or observation device per se but rather the inferences derived from test scores or other indicators—inferences about score meaning or interpretation and about the implications for action that the interpretation entails. (p.245)

Below a brief outline of what Kane's argument-based approach entails is given. This approach involves two stages. The first, known as the 'developmental stage' (Kane, 2006), requires the laying out of an *Interpretative Argument* (IA) which includes any claims about test scores and intended uses specified in some detail in order to justify test use. The second, or 'appraisal stage', involves an evaluation of the overall plausibility of the proposed interpretations and uses (Kane, 2012, p. 4). This is the *Validity Argument* (VA) for the test and it should specify score meaning and justify the theoretical framework which underlies the assessment. The framework includes a set of validity inferences which are essentially conclusions reached using some kind of evidence about certain aspects of the test's validity (Aryadoust, 2013). The validity inferences in an IA are depicted in Figure 8.

Figure 8: Links in an interpretative argument (modified after Bachman, 2005 and Kane, Crooks & Cohen, 1999). (Adapted from Xi, 2008, p. 182).



The validity inferences include:

1. The Evaluation Inference

This refers to test scores being fairly and consistently awarded and observations on tasks are evaluated in an accurate and relevant manner. This would mean that scoring needs to reflect the skills and abilities of candidates and should not be influenced by other factors than the construct being tested, such as poor quality of a sound file, insufficient time for reading and answering items or bias in favour of one particular group of candidates. As such, this inference can be investigated by conducting studies about construct irrelevant variance and score reliability, among others (Xi, 2008).

2. The Generalisation Inference

This inference involves the generalisation of observed scores to universe scores and "assumes that performance on language tasks is consistent across similar tasks in the universe, raters, test forms, and occasions" (Xi, 2008, p. 181). The generalisability of test tasks is likely to be questionable and therefore attention should be paid to this at the test development stage (Kane, 2006). Scores must be reliable and generalisable across parallel test administrations. Estimations of internal consistency and inter-rater reliability (for more subjective tasks) using classical test theory (CTT) could be used, as well as G-theory (Xi, 2008). In terms of test development, precise test/task specifications must be developed providing a framework to guide test development, making the creation of parallel test forms more likely. Also, statistical equating techniques should be used to compare test forms over different administrations to check that scores are consistent and interpretable.[40]

[40] However, Bachman and Palmer (2010) define 'generalisability' in a different way from that in the measurement literature. Their definition corresponds more closely to what Kane (2006) refers to as 'extrapolation', it therefore resembles notions of 'construct validity'.

3. The Explanation and Extrapolation Inference

These two inferences have often been grouped together (as in Figure 8) as they are highly related, they both rest on the quality of the test's theoretical underpinnings and degree of construct representation, and therefore this is often considered to be the most important link (Aryadoust, 2013). As such, this link is "commonly dealt with under the heading of construct validity" (McNamara & Roever, 2006, p. 27). The idea of 'cognitive validity' (Field, 2013a) could be seen as part of this inference,[41] along with traditional 'criterion validity'. This link "rests on the assumption that test tasks engage abilities and processes similar to those underlying performance on real-world language tasks indicated by a domain theory" (Xi, 2008, p. 184). Here again we can see how the argument-based approach links back into previous parts of a test development cycle, that is to say, we need to have previously provided a domain theory. Kane (2012, p.9) points out that if tests are based on such domain theories "we incur an obligation to provide evidence in support of the theory". We need evidence to show that candidates use the skills, knowledge, abilities and processes which would be used to complete tasks in the TLU domain and that another measure of the same construct would give the same results. Possible research methods to support this assumption would include verbal protocol analysis, logical and judgemental analysis of test tasks, structural equation modelling, and correlation analysis with other measures of the same construct (Xi, 2008). Kane (2006) recommends the use of combining analytic and empirical evidence to support this inference.

4. The Utilisation Inference

This final link deals with test use and consequences where decisions are based on observed scores that represent the ability of test takers in the TLU. According to Bachman (2005, p.9), this "should be the overarching concern in language assessment". In terms of a new test, its introduction should lead to positive washback and this will be discussed in some detail in section 3.5.3. Scores should be interpreted and used

[41] Indeed 'cognitive validity' has been incorporated as an important part of Weir's updated socio-cognitive validity framework (see Taylor, 2013).

appropriately and all stakeholders should be well-informed about how decisions will be made. This inference therefore rests on assumptions that scores are sufficient and useful, thet the decision-making process is appropriate and that there are no negative consequences of introducing the test (Xi, 2008). Bachman (2005) and Bachman and Palmer (2010) argue that the uses and intended consequences of the test should be the starting point for test design and evaluation. They emphasise the need to justify test use rather than score meaning. The question is whether or not the scores support the decisions to be made and therefore score reporting practices are important. Meaning can be added to scores by referencing them to external benchmarks, such as the CEFR (Kane, 2012). Here, the role of setting cut scores on tests has an important influence and needs to be based on informed research. For example, collective judgements of the many stakeholders involved. Test consequences or impact studies need to be carried out, especially if the introduction of the test was specifically designed with intended positive washback in mind.

One major study which was guided by and subsequently built upon Kane's argument-based approach was that of Chapelle et al. (2008), who reported certain advantages of using the approach over other validation approaches (see Chapelle et al., 2010). One major advantage is there is not such a great need to describe the test construct, something which had previously caused great difficulty. The basis of the score interpretation is the IA, rather than the construct definition as such and the IA should include both competency and task-based approaches such as the one developed for the TOEFL test. This IA was also used by Aryadoust (2013) to develop the validity argument of the IELTS Listening test. Validation research can be prioritised and defined through a systematic process of examining the inferences in the IA and is a move away from traditional checklist approaches to validity, which do not give us any guidance on what type of evidence should be collected in a particular context. Research results can then be integrated into the presentation of the VA in order to show how the IA is supported by evidence. The TOEFL approach is considered to be "an important move in language testing away from the highly abstract unified model of validity" (Bachman, 2005, p. 17). However, the study has been criticised because the authors did not provide evidence for

the last bridge in the validity-argument concerning decision-based inferences as proposed by Kane (2002, 2004), Bachman (2005) and Bachman and Palmer (2010). The study did however introduce an extra link in an IA, that of a domain definition based on an examination of the TLU domain. As has been seen, this link is present in the ALTE validity argument approach for CEFR-related tests.

It is therefore clear that any validity argument must be supported by accumulated evidence that scores from a given test can be used to make the correct decisions. The observable attributes we intend to measure in a test are those which are necessary to perform tasks in the TLU domain. The TLU consequently needs to be clearly defined, which for the present study means the specification of a broad TLU for CEFR B2 listening proficiency in the context of school leaving and university entrance, and the inclusion of a range of possible tasks belonging to the domain.

3.5.2 Target Language Use Domain (TLU)

The TLU is defined by Bachman and Palmer (1996, p.44) as a "set of specific language use tasks that the test taker is likely to encounter outside of the test itself, and to which we want our inferences about language ability to generalise". In Kane's terms, the TLU is defined in terms of a range of tasks and rules for scoring responses; in order for an observable attribute to be well defined there must exist a clear specification of the TLU. The target domain specifies what is meant by the observable attribute, although it does not necessarily indicate how to assess that attribute (Kane, 2006). In developing a test, we have some purpose in mind; as Fulcher and Davidson have written "scores on language tests are used to make decisions, and test design needs to be closely aligned to the types of decisions that need to be made" (Fulcher & Davidson, 2009, p124). In the case of the present study, that purpose is the evaluation of CEFR B2 listening proficiency in the context of school leaving and university entrance. We must therefore begin by looking at this TLU and identifying some of the core skills and aptitudes associated with success for this construct before proceeding to develop a test that measures these attributes (Kane, 2013).

The TOEFL validation studies (Chapelle et al., 2008, p.2-3) felt that language proficiency should be the basis of score interpretation, even with the problems of defining such a construct. They purport that language assessment specialists agree on two things:

1. Language proficiency needs to be broadly conceptualised and we need a model of communicative competence which includes all aspects of communication such as strategy use.

2. We must take into account the context of use as the context affects the nature of language ability.

As such, a TLU needs to be conceptualised in terms of both context and use. Kane also highlights the importance of test content:

> much of the evidence needed to support the interpretations of test scores as measures of observable attributes may be generated during test development as the content domain for the test is specified, data collection procedures are defined, and samples of tasks and conditions of observations are drawn.
>
> Kane (2006, p.131)

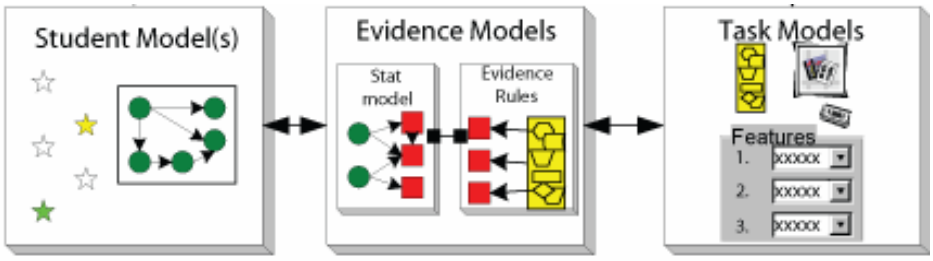
With this in mind, the TOEFL validation studies turned to tools provided by '*evidence centred design'* (ECD) put forward by Mislevy and his colleagues adding an extra inference to their IA—that of *domain definition*. Mislevy and colleagues (Mislevy, 2007; Mislevy & Haertel, 2006; Mislevy & Riconscente, 2006; Mislevy, Steinberg, & Almond 2002, 2003) proposed an evidence-centered design (ECD) approach to the design and development of assessments. In ECD, evidentiary reasoning is a key concept linking the intended score-based interpretations to different kinds of supporting evidence. Its aim is to establish a link between test developers' claims and the evidence supporting these claims. The validation framework comprises four stages: domain analysis, domain

modelling, conceptual assessment framework, and operational assessment. It has been argued, however, that this framework does not explicitly respond to test consequences and the social concerns of assessment (McNamara & Roever, 2006).

Chapelle et al. (2008) drew upon the first two stages of ECD, domain analysis and domain modeling, which provide a process to identify the theoretical underpinnings of a test—McNamara and Roever (2006, p.23) call this the 'thinking stage'. It is necessary to begin by asking what complex set of knowledge, skills, or other attributes should be assessed and what behaviours or performances would reveal those constructs and what tasks would elicit those behaviours (Mislevy & Haertel, 2006 p.16). This stage could be likened to a needs analysis study and could include investigations of documents such as the curriculum, textbooks used or typical language use situations in a particular context. In domain analysis we need to specify what we know about the TLU. For the present study, we would need theoretical backing of just what knowledge, skills and processes are used whilst listening, an undertaking which has largely been met with the literature review. In domain modelling, information and relationships drawn from the domain analysis are organised into three interrelated components: the student (or proficiency model), evidence models and task models. The student model lays out what the test designer wishes to measure expressed as variables that reflect test takers proficiency, the simplest model being a pass/fail decision. If primary purpose is proficiency, then there will be only one variable in the proficiency model; but if we also wanted to give diagnostic information, then we would need to include additional proficiency variables. Here, the student model could be represented by IRT modelling techniques (Mislevy & Riconscente, 2006). The task model addresses the context and is key to task design; decisions need to be made about just how the candidates performance will be captured. In the case of a listening test, for example, we could have MCQ items, open-ended items or a spoken response to some aural input. In language testing specifically, we could use Bachman and Palmer's (1996) taxonomy of task characteristics to establish a relationship between task characteristics and the TLU. The evidence model acts as a bridge between the two (see Figure 9). It has an evaluation component (evidence model) which tells us how the test will be marked or item scores and the measurement model which has

information about what statistics will be used to collect data across tasks. In the ECD framework, the layers of domain analysis and domain modelling require assessment developers to establish an assessment argument through clear documentation of the target domain.

Figure 9. Three central models of the conceptual assessment framework for evidence centred design. (Taken from Mislevy, Almond & Lucas, 2003, p.5)



This documentation is subsequently used to develop the conceptual assessment framework or CAF (which is to say, either the technical specifications for both tasks and the test, or detailed test specifications for providing the blueprint for test development). This framework contains a lot of technical detail and requires the development of detailed test specifications from which to develop the test. Chapelle et al. (2008) implemented these ideas for task design analysis in the TOEFL test and confirmed that it was helpful for clarifying just what is being measured in a test. They therefore added an extra inference to the IA, that of the domain description inference. Here, examples of evidence they include to support this inference rely on elements of the model at the test design stage, such as gathering corpus evidence of academic language (Biber, Conrad, Reppen, Byrd, & Helt, 2002).

Chapelle et al. (2008) thus added an extra 'stepping stone' to Kane's argument-based approach whereby *domain definition* becomes the first link between *target domain* and *observation*. Similarly, ALTE (2011) include this important link in their chain of reasoning for CEFR related tests, calling it the 'observation inference' —as can be seen in Figure 4 (page 16). However, Kane's final link (utilisation), considered by many to be

the most important, is not included in ECD. Following Chapelle et al. (2008) and ALTE (2011) the present study will include the domain definition/observation inference, as it is felt that this link is especially important for a CEFR-related test and that any such claims should be supported by evidence. Indeed, if the theoretical underpinnings of a test are not valid, the entire validity argument will fail (Aryadoust, 2013).

The ideas presented here are not new and ECD does not attempt to make traditional methods of test development obsolete. Rather, it provides a framework to formalise and document traditional test design processes in greater detail and to articulate the connections between elements of the test design more clearly, and is therefore particularly useful for measuring new constructs (Zieky, 2014).

Certainly, every test should produce detailed test specifications to guide test development (e.g., Alderson et al., 1995; EALTA, 2006). The test specifications provide the test 'blueprint' and help define the construct underlying the test as well as providing information about how these constructs will be tested. This allows test developers to make direct links between the theory on which the test is based and the test tasks (Alderson, 2000). Davidson (2012, p.201) states that specifications are supposed to be 'generative' and support the production of multiple, standardised items/tasks, so contributing to the production of parallel test versions. Normally, a number of versions of the test specifications would be produced for different audiences. Internal specifications would be specifically produced for test developers and alternative versions would be made available for the different stakeholders such as test-takers, test-users and teachers. An external evaluator of a test should be able to link tasks and items with descriptions of test content, that is to say with any claims of just what is being tested (Green, 2017).

Kane (2006) points out the role that judgements have to play in test constructs and proposed score uses. Content-related evidence about the relevance of observed performances to the proposed interpretation and use is largely produced during test development and tends to have a 'confirmationist bias', especially when judgements are made by the test developers themselves. Although content related evidence does not fully

justify the proposed interpretation, claims about construct relevance and representativeness are an essential part of overall validity. For example, Bachman (2005) highlights the 'relevance warrant' in a utilisation argument which should include both ability and performance in the TLU.

Kane (2006) argues that when developing a test and putting together a plausible and coherent IA, test developers must be creative. There is a potential for numerous measurement procedures to be developed, and it is up to the test developer to decide on appropriate measurement procedures for achieving the desired goal. As such, it is preferable that the test and the IA be developed simultaneously. Once a preliminary IA has been specified, test specifications can be developed to fit this IA, which includes inferences about proficiency in a skill in a certain domain. Such specifications would obviously include test tasks that require the use of the skill in the target domain.

In the present study, the TLU domain is represented by successful listening behaviour associated with CEFR B2 level school leaving/university entrance and the CEFR B2 listening descriptors outline the types of behaviour we would expect to find at this level. Topics and language use situations should be drawn from the personal, educational, academic and social situations (see CEFR p. 48-49) which would be typically encountered by L2 students of English in a university entrance situation. The present test is also meant to act as an achievement test for school leaving and thus the TLU domain should therefore both include the curriculum objectives of baccalaureate study, and be representative of baccalaureate classes (and hence of the course books used). All these concerns must be taken into account when drawing up test specifications and evidence should be provided to support the *domain definition* inference.

3.5.3 Test Consequences

As previously stated, test impact, consequences and washback are included in the 'utilisation inference': the use of the test needs to be justified and positive consequences should result. As the present study concerns a new test which is essentially being

introduced in order to create positive washback effects on teaching and learning within the Spanish education system, I will now look at this aspect in further detail. The belief that good tests bring about positive washback is supported by Messick (1996, p.247), who argues that positive washback can be associated with the introduction of more valid tests because "minimising construct-under-representation and construct-irrelevant variance in a test should facilitate good educational practices". Consequences (a term usually used in the measurement literature) along with impact and washback are all terms applied to the consequences of test use. Impact has been defined as the broad influences on individuals, education systems and society, whilst washback has been used to describe effects of testing systems on teaching and learning (Hamp-Lyons 1997; Hawkey, 2006). There are therefore both macro- and micro-considerations to test consequences. Following an argument-based approach to the implementation of a new test, including consequences as a part of validity means studies need to begin before the test is implemented. For example, Brown (2008) and Fulcher and Davidson (2007) advocate democratic testing which involves all stakeholders. Conversations with stakeholders are necessary, with the final design decisions about items and tasks being informed by the impact the test will have on stakeholders, thereby linking consequences to test design. Within the context of educational reform,[42] Chalhoub-Deville (2009) looked at the intersection of policy, validity, and impact in U.S schools in terms of 'social impact analysis'. She calls for all key stakeholders to be involved in the process of test development, arguing that there is a better chance that policy goals will be achieved if stakeholders work together proactively to inform educational reform. The evaluation of score uses requires an evaluation of the consequences of the proposed uses; negative consequences can render a score use unacceptable (Kane, 2013), and negative consequences should be anticipated within the social impact analysis framework. Similarly, Saville (2012), using the concept of 'impact by design', argues that test developers should anticipate the scenarios which will result from the implementation of a new test and try to mitigate negative impact.

[42] In particular the No Child Left Behind Act (2001).

Bachman (2005) emphasises the role of test uses and consequences in his AUA and believes that while test consequences form a major part of any validation endeavour (e.g., Messick 1989, Kane, 2002, 2012), there has been no systematic method to link score interpretations to test consequences. He argues that validity concerns about test usefulness, fairness issues, social consequences and impact have been addressed separately to a VA. Building on the work of Kane (1992, 2002) and Mislevy et al., (2003), Bachman proposes that test developers should use the argument structure not only for validity inferences (following Kane) but also for the uses of assessment. Pardo-Ballester (2010, p.140) likens the AUA to a meta-structure which allows for the consolidation of test design, development, scoring interpretations and intended uses within a single model and as such prioritising the consequences of a test use. Using a Toulmin approach (see section 3.5.4), Bachman (2005) and Bachman and Palmer (2010, p.103) articulate a utilisation argument in terms of the claims about test scores, which is essentially the decision we want to make, by defining four types of warrants which will differ from those in the VA. These warrants are:

1. Consequences of using an assessment are beneficial to stakeholders. An assessment will affect students, educational programmes and teachers. Washback should be considered as part of test consequences.

2. Decisions take into account values in society and are equitable.

3. The interpretations about the assessed ability are, meaningful, impartial, generalisable, relevant and sufficient. The interpretation of the score should be relevant and meaningful to the decision to be made and assessment tasks should correspond with those in the TLU in terms of both ability and performance. This warrant would also include the extent to which score information is conveyed in terms that test users can understand and relate to (Bachman & Palmer, 2010, p.114).

4. Assessment records are consistent.

By addressing consequences in this way, Bachman and Palmer (2010) believe that we follow a process of 'assessment justification', as each AUA should be tailored for particular local contexts to guide individual test development projects. This is important as "assessment development and use, and the process of justification are necessarily local" (Bachman and Palmer, 2010, p. 438). That is, we cannot have a one size fits all approach; rather, the whole process is context specific. However, Kane (2006, p. 8) questioned the extent to which all consequences of test use should fall under the heading of validity and Cizek (2016) goes as far as to argue that test score interpretation and justification for test use cannot be part of the same validity argument. Messick (1996) states that washback only impacts validity if it occurs as a result of test implementation and not other aspects relating to the educational system. Bachman's AUA (2005), however, extends Kane's validity argument approach and includes positive washback or intended consequences as a warrant which needs backing. However, he does qualify this by recognising that intended test consequences are not simply brought about by the test but also its influence on teaching, learning, the education system and society. His model clearly defines the responsibilities of test developers to ensure positive consequences.

Previously, impact studies have tended to be conducted separately and have not been included in the unified view of validity, and although they certainly provide evidence for test validity, they only really strengthen a validity argument once they are integrated to support a conclusion (Xi, 2008), making Bachman's AUA an important advance. In the context of the present study, the main aim of introducing the new test is to bring about positive washback in the educational system. Yet it may be the case that not all test consequences are either positive or intended. Indeed, for the present study we can only examine intended consequences on teaching and learning, such as stakeholder opinions about future washback, and actual consequences can only be studied once a new test is up and running—perhaps after a baseline study has been carried out. As previously mentioned, a few studies involving stakeholders have already been conducted in Spain, and for example we already know that students (as important stakeholders) believe that the present system needs to be changed (Fernández Álvarez, 2007). Despite the fact that most washback studies are carried out once a test is operational, considerations of

stakeholders can make contributions to a test before its implementation. With this in mind, one aspect of validation research for a new test would necessarily have to include dialogue with stakeholders.

Regarding the teaching and learning context, the new BFE is specifically planned at a micro level to be introduced by the Spanish education authorities in an attempt to achieve beneficial washback and respond to the real linguistic needs of Spanish students in this new communicative scenario (Amengual-Pizarro & Méndez García, 2012). It is to be hoped that including communicative competence in the four skills in test materials will mean that those teachers who teach to the test will necessarily change the content and style of their teaching. The hope is that the changes made to teaching practices will bring about a more communicative English language classroom in secondary schools. Certainly, the teachers surveyed cited the current PAU test as one main reason why oral skills were being ignored in the classroom (Amengual-Pizarro, 2009). Fernández Álvarez (2007) found that 63% of the high school teachers who were surveyed in his study declared that they spent at least one class per week on preparing their students for the PAU test. The washback effect has a direct impact on teaching (Wall, 2000), especially when it originates from a test which is directly related to university admission. If a new test has provided a construct of communicative competence, such a construct would necessarily be encouraged in the classroom, as it will influence "attitudes to the content, method, etc. of teaching and learning" (Alderson & Wall, 1993, p.120). The teacher is therefore an important stakeholder, and we would also need to consider their opinions about any new test.[43] In fact, it has been argued that language teachers themselves need to be involved in the test development cycle, as more positive washback can be promoted when the teachers are involved in any aspect of the test design process (Turner, 2001). Teachers feel themselves to be accountable for their students' marks and evaluated on their success, and as a result feel a strong pressure to teach to the test (Garcia Laborda, Gimeno Sanz & Martínez Sáez, 2008; Luxia, 2005). In turn, there further exists pressure on teachers from parents demanding that teachers prepare their children for a test (Choi,

[43] As previously mentioned studies canvassing teacher's views on test content have already been carried out in Spain (García Laborda & Fernández Álvarez, 2012)

2008). Certainly, many children in Spain attend extra-curricular English classes to prepare for Cambridge and Trinity exams, and many parents want their children to have an English language qualification. Parents would therefore be another stakeholder whose opinions should be canvassed.

Planned positive washback in the teaching and learning context needs to be empirically researched to see if intended washback has actually been achieved (Alderson & Wall, 1993). This type of impact is normally measured using longitudinal studies, starting from a baseline study (e.g., Hawkey, 2006; Wall & Hora´k, 2006). Many washback studies have shown how tests have brought about 'micro' changes including changes in the content being taught in classrooms, teaching methods and style, and teachers' perspectives about a test (e.g., Alderson & Wall 1993; Cheng 2005; Green 2007; Hawkey, 2006). These changes have been explained by the high-stakes nature of the tests under investigation. Some recent impact studies offer a comprehensive overview of the effects which *Cambridge English: Young Learners* is having on bilingual schools in specific areas of Spain. For example, Ashton et al. (2012) looked at policy maker intentions, exam implementation and reception by teachers, learners and parents. However, the published findings of this study have certain limitations. First, the results of this research are mainly obtained through questionnaire data, which entails the risk that teachers' own opinions are not accurately reflected. Secondly, the study focuses on the Bilingual English Development and Assessment (BEDA) project carried out by Cambridge English in conjunction with the Federation of Religious Schools in the Madrid area (FERE Madrid) since 2008. It therefore covers just one specific programme in which Cambridge English was involved from the outset, and is therefore not entirely representative of the country as a whole.

Froetscher (2016) found that the transition in Austria from a teacher- developed exam to a standardised, professionally-developed national school leaving exam had a clear effect on classroom teaching and recommended that any assessment reforms should therefore consider the washback implications involved. If teachers 'teach to the test', such a test should clearly represent those competencies which we wish our students to

obtain. As Messick (1996, p. 241) claims "ideally, the move from learning exercises to test exercises should be seamless".

One important consideration in the present study would therefore be how the test would affect the teaching and learning of listening in schools in Spain, as Wagner points out:

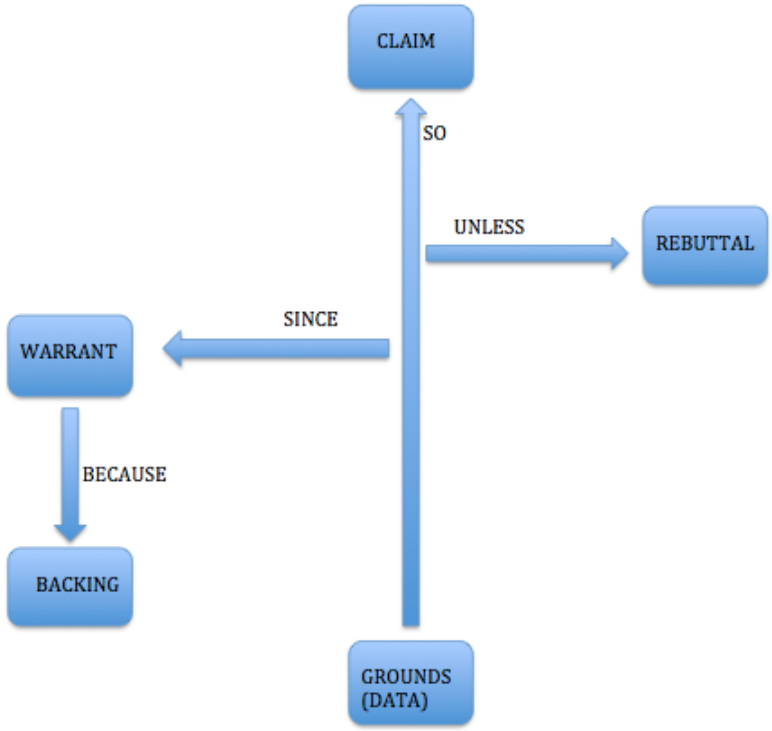
> it seems obvious that teachers and testers should be interested in L2 learners developing the ability to listen to and comprehend authentic spoken discourse, which usually includes things like connected speech, reduction, phonological modifications, vernacular language, language variation, and nonverbal communication.
>
> (Wagner, 2013a, p.13)

The development of such focuses is precisely what the present study will propose; the test construct must be broad enough to promote teaching to the test which includes all the necessary competences for successful listening. In short, the more the test reflects real-life listening tasks, the more beneficial the washback will be (Vandergrift & Goh, 2012).

3.5.4 Toulmin's approach to logical reasoning

The evidence supporting practical arguments needs to address (1) the appropriateness of various lines of argument in specific contexts, (2) the plausibility of assumptions, and (3) the impact of weak assumptions on the overall plausibility of the argument (Kane, 1992, 2002, 2012) and here Toulmin provides a convenient framework and an established vocabulary for discussing interpretive arguments (Kane, 2004). Indeed, Kane, Bachman, Bachman and Palmer, Chapelle et al. as well as Mislevy and colleagues all follow Toulmin (1958/2003) in using evidentiary reasoning in their validity frameworks. Consequently, Toulmin's criteria for evaluating practical arguments will briefly be outlined below.

Figure 10. Toulmin's approach to logical reasoning



Toulmin articulated a model of argumentation using a chain of reasoning in order to build a case for a particular conclusion; the person making the claim has an obligation to provide relevant supporting evidence. Toulmin (2003, p.8) states that "a sound argument, a well-grounded or firmly backed claim, is one which will stand up to criticism, one for which a case can be presented coming up to the standard required if it is to deserve a favorable verdict". Toulmin describes the structure of an argument as the sum of a claim, grounds (data), warrants, backings and rebuttals as shown in Figure 10. We state our claims and then evaluate the credibility of those claims and the inference we want to make must be plausible or supported by evidence (Kane, 2012). Conclusions drawn about test takers are referred to as *claims* because they state the claims that the test designer wants to make about a student (for example, the statement 'the test taker has CEFR B2 listening proficiency' would be a claim). Claims are made on the basis of data or observations that Toulmin referred to as *grounds (*for example 'the test taker has passed the BFE CEFR B2 listening exam'). *Warrants* are general statements or

assumptions, specific to the context, stating how and why the data support the claim (for example, ‘ People use CEFR B2 competences to solve test items’). *Backings* are general statements which support the warrants (for example ‘Test tasks were developed based on detailed CEFR B2-related test specifications or experts agree that test tasks are representative of the CEFR B2 listening domain’). A *rebuttal* is a challenge or a counter claim, an alternative explanation (for example ‘In order to solve test items test takers use test wiseness strategies, not CEFR B2 competencies’). Possible rebuttals to a claim would therefore constitute weaknesses in an IA and could be formulated as ‘What if ?’ type questions which would then form the basis of validity evidence studies in order to present a validity argument for any given test. In the above example, research would be necessary to discover if indeed test takers use test-wiseness strategies or not in order to refute the rebuttal. In this framework, any assumption which is questionable must be supported by evidence (Kane, 2012). It is the way that the IA is specified which makes clear how the validity argument can be questioned or what research is necessary in order to refute rebuttals (Aryadoust, 2013). Such an argument could be made at any level of the test development process and indeed, Fulcher and Davidson (2007) state that we can produce a validity argument at item-level, task-level or test-level. Mislevy et al. (2002, 2003) propose that grounds for an IA should include a statement of the students’ performance and a statement of the task characteristics used to elicit the performance, thereby reconciling task-centered and competency-centred approaches to test design (Chapelle et al., 2008).

A warrant is defined as a general statement which provides legitimacy of a particular step in the argument (Toulmin, 2003, p. 92), it is the link between the claim and the grounds. Once a warrant has been well supported and rebuttals have been refuted, the inference we want to make can be accepted. Kane (2006) says that a warrant can be viewed as a ticket which permits the crossing of an inferential bridge in the IA; the ticket must be valid and it is the backing that validates the ticket. Backing can be based on theory, prior research, or evidence collected during our validation process to support our warrant. Using this approach to providing the validity argument for a test brings us an advancement in professional knowledge by giving us both guidance and a conceptual

infrastructure to reach conclusions about test score interpretations and uses (Chapelle et al., 2010).

3.5.5 A validity argument approach in practice

Most studies about language assessment are concerned with some aspect of test validity. However, a review of test descriptions by Cizek, Rosenberg and Koons (2008), concluded that reported validity information does not follow current views about validity, with consequential evidence in particular noted as lacking. In other words, developments in educational measurement have not followed through to language testing practices. Despite this reported gap, however, a number of argument-based validation studies have in fact been carried out and these will be reviewed below.

Here I will present a number of studies which have used the approach to show how test purpose should be the guiding force for asserting claims and directing the specific types of validity evidence to be collected. Examples will be given to illustrate just how an argument-based approach to test validation is applied in practice. As the process is context specific, judgements about the justification of test use are also local (Bachman and Palmer, 2010, p.438) and will be influenced by a number of contextual factors, such as the types of stakeholders, the availability of resources, and value systems in society. Each test validity argument will be unique and depend on test use and the decisions to be made within a given context. The presentation of these studies is useful as the review is aimed at visualising the implementation of the argument-based validation approach and the types of evidence which can provide backing for warrants in an IA.

Perhaps the most detailed study is that of Chapelle et al. (2008), who adapted Kane's IA to develop the TOEFL ibt interpretive argument. This study provides a model for the future argument-based validation studies in language testing and assessment. Chapelle et al. applied this approach while they were still designing and trialling the test, before it went operational, calling this stage of test validation 'design validity'. They used an interactive approach, developing IA's as the test development process progressed and

gradually adding evidence in order to produce the overall validity argument. They adapted ECD, with specific reference to the skill of listening in order to form the basis for their task design analysis and finalise their test specifications—and ultimately the test blueprint. As the unnaturalness of the previous TOEFL listening tests had previously been criticised as a poor representation of natural speech, domain definition drew on research from academic corpora, so as to improve the authenticity of the language used in sound files.

Chapelle et al. also investigated the impact of using non-native accents (Major et al., 2002), as logically the TLU situation would include interactions with non-natives, but concluded that such a move has potential for introducing bias. Overall, their validation project found that specifying a task framework was useful for generating new relevant task types.[44] Note-taking was introduced as being more representative of the TLU after a study by Carrell, Dunkel, and Mollaun (2002, 2004), who also studied the effect of topic and length of sound file. In order to provide evidence for the explanation inference, an investigation explored tasks which could assess pragmatic understanding and integrating information, two important abilities which were included in the new construct definition. Some of the prototyped integrating information type tasks showed themselves to be testing memory and would not have supported the validity argument, and these item types were subsequently dropped. Some studies were carried out concerning speech rate and sentence structure (fragments versus complete grammatical sentences) to see how they affected task difficulty. Results showed that speed of delivery and sentence structure did not have any significant effects. Throughout the design stage, theoretical analysis and empirical data were used to serve as backing for inferences in the IA and the main advantage of the approach was reported as being the guidance given to the type of research which was necessary.

Wang, Choi, Schmidgall and Bachman (2012) applied an AUA framework to the Pearson Test of English Academic (PTEA), a relatively new test which aims to measure

[44] However, one of the proposed task type was found to be difficult to produce and was subsequently dropped from the test specifications.

communicative English language proficiency. Their analysis, however, is thwarted by the fact that the publishers present validity evidence of a general nature, rather than with a specific test use in mind. However, even though this was a retrospective study, it showed that structuring existing research findings using an IA framework can "organise the evidence and its implications" (Chapelle et al., 2008, p. 23). Types of evidence presented for *meaningfulness* are an analysis of test specifications, where test task characteristics must be well specified and detailed item writer guidelines have been produced. A warrant for the test administration procedures was supported by evidence from test taker feedback (Zheng & De Jong, 2011), in which claims that the test/task instructions are understandable and that problems with test administration have been addressed are supported. However, these results—while stated—were not actually reported in Zeng and De Jong (2011). A warrant stating test tasks engage the test takers in the relevant knowledge, skills and abilities (KSAs) was supported by statistical evidence allowing for the removal of both poorly performing items and items where non-native speakers outperformed native ones (Zheng & De Jong, 2011). However, these results are taken from two large pilot field tests (administered in 2007 and 2008) and statistical analysis is not reported to be part of an ongoing cycle. The warrant that scores can be interpreted as indicators of the construct was supported by correlation evidence of scores between PTEA and two tests (TOEFL iBT and IELTS) which aimed to measure the same construct, also reported in Zheng and De Jong (2011). For the *generalisability* inference (this would be the extrapolation inference in Kane's framework), the warrant that test task characteristics match the TLU (in this case in English university academic setting) was supported by evidence from test specifications and the use of authentic tasks. For the *relevance* inference, the warrant that information provided is relevant to academic admissions and for professional and government organisations when a CEFR level is required was supported by evidence of the CEFR linking process, having used both a test taker-centred and an item-centred approach. This claim is reported by Wang et al. (2012) to be supported by the Zheng & De Jong (2011) study on page 32, yet no mention is made of CEFR linking. A CEFR linking study has, however, been carried out for the PTEA and is reported in Pearson (2010). For the *decisions* inference, the claim states that PTEA scores are equitable and sensitive to educational and societal values, although

there was no evidence to support such claims. For the *consequences* inference, the claim is that using the PTEA scores is beneficial to stakeholders and society. Evidence is provided from interviews and focus groups with test takers who reacted positively. For the warrant that test scores are reported in a way that provides clear and understandable information to all stakeholder groups, evidence comes in the form of the documents provided by Pearson to aid interpretation of the score report. For the warrant that the PTEA test has beneficial washback in the classroom, evidence is provided by stating that item types are authentic and require integrated language skills for communication. However, such a claim may be strongly rebutted as no evidence has been collected in language classrooms. Overall, the validity argument for PTEA reported by Wang et al. highlighted a number of potential rebuttals and recommended that future studies be carried out to address these; therefore, the framework was useful in providing guidance on the types of validity evidence which needed to be collected.

Aryadoust (2013) used the six stage IA framework approach proposed by Chapelle et al. (2008) to provide the validity argument for the IELTS listening test. He proposed a Rasch-based VA using Rasch analysis to investigate each of the five inferences, citing economy as one of its most important advantages. For example, Rasch measurements can be used by examining the item variable map for construct representativeness/coverage or using a principle components analysis to discover whether a test is unidimensional and does not contain construct irrelevant variance, thereby informing the explanation inference. However, Aryadoust does state that Rasch alone is not sufficient for the construction of a VA. While the present study will draw upon some of the uses of Rasch measurement employed by Aryadoust, they will not form the complete basis of the VA and other research methods will also be used in order to further strengthen any claims made.

Schedl (2010) followed an ECD approach to revise the TOEIC reading and listening test. This project began with construct identification, in an attempt to identify those KSAs candidates would need in the real world. Subsequently, they tried to incorporate these in order to improve the old version of the test so that more information could be provided

about candidates. The ECD approach employed here allows performance on test items to be linked to evidence about candidates' language abilities. The test constructs were clearly defined and articulated in terms of abilities,[45] which are written in terms of claims, and tasks were then developed to link to the required abilities. This approach is a useful example for the present study, for which listening abilities are defined in terms of 'can do' statements both in the CEFR and in the Spanish national curriculum. The language abilities underlying the claims were identified on the basis of current language theory and research. The redesign team then looked at variables thought to affect the difficulty of performance on test items measuring these abilities. Prototype items were piloted on small groups to discover candidate reactions. In the TOEIC test, varied accents were not found to be problematic and different task formats were used to allow candidates to express their preferences. The test design therefore takes into account the opinions of the test takers, arguably the most important stakeholders.

The argument-based approach can also be applied to low stakes tests; indeed, any test which has intended decisions can be justified using the approach. Llosa (2007) describes a study which built a validity argument for the interpretation and use of a classroom-based assessment aligned with state-mandated proficiency standards. In Chapelle et al. (2010) an argument-based approach was used to examine a computer-delivered test of productive grammatical ability. This test is at the development stage and so the researchers do not address the utilisation link, but they do highlight the use of both qualitative and quantitative methods to support the other inferences in the IA.

Pardo-Ballester (2010) applied an AUA to a new online listening test, the Spanish Listening Test (SLE), a university placement test. Warrants specifically addressing the qualities of consistency, construct validity and authenticity were proposed based on Bachman and Palmer's (1996) test usefulness framework, but not as such strictly adhering to the AUA approach. Neither did the authors address the utilisation argument.

[45] For example, for listening, claims were based on types of listening such as 'Examinee can understand details in talks and conversations on workplace and social topics and in descriptive sentences about photos'.

Nevertheless, they did show how a Toulmin approach could be used successfully to guide research decisions for the provision of relevant evidence.

Chapelle, Jamieson, & Hegelheimer (2003) conducted an argument-based validity study of a low stakes web-based ESL test. They illustrate well how test purpose can be used to identify sources of validity evidence. Like Pardo-Ballester (2010), they employed Bachman and Palmer's test usefulness framework. The study demonstrates how testing consequences may be taken into account through the direct integration of intended impact as an integral part of test purpose at the design stage. The authors also examined both positive and negative attributes of the test; here, using a cyclical approach allowed any negative attributes to subsequently inform future steps towards improving the test.

Wang (2010) used Bachman and Palmer's AUA framework to justify an additional use of a college level proficiency test. Due to the change in use of the test, new warrants had to be identified. The study resulted in a recommendation that test developers should focus on identifying and addressing construct-irrelevant variance. This study highlights the fact that tests cannot be used for a purpose other than the one they were designed for; if a new use is introduced, a new validity argument should be constructed.

Jia (2013) employed a AUA approach to investigate the claims about the GSLPA Spoken Language Test, a graduate programme exit test in Hong Kong. Their approach was found to be practical, and is described as 'a powerful framework' which guided the justification process well due to its clear articulation of exactly what types of evidence need to be collected for which claims or warrants. However, the caveat was made by Jia that the size and complexity of a justification study may provide a big challenge for a single researcher. Indeed, validation studies for high stakes tests can be laborious and costly and expertise is required to carry them out.

As well as the lessons which can be learned from these studies, in an attempt to guide test developers and researchers, Kane (2001, p. 330) gives useful advice by outlining the steps to be taken when applying an argument-based approach to a test validation project:

1. State the proposed IA.
2. Assemble all available evidence relevant to the inferences and assumptions in the IA to give a preliminary VA. This should alert us to the most problematic areas of a test or to any weaknesses in the IA.
3. Evaluate the weak and problematic assumptions empirically and/or logically. At this stage we may reject or improve the IA.
4. Restate the IA and VA and repeat step 3 until all the inferences in the IA are plausible.

Chapelle (2012, p. 26) specifically addresses lessons learnt from using an argument-based approach to language assessment and summarises how issues have been taken into account by this approach along with implications for language assessment. These implications for test developers will be taken on board in the present study.

3.5.6 Summary

Contemporary views see validity as a unitary concept, that is to say, as an integrated unified argument provided about a given assessment to support the intended interpretations, uses, decisions and consequences. Validity evidence must be integrated into a coherent argument which supports the intended test uses. Here, the argument-based approach to test validation provides a clear and sensible guiding framework (Chapelle, 2012), moving away from traditional checklist approaches and allowing for multiple sources of evidence using both qualitative and quantitative techniques to be appropriately integrated, thereby strengthening the validity argument. As Bachman and Palmer (2010) state, previous validation work has been heavily based on quantitative measurement, and there is a need for other sources to be included if we are to justify the use of language tests and convince stakeholders. Kane (2001, p.328) states validity issues are concerned with "the adequacy and appropriateness of the interpretations and the degree to which the interpretation is supported by the collected evidence". Indeed, as Kane et al. (1999, p.15) point out "the overall (validity) argument is only as strong as its weakest link". First a clear IA needs to be articulated in order to decide what evidence needs to be collected to

provide backing for any claims made. Only if sufficient evidence for each bridge in the IA is provided, can test scores be seen as valid in terms of the domain of interest.

Despite the provision of frameworks outlining the network of inferences, each particular assessment context will need to identify the pertinent assumptions to make sure claims follow logically from the specified assumptions (Xi, 2008). In language testing, there is increased awareness and concern not only about test consequences and impact, but also about ethics and fairness (Bachman & Palmer, 2010). The goal of providing beneficial consequences plays a major role and should itself be considered as a warrant and be backed by evidence in any validity argument, though it must be remembered that intended consequences cannot be brought about by the test alone and it may be difficult to argue that all test consequences are part of the test quality.

The validity argument has an audience (e.g., test takers, government bodies and institutions which accept the test), and must be plausible to convince this audience (Chapelle, 2012). The introduction of a new test may be resisted and so the active involvement of stakeholders in the test development process is recommended. In the present context, such resistance to change has already been mentioned and it is hoped that by providing sound validity evidence for the proposed new test, conversations with stakeholders will be more productive and contributions can be taken on board in the future.

The present study is based on a test which needs to be adapted to both curriculum and European standards. Such a criterion-related assessment will obviously require a validity argument that very much takes these standards into account. Both the test development cycle and the test validation process are cyclical and iterative, where test specifications are not set in stone and can be updated as and when new evidence comes to light or stakeholder demands are taken on board.

I will now proceed to develop both test specifications and an initial version of the proposed BFE listening test. A VA approach will be used to guide the study and to this

end, the IA will be outlined and research questions will be developed using a Toulmin logical reasoning approach in order to determine the evidence which needs to be collected so as to present the final validity argument for the test under proposal.

Chapter 4. Conceptual framework and research questions

Any test development project must begin with an initial plan, which will then be used to produce the assessment instrument (Bachman & Palmer, 2010). Consequently, I will now present the test specifications for the proposed test, which will act as the test blueprint (Alderson et al., 1995) and allow me to operationalise the BFE listening construct, thereby producing a version of the test which will act as the measurement instrument. I will then go on to present the initial BFE interpretative argument and consider possible rebuttals, which will be presented as a series of 'what if ...?' type questions; these rebuttals will subsequently become my research questions and guide the research necessary for defining and presenting the BFE validity argument. This chapter will therefore present both the test and its theoretical underpinnings, which are drawn from substantive theory. The initial IA for the test will also be outlined and Toulmin's logical reasoning approach will be used in order to decide upon my research questions.

4.1 Phase 1: Development of test.

As has already been argued, test design is central to test validity. Following both the recommendations of ALTE (2011) and an interpretative argument approach to the BFE test, the first requirement is the identification of the target language use domain (TLU), in order to produce a detailed domain analysis and, subsequently, the test specifications. This process will draw upon my literature review in order to provide the theoretical backing for the observation inference and answer the questions 'what to observe and how?' posed by ALTE's chain of reasoning (Figure 4, p.45). Test specifications can then be drawn up to provide the blueprint for task and item development. As previously stated, this project has considered the importance of creating positive washback and it is therefore hoped that the presentation of detailed test specifications would be useful for stakeholders to have a clear understanding of how listening ability will be assessed. Detailed test specifications should be available to stakeholders if transparency is to be achieved, an important consideration according to EALTA guidelines (2006). The format of the test needs to be outlined and the listening processes which are intended to be elicited by the test tasks should be stated. Following Green (2017), if we make the construct accessible, an external evaluator should be able to link test tasks and items to the description of the test. Green also highlights the fact that test specifications are not set in stone and their development should be viewed as iterative, dialogue with stakeholders should be encouraged and final test specifications should be consensus based. It is therefore emphasised that the present study provides initial test specifications for the proposed test and that these could clearly be further debated. As an experienced item writer who has worked on high-stakes CEFR related tests in Spain, the UK and the USA, I decided to develop the test specifications and tasks myself and a detailed description of this process follows. The resulting test will be the measurement tool, which will then be evaluated as to its validity following the test's IA using the results of my research questions.

4.1.1 Test Specifications

The measurement instrument is intended to assess CEFR B2 listening ability for school leaving/university entrance and the test specifications must be drawn up to reflect this construct. The specifications are 'generative' (Davidson, 2012, p.201), but should include enough detail to support the production of comparable standardised test versions, which will in turn allow for score consistency. By standardising testing conditions such as, settings, time limits and instructions, we narrow the universe of generalisation and enhance generalisability (Kane, 2013). A key requirement is the identification of tasks which reflect and measure CEFR B2 listening ability in the TLU domain, and it is therefore considered important that CEFR B2 listening descriptors be included in the test specifications. If a test taker is granted B2 listening proficiency as the result of taking a test, the test provider is in effect claiming that that the test taker is capable of performing those descriptors beyond the limits of the test. The test is criterion-referenced and as such the criteria to be assessed will dictate the characteristics of the test tasks (Wagner, 2013a). The CEFR 'can dos' specifically describe the types of communicative listening activity to be expected at each proficiency level and should be adhered to at the task development stage; thus, the CEFR descriptors will be embedded in the test construct. In the BFE test specifications, it is these descriptors which will therefore provide the specific purposes for test tasks, and which should be well sampled on each test administration in order for the test to have good construct coverage. Indeed, in order to sample the construct well, it was decided that the test should include four tasks and have a total of 28 items. Such a number of items should mean that the test is long enough to provide good construct coverage and be a reliable measure of listening ability (R. Green, personnel communication, April, 2011). Different task types should be included in order to minimise the task effect (Alderson et al., 1995) and allow the test takers to have a number of new opportunities to restart. In order to provide enough input, redundancy and context it was decided that each sound file should be between 3 and 5 minutes long, should be played twice, and that the total test should last approximately 35 - 40 minutes.

4.1.1.1 Domain Analysis

As previously noted in the literature review, a number of important decisions need to be taken regarding the sound files to be used and the types of task to be developed in order to elicit the listening behaviours outlined in my listening ability model (Figure 6 p.67). These decisions will, in effect, be a representation of the *student model* using terminology from evidence-centred design, and will lay out the knowledge, skills and processes which will be assessed.

In accordance with my literature review, the first important decision I have taken regarding the sound files to be used is that only one-way transactional listening will be assessed and that the mode of delivery will be audio only. Likewise, all sound files must be authentic, either sourced from the Internet or produced intuitively by speakers in response to prompts in order to obtain samples of non-adapted natural speech. General focuses will include the different types of listening which have been outlined in the literature review. Again it is hoped that such practices will create positive washback, as teachers will need to focus on these different listening skills and so provide students with the 'core skills for listening' outlined by Vandergrift and Goh (2012 p.169). To this extent, we will be able to help facilitate a seamless transition from classroom activities to test tasks.

Besides the general guidelines provided by the CEFR—which gives examples of source materials from public, personal, educational and occupational domains—it was decided to consult the baccalaureate curriculum and textbooks used in these classes. This practice follows Kane (2006), who argues that in order to ensure that a test contains a representative sample of the TLU domain, a serious effort should be made to analyse that domain. The listening ability model has already been provided by the extensive literature review, which also includes guidance concerning contextual features of test tasks. Clearly, a balance also needs to be struck between the two functions of the test:

1. An achievement test for school leaving

2. A CEFR B2 proficiency test for university entrance

The new school curriculum for baccalaureate study (BOE, 2015) includes all aspects of communicative competence, and as such reflects the CEFR. The curriculum content for listening specifically mentions those meta-cognitive strategies which have been included in the present BFE listening ability model (see BOE, 2015, p. 442), along with assessment standards which, as previously stated, do not seem to be aligned with a specific CEFR level. However, the first descriptor under the heading 'evaluation criteria' does resemble the CEFR B2 global descriptor for listening (CEFR, p. 26). However, as already stated, some of the evaluation criteria for listening are more reflective of a level higher than CEFR B2 listening proficiency (descriptors mention the understanding of animated conversations between various interlocutors as well as understanding irony and humour—neither of which belong to B2 or below proficiency).

Topics covered by the various English baccalaureate textbooks examined were found to share a high degree of similarity, and are designed to be of interest to 16-18 year olds.[46] The listening exercises found in the textbooks examined included a variety of the listening types already mentioned, such as *main ideas* and *specific information*. However, one notable feature of all the textbooks examined is the fact that listening audios were not authentic, instead they followed the traditional format of scripted recordings performed by actors. As argued in the literature review, this is an issue which would have to be addressed. Drawing on the literature review the final specifications are presented below in Table 2.

[46] Examples of topics covered include: travel, shopping, law and justice, diet, education, relationships, personality, work, special occasions, places, technology, news and crime, sport, beliefs, music.

Table 2. BFE test specifications (version 1)

General Purpose	To establish CEFR B2 listening ability for school leaving/university entrance.
Specific Purpose/Test Construct	**CEFR B2 Descriptors:** **OVERALL LISTENING COMPREHENSION** (CEFR, p. 66) 1. Can understand the main ideas of propositionally and linguistically complex speech on both concrete and abstract topics delivered in a standard dialect, including technical discussions in his/her field of specialisation. 2. Can follow extended speech and complex lines of argument provided the topic is reasonably familiar and the direction of the talk is sign-posted by explicit markers. **UNDERSTANDING CONVERSATION BETWEEN NATIVE SPEAKERS** (CEFR, p. 66) 3. Can with some effort catch much of what is said around him/her, but may find it difficult to participate effectively in discussion with several native speakers who do not modify their speech in any way. **LISTENING AS A MEMBER OF A LIVE AUDIENCE** (CEFR, p. 67) 4. Can follow the essentials of lectures, talks and reports and other forms of academic / professional presentation which are propositionally and linguistically complex. **LISTENING TO ANNOUNCEMENTS AND INSTRUCTIONS** (CEFR, p. 67) 5. Can understand announcements and messages on concrete and abstract topics spoken in standard dialect at normal speed. **LISTENING TO AUDIO MEDIA AND RECORDINGS** (CEFR, p. 68) 6. Can understand most radio documentaries and most other recorded or broadcast material delivered in standard dialect and can identify the speaker's mood, tone etc. **IDENTIFYING CUES AND INFERRING** (CEFR, p.72) 7. Can use a variety of strategies to achieve comprehension including listening for main points checking comprehension by using contextual clues. (This descriptor should be used to complete all the tasks on the test). **General focus: Type of listening**

	Gist (G) Main ideas with supporting details (MISD) Specific information and Important details (SIID) Listening to infer (propositional) meaning (IPM)
Target Population	Second year baccalaureate students who are planning to enter university. 17/18 year olds, both sexes, from various socio-economic backgrounds.
Test Length	Approximately 35- 40 minutes
Number of tasks	Four
Number of items	**28** - one point per item, all items carry equal weight. (Spelling and grammar mistakes will not be taken into account on constructed response tasks)
Mode of Delivery	Only audio, each sound file to be played twice.
Audio files	B2 level authentic sound files sourced from the Internet or constructed using 'prompts' in order to obtain samples of real non-adapted speech which reflect the real life listening event. - 3 to 5 minutes each - International English (Including one second language speaker) - Speed of delivery approximately 180 words per minute. - Topics taken from Personal, Public and Educational domains (see CEFR p. 48-49) to be relevant as well as accessible, interesting and motivating for the target population. No culturally/technically or academically specific topics should be used. No topics which may cause offence or emotional distress should be used. - Discourse type: narrative, descriptive, argumentative, problem / solution, expository, and persuasive. - Both monologues and dialogues. - Content should be both concrete and abstract
Test method	1. Multiple match 2. Multiple choice (4 options) 3. Multiple choice (4 options) 4. Note form (including table completion) See separate task specifications.
Task rubrics	Written in English at CEFR B1 level and should include an example for each task.
Task specifications	
Task 1	
Task description	Short utterances (approx. 30 seconds each). This could be different speakers giving an opinion on something or one speaker answering questions about a given topic.
Response format	Multiple Match (MM). 6/7 items with 1 extra distractor.
CEFR Descriptor(s)	1 and 3
Type of listening	Main focus: G / MI
Rubrics	1) Listen to some people talking about ? Choose the correct speaker (**1-7**) for

	each sentence (**A-I**). There is **one** extra sentence that you do not need to use. There is an example (0) at the beginning. 2) Listen to ? answering questions about ? Choose the correct answer (**1-7**) for each question (**A-I**). There is **one** extra question that you do not need to use. There is an example (0) at the beginning. First you have 45 seconds to study the questions. Then you will hear the recording twice.
Task 2	
Task description	Monologue or Dialogue. This could be a person giving a talk or presentation or a conversation between 2 speakers.
Response format	MCQ – 4 options. 7/8 items
CEFR Descriptor(s)	4/2 or 3
Type of listening	Main focus: MISD / IPM. Includes following logic of an argument e.g. cause effect links, infer opinions and attitudes.
Rubrics	Listen to ? talking about ? Choose the correct answer (**A, B, C** or **D**) for questions **8- 14**. There is an example (0) at the beginning. First you have 45 seconds to study the questions. Then you will hear the recording twice.
Task 3	
Task description	Radio Interview or Talk. Monologue or Dialogue
Response format	MCQ – 4 options 7/8 items
CEFR Descriptor(s)	6 (3)
Type of listening	Main focus: MISD / IPM. Includes following logic of an argument e.g. cause effect links, infer opinions and attitudes as well as speakers mood.
Rubrics	1) Listen to a radio interview with ? talking about ?. Choose the correct answer (**A, B, C or D**) for questions **15-21**. There is an example (0) at the beginning. First you have 45 seconds to study the questions. Then you will hear the recording twice. 2) Listen to a radio programme about ? Choose the correct answer (**A, B, C or D**) for questions **15-21**. There is an example (0) at the beginning. First you have 45 seconds to study the questions. Then you will hear the recording twice.
Task 4	
Task description	Monologue. Instructions/announcement or a presentation/lecture
Response format	NF or NF table completion.7/8 items.
CEFR Descriptor(s)	5 or 4 (2)

Type of listening	Main focus: SIID / Selective listening
Rubrics	1) Listen to ? talking about ? Listen and answer the questions **22-28** in a maximum of *THREE* words. There is an example (0) at the beginning. 2) Listen to ? talking about ? Listen and complete the table for the questions **22-28** in a maximum of *THREE* words. There is an example (0) at the beginning. First you have 45 seconds to study the questions/table. Then you will hear the recording twice.
On every test form a representative sample of CEFR descriptors and topics should be included. On every test form there will be **one second language speaker.**	

The domain analysis is well supported by the literature review and draws on content standards presented in the CEFR as well as those presented in the baccalaureate curriculum. For example, each test form must include a range of types of listening (Gist, Main ideas, Specific Information). Task topics should be drawn from the TLU domain based on CEFR descriptors and the baccalaureate curriculum. The previous discussion showed that CEFR descriptors, purpose for listening and domain analysis provide the main parameters for linguistic demands. While the CEFR has been used as a guidance document, it has not provided a super-specification (North, 2004); instead, an extensive review of listening research has informed the test specifications for the present study.

4.1.2 Task Development

In accordance with the test specifications, I needed to develop the four tasks in order to operationalise my construct. Bachman (2002, p.471) states that test developers must fundamentally identify tasks corresponding to real-world communicative events which engage candidates in language use. That is, tests must be developed by integrating task and construct. By using authentic sound files with purposeful items and basing the listening activity on expert behaviour, it is hoped that the real-life cognitive processing demands will be incorporated into the testing situation. As a large part of an a priori validity argument is based on the item writing process (Zheng & De Jong, 2011), a description of this process will now follow in order to provide a priori validity evidence.

In response to the general call for authentic input, texts must be either sourced from the Internet or constructed in response to prompts in order to produce real, non-adapted, connected speech. This is a process which takes time, but was certainly made easier by the fact that I was able to follow detailed test specifications. The following four sound files were finally decided upon as suitable for exploitation in terms of the topics and test specifications:

Task 1. Opinions about sport. This sound file was constructed in response to prompts about controversial topics related to sport. I gave the male speaker a list of topics (17 in total) and asked the speaker to simply give his opinions with reasons. As such no script was followed; the resulting sound files can be considered natural. Sport is a topic found in all the surveyed textbooks. Of the seventeen utterances collected I chose eight which were propositionally and linguistically complex and contained abstract as well as concrete ideas in accordance with specific purpose 1 in the test specifications. This sound file also covers specific purpose 3, as the extracts are quite complex and the listener would simply have to 'catch' the main idea. The sound file is therefore considered appropriate to test for Gist/Main ideas.

Task 2. Moving to the USA. This sound file was sourced from the Internet and copyright permission to use the sound file was sought and granted. The sound file is a talk about moving to the USA from Mexico given by a Mexican, who is a proficient English speaker. The audio therefore covers specific purpose 4 and introduces a second language speaker to the test as per the test specifications. The speaker explains about his move to the USA and includes opinions and attitudes as well as cause and effect links and so is felt to be suitable to test MISD and IPM.

Task 3. **Text messaging**. This sound file was sourced from the Internet and copyright permission was sought and granted, although it would still need to be

further granted for the purpose of large scale use. The sound file is a radio interview with an academic about her research into language use in text messaging. It therefore fully covers specific purpose 6 and, as a dialogue, to some extent also covers specific purpose 3. The sound file also lends itself for testing MISD and IPM.

Task 4. Geography trip. This sound file was constructed in response to prompts which contained instructions about a forthcoming geography trip. I asked a female English teacher to convey the information as if she was talking to a class of students, something she is accustomed to doing. A number of attempts were made before the audio was considered suitable. The recording was made in a classroom and so mirrored the real life situation to a certain extent; although students were not present and I acted as the 'live audience'. The audio is therefore considered appropriate to cover specific focus 5 and to a certain extent specific purpose 4. The information conveyed was quite dense and so it is felt that the audio is appropriate to test selective listening in order to extract specific information and important details.

The two recordings which I gathered myself were made using a Sony ICD-UX200 digital voice recorder and their quality is considered to be of a high enough standard. Nevertheless, if this process were to be followed on a national scale the recording quality could be perhaps improved upon through the use of professional studio recording techniques. The two sound files taken from the Internet were also considered to be of adequate quality and representative of the quality candidates would encounter in the TLU. Having decided on the appropriate sound files and the types of listening which they could be used to test, I proceeded to the next stage of task development, that of *text-mapping*.

Text-mapping is a process which makes no reference to a transcript, rather it places the spoken word centre stage and therefore contributes evidence to the claim that the test is one of listening (see discussion on authenticity, section 3.4.1.1). The text-mapping

process should be carried out by a small team of test developers. This was not possible for the present study as it is not a collaborative project. I did, however, carry out the process myself and a brief explanation of the process follows.

Initially, the task developer makes a decision on the 'purpose for' and 'type of' listening which an expert listener would normally employ (e.g. listening to follow the main ideas and supporting details). The sound file is then sent to the rest of the test development team for text-mapping. The process is explained in detail in Green (2017), while a similar process for reading is outlined in Sarig (1989) and Urquhart and Weir (1998). The process is recommended by Weir (2005b, p.101) in order to replicate one type of listening when developing listening tasks. In the text-mapping process, each member of the team listens once to the sound file and notes the salient ideas which they have taken away from the text. In this way the developers form a consensus of agreement of just what an expert listener has understood from the text (the extracted meaning), and items are subsequently developed in order to test the understanding of only that information which has been noted by the majority of team members. Consensus is determined as being *n-1,* where *n* is the total number of test developers. As such, it is proposed that the development process responds to Field's (2012a) suggestion to test developers that they provide cognitive validity evidence by modelling the skill on expert listeners. The consensus provided is especially useful for higher-level proficiency tests as there may be different interpretations of implicit or pragmatic meaning and the overall discourse model.

After text mapping, I developed the items for each of the four tasks following the test specifications. An attempt was made to develop more items than were necessary as trialling may show that some of the items do not work as expected and need to be dropped from the test.[47] Table 3 below gives a description of Version 1 of the test. It can be seen that the construct has been well sampled. Following the test specifications, the

[47] However, task 1 is a MM task which will be greatly altered by adding an extra option so only 7 items plus an example were included in the original trial task.

test includes a variety of topics, specific CEFR focuses, types of listening behaviour, and response modes.

Table 3. Description of BFE CEFR B2 listening test (version 1)

Task (number of items)	**CEFR B2 Descriptor/ Specific focus**	**Type of listening and response mode**
Task 1 Opinions about sport (7 items)	1 and 3	**Gist/Main ideas** (G/MI). **Multiple Match** (MM)
Task 2 Moving to the USA (9 items)	4 and 2	**Main ideas with supporting details** (MISD)/ **Listening to infer (propositional) meaning** (IPM). **Multiple choice** (MCQ)
Task 3 Text messaging (8 items)	6 and 3	**Main ideas with supporting details** (MISD)/ **Listening to infer (propositional) meaning** (IPM). **Multiple choice** (MCQ)
Task 4 Geography trip (10 items)	5 and 4	**Selective listening / Specific Information and Important details** (SI/ID). **Note form** (NF)

Once the tasks had been developed, the complete test was put together, with the addition of a context significant image included to help test takers activate their relevant schemata. The sound files for individual texts were then combined into a single final sound file, which also included necessary pauses for reading questions and recorded rubrics taken from the text specifications. The inclusion of recorded instructions which match those on the question paper has been noted as reducing test taker anxiety, as they

allow the candidate to more easily follow the rubric and contextualise the task (Green, 2017).

As a large-scale pilot study takes a lot of time and effort, care must be taken that as many problems as possible are identified before it is carried out. It was therefore decided to carry out a small-scale trial before going on to the main study. However, a large testing body should have the resources to go through a number of rounds of revision and review before going to pilot. I will now report on the results of the small scale pre-pilot study of the original 36 items along with the conclusions which were drawn from this study.

4.1.3 Pre-pilot study

For the first part of the pre-pilot study, the test was given to a group of B2 students at the Centro de Lenguas Modernas (CLM), University of Granada studying a familiarisation course for the CLM CEFR B2 exam and who were all motivated to take part. The 16 participants had all undergone an internal level test and could all be reasonably considered to be at or around CEFR B2 level. An initial analysis of results was run using classical test theory in SPSS, which can be seen below in Table 4. This initial analysis provides some useful information, particularly the FV (facility value). The FV shows the percentage of test takers who got the item correct and, even though my trial sample was very small, it can give some guidance as to whether the test is the correct level of difficulty for the target proficiency group. With such a small sample, the discrimination index (DI) could be affected by a few random responses and so should not influence decisions too much. For example, item Q3.1 seems to show negative discrimination due to one correct answer given by the weakest participant. It can be seen at a glance that task 4 was the most difficult, with most items having very low FVs and no test takers answering item Q44 correctly. Conversely, item Q26 in Task 2 was answered correctly by all participants. Such items effectively give no information about our target audience and are therefore not useful to include in a proficiency test.

Table 4. Classical item analysis for pre pilot study (n=16)

	Facility value (FV)	Discrimination index (DI)
TASK 1 Sports		
Q1	.88	.487
Q2	.75	.583
Q3	.81	.227
Q4	.50	.403
Q5	.44	.516
Q6	.88	.349
Q7	.94	.443
TASK 2 Move USA		
Q2.1	.63	.328
Q2.2	.56	.446
Q2.3	.44	.323
Q2.4	.75	.472
Q2.5	.69	.042
Q2.6	1.00	.000
Q2.7	.56	.373
Q2.8	.44	.665
Q2.9	.50	.056
TASK 3 Text messaging		
Q3.1	.56	-.064
Q3.2	.81	.286
Q3.3	.94	-.338
Q3.4	.94	-.023
Q3.5	.38	.395
Q3.6	.56	.116
Q3.7	.50	.598
Q3.8	.75	-.006
TASK 4 Geography trip		
Q4.1	.13	.217
Q4.2	.50	.238
Q4.3	.44	.252
Q4.4	.00	.000
Q4.5	.13	.425
Q4.6	.06	.354
Q4.7	.19	.298
Q4.8	.06	.307
Q4.9	.63	.576
Q4.10	.13	.425
Q4.11	.44	.516
Q4.12	.50	.673

For the second part of the pre-pilot study, I carried out two verbal reports with volunteer students from my own Cambridge FCE preparation classes at the CLM (one boy and one girl), who I considered to have a CEFR B2 proficiency level in listening. The purpose of this part of the study was twofold: (1) to collect information about any difficulties encountered whilst doing the test, and (2) to pilot the methodology before using it in the main study.

After combining information from the two data sources, the following changes were made to the test tasks:

1. Sports opinions: All the items discriminated well but Q7 was very easy for this population, with only one person choosing a distractor. Q7 has a FV of 94% and it was seen from the verbal reports that this item could be guessed correctly by simply 'word spotting' and using inference, that is, by not understanding very much of the audio as shown by the following extract.

> '*This is strange, but I've heard something about 'flexibility'.....this is strange, the most strange question. I think that he was talking about yoga.*'
>
> Participant 1 (Task 1, Q7)

It was therefore decided to use this extract as the example and substitute the item/sound file with another extract (from the original 17) which is more abstract and contains less obvious vocabulary. Although Q1 also appears to be relatively easy, it has a good DI and an easy first question on a test could reduce anxiety for the test takers. Q6 was also an easy item, and it was seen that the sound file repeated key information leading to the choice of the key, this repetition was therefore removed from the sound file with no effect on the overall naturalness of the utterance. The order of the items was also changed as Q6 appeared opposite the correct place in the answer key, perhaps making this item easier. It was also seen that distractor G was very strong for item 5 so the word 'women's' was removed from the item in the hope that it would now work as a more general distractor for all the items.

2. Move to USA: The items appear to be targeting the difficulty level very well, with FVs ranging from .44 to .75—apart from Q2.6, which all participants got correct. This item was based on a part of the sound file that was well explained and exemplified. The information was explicit and would therefore better represent a CEFR B1 listening item. I therefore decided to drop the item from the test. The verbal reports showed that Participant 2 did not choose the correct answer for item Q2.1 simply because she did not understand the key, the word 'straightforward', even though she had understood the input text. Indeed, on further analysis using tools developed to analyse CEFR-related vocabulary it was found that the lexical item 'straightforward' is shown to be B2 in Cambridge *English Vocabulary Profile* and B2 spoken in the *Longman 3000* vocabulary lists (although it does not appear in the written corpus). As this is not a test of reading, test takers should not get the item incorrect because they cannot understand the written question, and I therefore changed 'straightforward' to 'relatively easy'. Q2.5 has a low discrimination index and the verbal reports showed that this may not have simply been due to careless mistakes; the results showed that option A was an extremely strong distractor, which was explained in the verbal reports.

> '*It's boring because he claimed he was from a big city and Charlottesville is a small village and there isn't much he can do... he's from a city that is 24/7'*
>
> Participant 2 (Task 2, Q5)

This response shows that higher order skills of inference could be applied to the item and that the distractor 'boring' could be reasonably considered to be implied. I therefore simply changed the distractor 'boring' to 'interesting', which was definitely not stated in the sound file, yet which still remains a plausible distractor. Q2.9 has a good FV of 50%, but a very low DI. Further investigation showed that the two weakest students got the item correct. Yet the two strongest students also got the item correct, as did both participants who did the verbal report. This could therefore be interpreted to simply be the result of a couple of lucky guesses by the weaker students and it was therefore decided to keep the item.

3. Text messaging: Most of the FVs are within range, but both Q3.3 and Q3.4 were easy for the target population. Both items had low DIs and relatively high ability participants got the items incorrect. While this could have been due to careless mistakes, I nevertheless examined the items in more detail and made a few minor changes. In Q3.3, I removed 'took' which was used in the sound file and changed the distractors in the hope of making them stronger, as well as making them more similar in length. For Q3.4, I used exactly the same idea from the sound file (Text Map) but reversed it as the idea of 'age' was very explicitly stated in the audio. Item Q3.1 showed distractor B to be too strong, and consequently it was changed.

4. Geography trip: This was the most difficult task and most items had very low FVs. However, the DIs were good and show that it was the better test takers who got the items correct. Although it is felt that the information was not too difficult, the low FVs may possibly be explained as the result of high cognitive load due to the large number of items (12), which is probably too many. Furthermore, the verbal reports showed that the participants did not find the speaker difficult to understand.

> *This was more difficult ... but I think it's easier to understand this woman when she talks, I understood nearly everything, she speaks more clearly but you need to take the ideas and write them down. It was difficult for me to write and listen, I think that I know some of the answers but I need to listen again and try to arrange all the ides I have*.
>
> Participant 1 (Task 4)

Indeed, the verbal reports suggested that the problem with this task was that there was too much information for the test takers to process. Although they believed they were answering the items correctly, in many instances they put the incorrect answer.

It was therefore decided to give the task extra scaffolding by presenting some of the information already in the task. In this way, the task looks like notes taken down during the talk and the test taker simply has to complete the missing information. I directly removed item Q4.4 which had a FV of 0% and so was too difficult, and reduced the

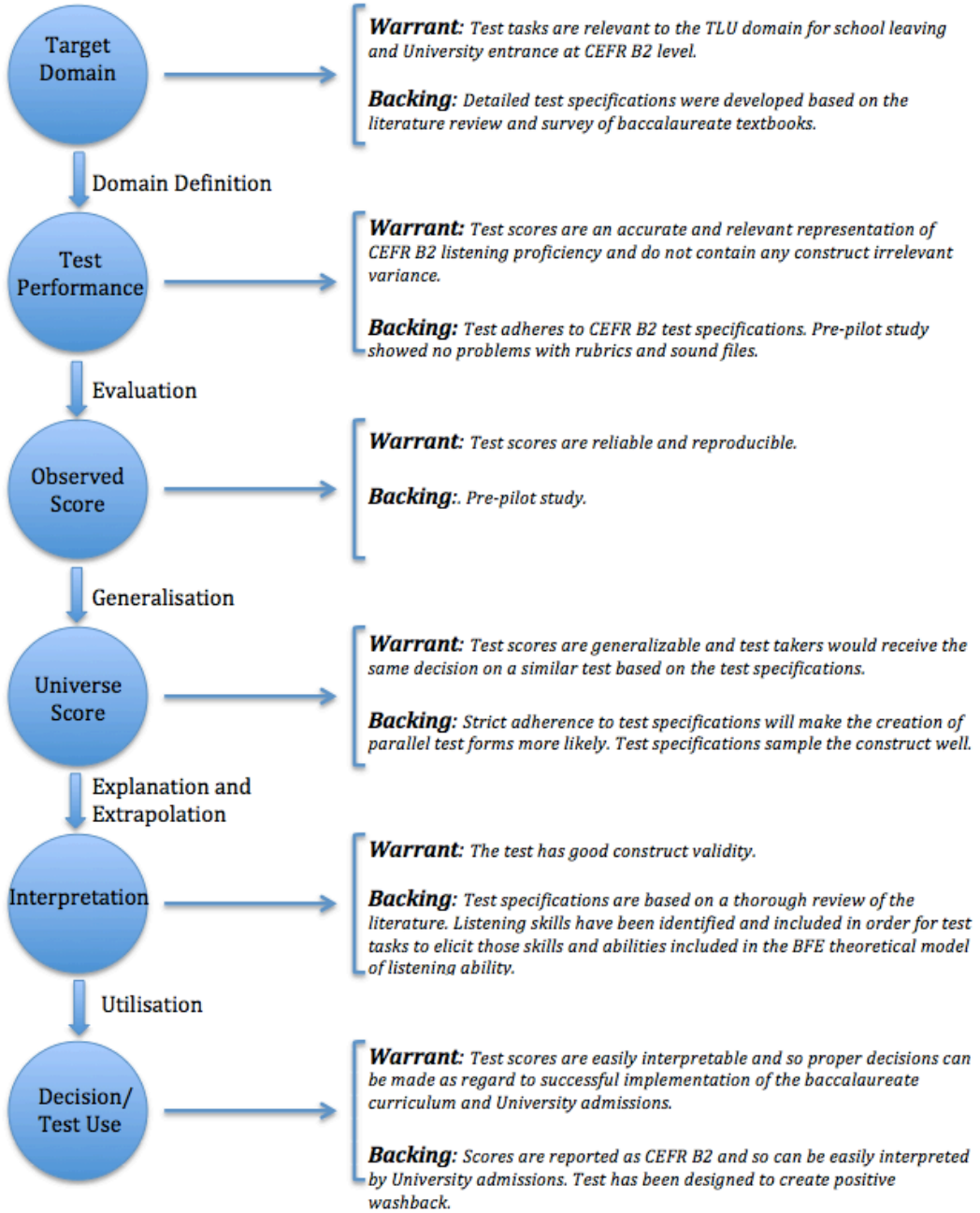
number of items to ten, allowing for a further two items to easily be removed after trialling.

After all the adaptions, a 33-item test which appears to be targeting the correct level in terms of the test specifications and difficulty was produced. The pre-pilot results, both quantitative and qualitative, were helpful in making these changes. Indeed, here I have followed the advice of Fulcher and Davidson (2007), who recommend the use of verbal report methodology at the task development stage for reading and listening tests. It was also useful to pilot the verbal report methodology as this helped in its design and final implementation for the main study (see Methodology section). As a final check, I also decided to ask a small group of native English speakers to do the test, and no further problems were found.

4.2 Presentation of the BFE CEFR B2 Interpretative Argument

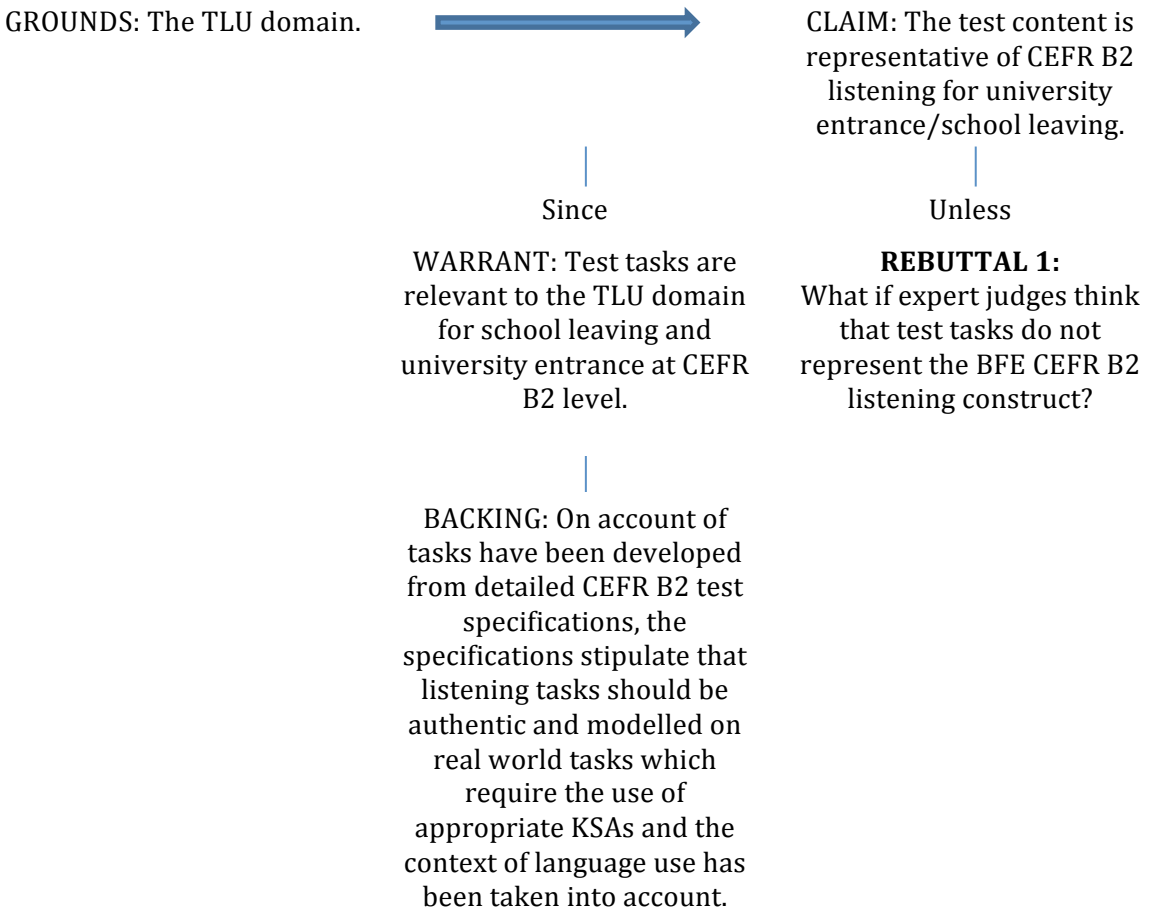
At this stage in the test development process, I can now present the initial IA. I have carried out the pre-planning stages of test development and have a 33-item test which will be used as my measurement instrument. A simple initial IA is presented in Figure 11.

Figure 11. Initial interpretative argument for BFE test



It can be seen that the evidence on which claims for the test's validity rest are mainly based on the fact that the test specifications were drawn up based on an exhaustive literature review and a domain analysis. The pre-pilot study also offers some validity evidence towards the test; however, this evidence cannot be considered as strong. I will therefore now outline Toulmin's more detailed argument approach for each of the validity inferences of my IA in order to identify possible rebuttals and draw up my research questions. Each argument targets one inference (or bridge) in my IA and is based on Toulmin (2003, p. 97).

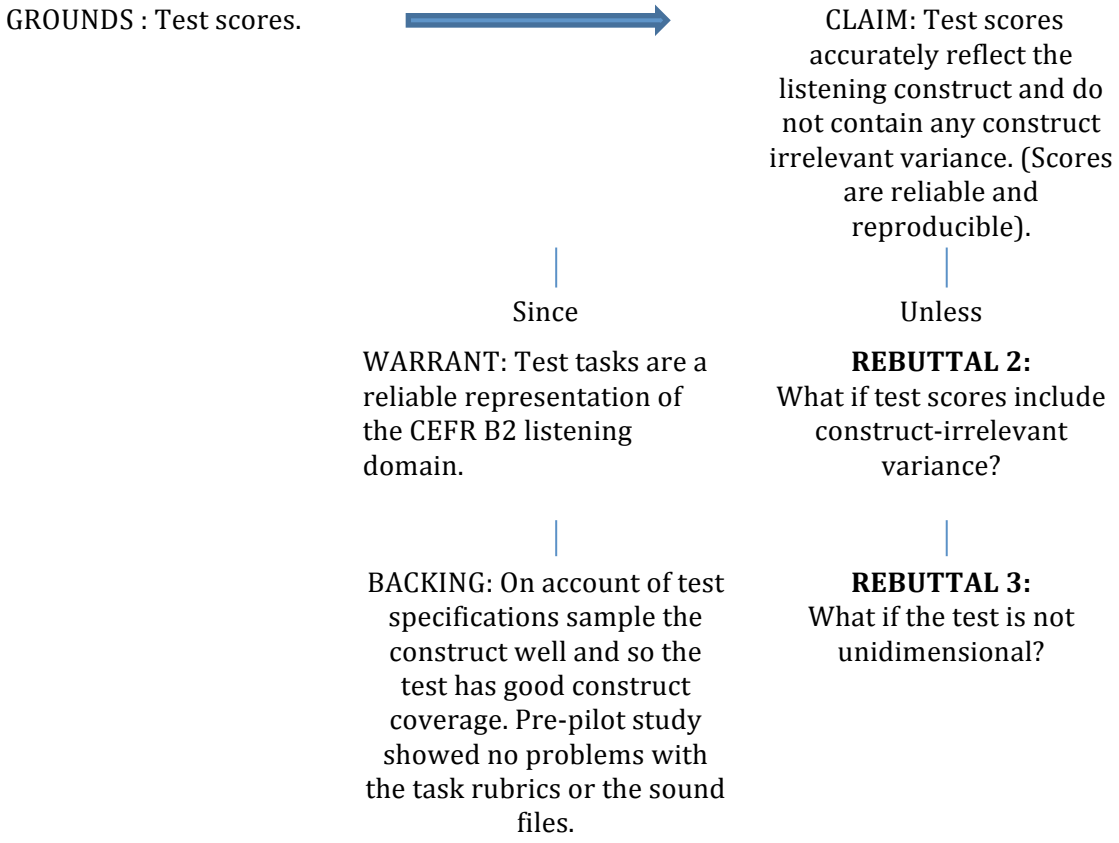
Domain definition (observation inference)



Research question:

Do expert judges believe the test tasks to be an accurate representation of the BFE CEFR B2 listening construct?

Evaluation Inference.



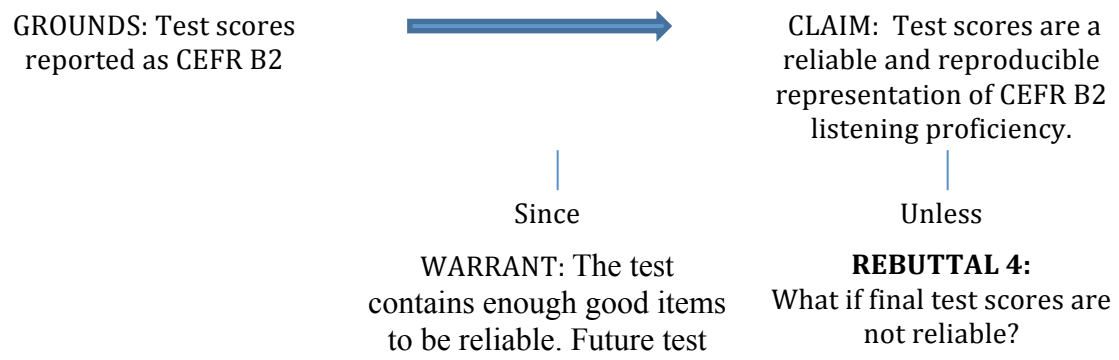
Research questions:

(What are the statistical properties of the test items?)

Are test scores unidimensional?

Do test scores include any construct-irrelevant variance?

Generalisation Inference.



forms will be a measure of the same construct.

BACKING: On account of test specifications are well-defined enabling parallel test forms to be produced.

REBUTTAL 5: What if parallel tests cannot be produced?

Research question:

Are scores on the final test form reliable?[48] (Can parallel test forms be produced?)

Explanation and Extrapolation Inference.

GROUNDS : Test scores.

CLAIM: Test scores reflect the CEFR B2 listening construct. Test takers use relevant KSAs to solve test items. Test takers would receive the same score on a similar test of the same construct.

Since

WARRANT: Observed score is attributable to the relevant KSAs for a CEFR B2 listening construct.

Unless

REBUTTAL 6: What if test takers use construct irrelevant strategies to solve test items?

BACKING: On account of strict adherence to test specifications and text-mapping procedure was used during task development to ensure listening behaviour reflects that of an expert.

REBUTTAL 7: What if test takers received a different score if a different measure was used?

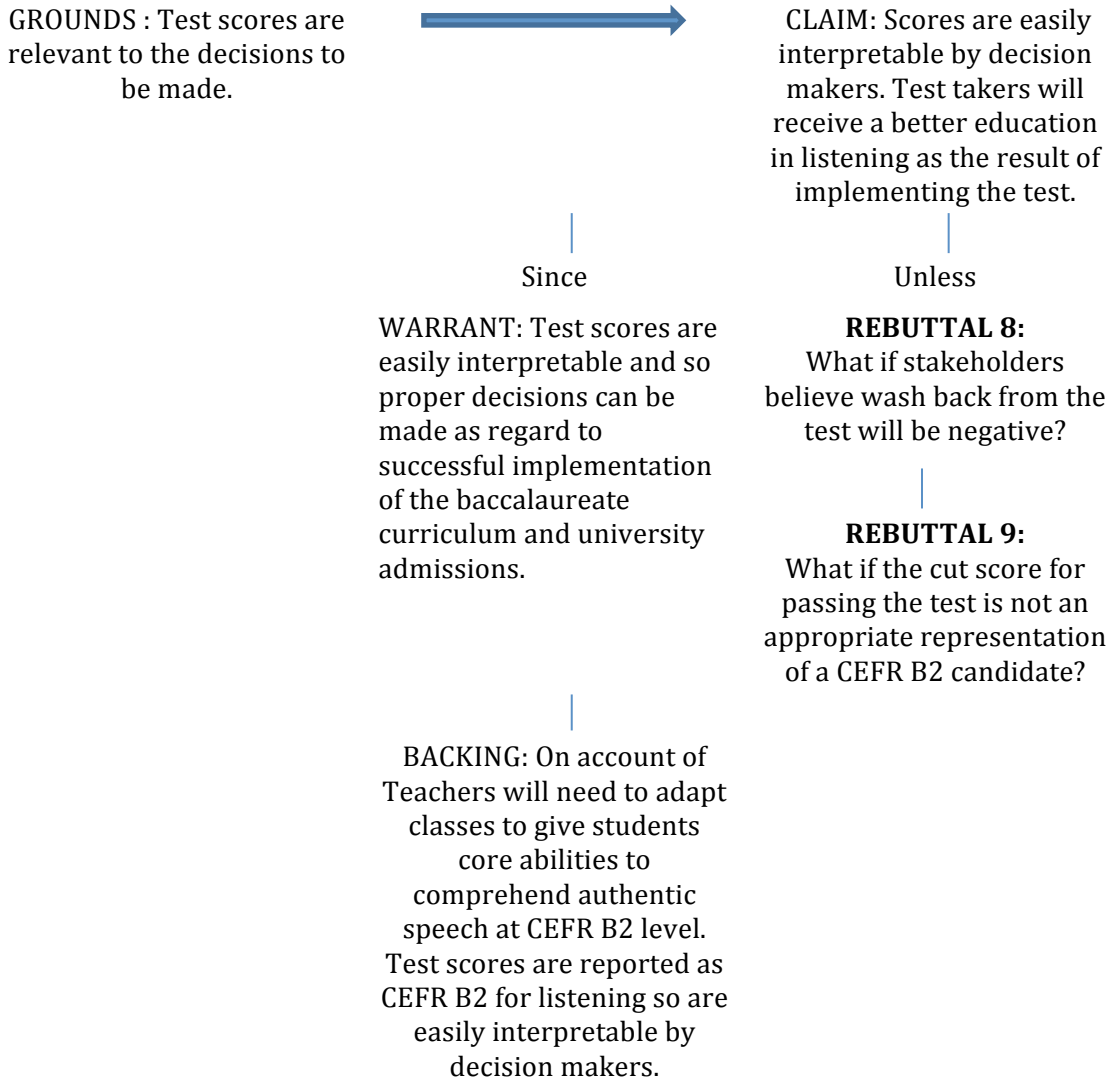
[48] In effect, this research question also provides backing for the evaluation inference showing whether scores are reliable and relevant.

Research questions:

Do test takers use the relevant knowledge, skills and abilities to solve test items on the BFE listening test?

Do test scores correlate with other measures of the same construct?

Utilisation Inference.



Research questions:

What are candidates' opinions of the BFE listening test? Do candidates believe that the test will have positive washback?

What should the cut score representing a BFE CEFR B2 ability level on the test be?

4.3 Final research questions.

It can be seen that most of the backings which already exist are based on the literature review, test specifications and test development procedure. Key research questions have been identified based on possible rebuttals. In order to answer these questions, a mixed-method approach will be applied using both qualitative and quantitative research methods, the triangulation of results will improve the validity of the study. It is therefore felt that the research questions need not be addressed in the order that they appear for each validity inference, as information from the different sources can be triangulated and fed back into the results obtained from the different data sources. Indeed, Bachman (2006) presents a 'research-use argument', which advises for a move to integrate different data sources in order to give more meaning to the data and link observations with interpretations, as interpretation supported by evidence from several sources will be more convincing.

Therefore, my final research questions are as follows, presented in the order in which they will initially be addressed:

Research question 1 (R1): What are the statistical properties of the test?

Research question 2 (R2): Is the test unidimensional? Do test scores include any construct-irrelevant variance?

Research question 3 (R3): Do test takers use the relevant knowledge, skills and abilities to solve test items on the BFE listening test?

Research question 4 (R4): Are scores on the final test form reliable?

Research question 5 (R5): What are candidates' opinions of the BFE listening test? Do candidates believe that the test will have positive washback?

Research question 6 (R6): Do expert judges believe the test tasks to be an accurate representation of the CEFR B2 listening construct?

Research question 7 (R7): What should the cut score be on the test in order to provide an accurate evaluation of a CEFR B2 candidate? (Can parallel test forms be produced?)

Research question 8 (R8): Do test scores correlate with similar measures of the same construct?

Chapter 5. Methodology

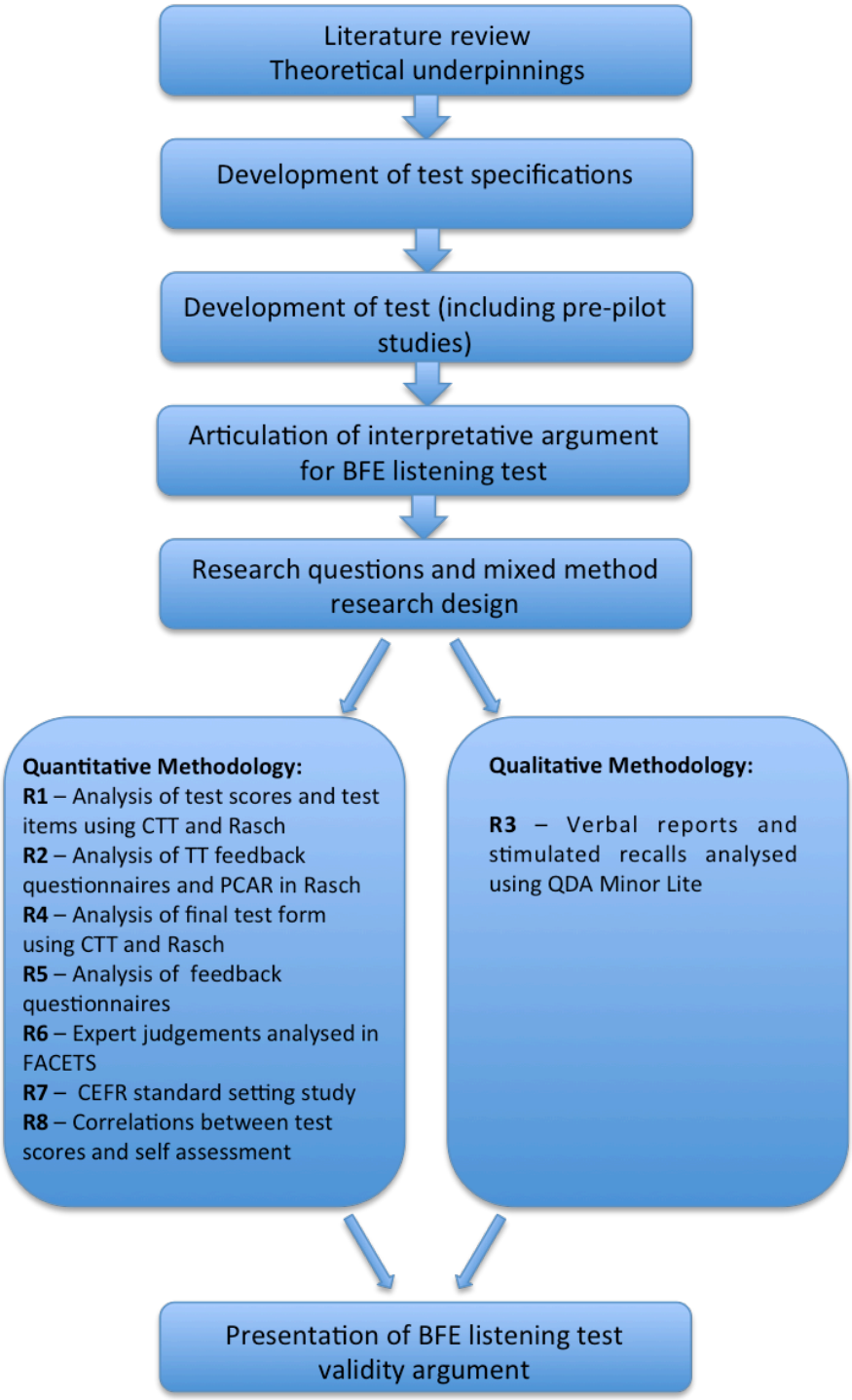
5.1 Research design.

In language testing we need to build a validity argument for our tests; consequently, investigation will normally draw on both quantitative (QUAN) and qualitative (QUAL) approaches. Creswell (2012) argues that mixed method research is an inevitable extension of studies which use different methods for triangulation purposes. Indeed, mixed-method research often relies on more than one QUAN methodology. Here, each research question is to be answered using the research methodology which is considered to be the most relevant and useful. In the present study, my initial research design was conceptualised in this way, as can be seen in Figure 12. The flowchart includes the steps which have already been taken in order to develop the measurement instrument.

However, it soon becomes apparent that such a hierarchical distinction is difficult to make, and evidence from different data analyses can be used to contribute to the answer to more than one research question. The QUAN/QUAL divide is also difficult to substantiate; for example, data obtained during a standard-setting study is essentially a qualitative judgement which is analysed using quantitative techniques. Certainly, some methodologies, such as verbal reports, require an initial QUAL analysis and once data has been coded, it can be ‘quantified’ and a more QUAN approach can be applied to the

data. Such designs have been called ‘a sequential mixed-methods design’ (Crewell & Plano Clark, 2011).

Figure 12. Flowchart of research methodology



Dornyei (2007) argues that there exist a number of approaches to research designs which use different methodologies; for example, each analysis could be carried out separately and any mixing of results left until the end. Alternatively, we can start integrating data at the analysis stage of the research project. Such approaches can guide the researcher and any given interpretation is made stronger if it is supported by evidence from several sources. In short, it is argued that a true mixed-method approach is not only feasible, it is desirable. It will be seen then, that my research design is not quite as linear as Figure 12 suggests and that although I begin by answering each research question separately, results will be combined during my discussion. I will now go on to discuss the different research methodologies employed in this study.

5.2 Research Methodologies

If we wish to generalise our results beyond collected observations, our interpretations must be meaningful—they must be well-grounded and based on multiple sources of information. Consequently, I have taken the decision not to outline the methodology for each separate research question, but rather to explain the methodologies I will be using and later present the results of each analysis separately. I will then attempt to answer my research questions by integrating my results during my discussion. In short, a pragmatic approach is taken which does not divide my methodology by research question. I will therefore discuss the methodologies which will be used in the study, before going on to outline my data collection and analysis procedures.

5.2.1 Analysis of test scores

Before the test is administered in a live situation, it must be piloted to check that the items have worked as the test developer intended. The results of this analysis give information about test difficulty, item and distractor functioning and test reliability. There are two psychometric theories which can be used for this analysis and these shall be discussed here.

5.2.1.1 Classical Test Theory (CTT)

ALTE (1998, p.138) define CTT as "a measurement theory which consists of a set of assumptions about the relationships between actual or observed test scores and the factors that affect these scores, which are generally referred to as error." Notions of true score and error are central to CTT, as the observed score is considered to be the *true score* plus error. In terms of language testing, a candidate's raw score on a test is therefore a combination of their true score plus error, which is due to effects that are not being measured by the test. This means that on any test which uses a CTT psychometric theory, the scores will contain *error variance*. "This notion that observed score variance is made up of true score variance plus error variance underlies the entire framework of CTT." (Brown, 2012, p. 324).

A number of computer programmes exist which can be used to run CTT statistics, such as SPSS, the programme which will be used in the present study. Descriptive statistics are useful for describing the population which took the test. Information about means, standard deviations, variance, skewness and kurtosis tells us about the distribution of the population and can confirm whether this distribution is normal, and hence which inferential statistics need to be used for further analysis.

Two useful indices from a CTT analysis in language testing are the *Facility Value* (FV) and the *Discrimination Index* (DI). The FV reports the percentage of test takers who responded correctly to an item and is therefore a measure of item difficulty (e.g., an item with a FV of 20% is much more difficult than an item with a FV of 80%). This information is therefore only useful at the test development stage if the test is piloted on the correct population, as obviously C1 candidates would find a B1 test very easy and the items would probably have FVs close to 100%, thereby giving little useful information to the test developer. The present study has access to very little information about candidates' proficiency level and so using CTT may not be the most appropriate methodology. Nevertheless, if a test is well targeted, items with a FV between 20% and 80% would be most productive (Green, 2013). By choosing items which correlate the

best with total raw scores, the items are more likely to correlate with true scores and therefore be capable of discriminating between lower and higher ability candidates. Using the statistical package SPSS, the DI is represented by *item-total correlations* and tells us to what extent each item is testing what the more reliable total score is testing (Brown, 2012). The DI illustrates the items which are able to discriminate and recommended values are as follows (Popham, 2000, cited in Green, 2013, p.29):

.40 and above	Very good items
.30 to .39	Reasonably good items but possibly subject to improvement
.20 to .29	Marginal items, usually needing and being subject to improvement
.19 and below	Poor items, to be rejected or improved by revision

The higher the DI, therefore, the better the test is able to differentiate between higher and lower ability candidates and the more reliable the test will be as a measure of candidate ability.

In CTT, reliability can be defined in terms of both true score variance and error variance, and a typical measure is Cronbach Alpha, which takes into account the number of items on a test, standard deviation (SD) of test scores as well as the variance of each item. It is not necessary to do complicated calculations, as they are done internally by SPSS. The resulting figure tells us the amount of observed score variance that is true score variance and a higher Alpha is therefore better. It is a measure of internal consistency, that is, of whether the items are all measuring the same underlying ability. Recommended values for alpha are above 0.7, but values above 0.8 are preferable (Pallant, 2007, p.989), though the language testing literature often cites values of above 0.9 as desirable for high stakes tests. However, these internal consistency estimates are computed using nonlinear raw scores and extreme scores (which have no error variance) are included in the calculation.

Test reliability estimates do however contain error and this error should always be reported if we want to have the full picture of how reliable test scores are. The standard error of measurement (SEM) is an important indicator of the consistency of test scores, especially for pass and fail decisions around the cut score. It is arrived at using the following equation:

$$\text{SEM} = \text{SD} \times \sqrt{1 - Reliability(\alpha)}$$

The SEM value tells us how much a candidate's score can vary due to error. However, as the calculation is based on the complete test, it will actually differ depending on the test score achieved and will be larger at or near the mean (Bachman, 2004; Brown, 2012). This means that any reported error will be overestimated for candidates with low and high scores. In order to have 95% confidence in a score we must apply 2 (1.96 to be exact) x SEM. Here, for example, if the SEM is 3 and a candidate scores 20 on the test, we can have 95% confidence that the true score is between 14 and 26. This has obvious implications when making pass/fail decisions around the cut score, and lower values for SEM are therefore desirable.

The advantages of using CTT is the fact that it is easy to use (especially using a statistically package such as SPSS) and it is understood by a wide audience. The major disadvantage is that the results are population dependant, and only apply to the population who took the test with those items that appear on the test. It is difficult to generalise results to a different or wider population. Also, because CTT is very much dependent on correlation analysis, extreme scores can greatly influence results. Indeed, in the context of second language proficiency, it could be argued that in order to provide evidence of a reliable test, simply give that test to a group of natives and a group of beginners. Indeed, Brown (2012, p.333) specifically provides test developers with the following warning: "Don't use CTT for criterion-referenced purposes because CRTs are not designed to spread students out and are therefore not referenced to a normal

distribution." Indeed, any test which purports to be CEFR-related is a criterion referenced test and so it is highly recommended that such tests should be developed based on *Modern Test Theory*.

5.2.1.2 Modern Test Theory (MTT): Rasch

To overcome the aforementioned limitations of CTT, most testing bodies now rely on *item response theory* (IRT) models for test development, the great advantage being that a robust banking system for test tasks can be developed. Using a system of linking, common item equating and anchoring, all test tasks can be placed on the same measurement scale (see for example, Kolen & Brennan, 2014; North & Jones, 2009; Wright & Stone, 1979). This is extremely important if we want our tests to be parallel in difficulty between administrations. Once a standard-setting study (see 6.7) has been carried out, the pass mark to reflect CEFR B2 listening proficiency can be placed at exactly the same candidate ability level on every version of the test. However, it should be noted here that a similar possibility also exists in CTT via the transformation of raw scores into z-scores (using the mean and a SD of 1) or percentile ranks (Wu & Adams, 2007), even though such transformations will not provide the same robustness of common-interval linear scale and sample free estimations as the Rasch model.

There is a family of IRT models which can be used to analyse test data and which are normally classified as one, two and three parameter models. The Rasch model (George Rasch, 1960) is a one-parameter logistic regression model and is often the model of choice in language testing. However, discussions amongst Rasch practitioners both in online communities and in the literature would dispute the fact that the Rasch model belongs to the family of IRT models. Although similar, the underlying philosophy behind Rasch measurement theory is very different. IRT aims to fit the model to the data, whereas Rasch aims to fit the data to the ideal model (Bond & Fox, 2015). Indeed, there are conceptual differences between the Rasch model and one-parameter IRT models (Linacre, 2005). The following discussion will therefore always refer to the Rasch model and not IRT. A user-friendly description of the Rasch principle is given by George Rasch,

> a person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one.
>
> (Rasch, as cited in Bond & Fox, 2015, p. 11)

The Rasch model is a probabilistic model which at its most basic is dichotomous—either correct or incorrect. Persons and items (two sources of measurement error) are given probability estimates in order to estimate item difficulty and person ability. Sick (2008) highlights that the major difference between CTT and Rasch is that CTT is descriptive and sample population dependent whereas Rasch is probabilistic and inferential. The person and item difficulty estimates can be used to predict the performance of any candidate at a given ability and of any item at a given difficulty. As McNamara states, we can

> make generalisations from the performance of a particular sample of subjects on a particular sample of items to enable us to estimate the ability of candidates in relation to the entire universe of such items and the difficulty of the items for the entire population of prospective test takers.
>
> McNamara (1996, p.153)

Rasch measurement is therefore said to be sample independent and this characteristic means that the Rasch model lends itself to measuring a proficiency scale such as the CEFR because item difficulty and person ability are measured on the same scale. This is made possible because both measurements are converted to the unit of measurement used by the Rasch model, 'logits' (a contraction of 'log odds unit'), hence both measurements can be placed on a common scale. This is in contrast to CTT, where raw scores are used. As Wright and Moc (2004, p.3) put it, this is important because "in order for measurement to be useful for inference, it needs to be linear and reproducible".

The Rasch model, therefore, has the added benefit of being able to easily relate test

scores to the construct, in the present case listening ability. As Wu and Adams (2007, p. 19) state "the notion of a construct has a special meaning in item response theory". The Rasch model attempts to measure the latent trait, which is not directly observable; instead, responses to items must be measured and converted into test scores, which are then a reflection of the latent trait. In the Rasch model, the latent trait is not a reflection of the item responses, rather it is the item response data which reflects the latent trait. The reader is directed to Bond and Fox (2015) for an in-depth discussion of the philosophy behind the Rasch model and its construction based on a matrix of observations and probability theory. The following, more detailed, discussion of the Rasch model will instead be centred around the *Winsteps* program (Linacre, 2017a), which will be used to perform Rasch analysis in the present study (version 3.71.0.1).

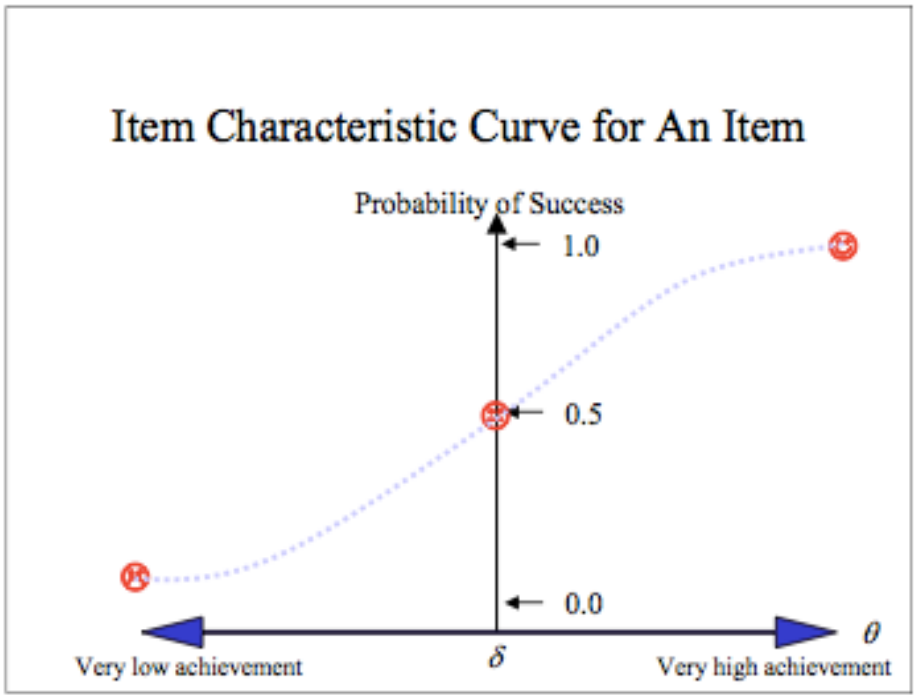
Once data has been entered, an iterative process—known as convergence—begins, where tentative person abilities and item difficulties are used to fit the model. This process may take a number of cycles before the data and the model converge. Once the item parameters have been estimated, we can model the response pattern of a particular item on a test using the equation:

$$P = 1/(1+\exp(-(\theta - \text{difficulty})))$$

Where θ is the ability level (or proficiency on the latent trait).

The *item characteristic curve* (ICC) represents the probability of getting an item correct across all ability levels on an item. An example of an item characteristic curve is given in Figure 13, where θ is a person-parameter representing person ability and δ is an item-parameter representing item difficulty on the same latent variable scale.

Figure 13. An item characteristic curve (Taken from Wu & Adams, 2007, p. 28).



Where item difficulty and person ability are the same, the person has a .5 (or 50%) probability of getting the item correct. As person ability increases, so the probability of answering the item correctly increases. With any given set of data, we have more information about an item where it is matched to the ability level of the persons, and this can be seen more clearly using the *Item Information Function* (IIF). The sum of all IIFs gives us the *Test Information Function* (TIF). The TIF can be used in test design by demonstrating areas on the ability continuum where there are few or no items. If necessary, items can then be added to the test to target these ability levels.

Once the parameters for item difficulty and person ability have been calculated in *Winsteps,* a useful output table is the item variable map. This table shows us the hierarchy of item difficulty calibrations at a glance, together with the corresponding person ability scale showing the level of proficiency reached by each person. Here, any mapped person has a 50% probability of correctly responding to a mapped item at the same point on the common scale. In a trial situation, the variable map shows us whether or not the test does

in fact have the correct level of difficulty for the sample population. In this context, the reported *standard error* (SE) by the Rasch model is superior to CTT, as it provides estimates of modelled error for every person ability and item difficulty (Wright & Stone, 1979)—a much more precise estimate than in CTT. These error estimates should not be higher than 0.3 (Linacre, 2017a). It will be seen (section 6.7) that the information about the continuum of item difficulty is extremely useful for conducting standard-setting cut score studies, using the Bookmark Method.

The notion of model fit is central to Rasch measurement theory, as we need to demonstrate that all the items do indeed tap into the same latent trait, i.e., that they are all measuring listening ability.[49] In order for the data to fit this model, all the items must therefore be measuring the same latent trait or construct and persons must be behaving in the same way. In test development, items which fit the model well are then considered to be interchangeable, they are all measures of the same construct.[50] This means that the model is extremely useful at the piloting stage of test development and can be used to select the well-fitting items to be used on a final test form in order to produce a test with a valid scale construction. In the present study, the trial data will be analysed using the Rasch model and any misfitting items will need to be removed from the final test form.

Fit statistics tell us about residuals, which are the differences between the predicted and the actual data (McNamara, 1996). Winsteps output includes fit statistics that provide information about *misfit,* data which does not fit the model. Misfits can arise for numerous reasons. For example, person misfits can be due to a candidate having a higher ability than the test items, which may result in that person being bored and not taking the test seriously. Items are also shown to misfit if they do not test the underlying construct (e.g., items involving mathematical calculations on a listening test). During a Rasch analysis, person fit statistics should first be scrutinised and greatly misfitting persons

[49] The data must fit the model rather than the other way round, as we are not searching for a model to fit our data as in many other research contexts.
[50] Indeed, one of the major drawbacks of CTT is that the inferences made about test scores are only interferences about that particular set of scores, they are not inferences about one underlying construct (Wu & Adams, 2007).

removed from the data set (easily achievable in Winsteps using the PDELETE command). Winsteps provides both *Outfit* and *Infit Mean Squared* (MNSQ), which are the Chi-squares divided by the degrees of freedom. Here, Linacre (2017a) recommends first investigating Outfit MNSQ as this data is more sensitive to outliers and is therefore easier to investigate, whereas the Infit MNSQ is weighted to give more importance to item responses at the item difficulty level and so is more sensitive to unexpected patterns of observations by persons on items that are targeted at their level (Linacre, 2017a). The nearer the fit statistics are to 1 the better, as this value means that the observed variance is exactly the same as the predicted variance. Various recommendations have been put forward for MNSQ values: for example, Linacre (2017a) says in order for an item to be productive for measurement MNSQs should be in the 0.5 - 1.5 range, while McNamara (1996) proposes a more conservative 0.7 - 1.3. Here, *underfit* is more problematic than *overfit,* as underfit shows serious deviation from the model, whereas overfit simply represents overly predictable behaviour. A transformed standardised fit statistic, *Zstd*, is also reported, which tests the null hypothesis that the data fit the model once predicted randomness has been taken into account (Linacre, 2017a). Here, values of ± 2 are considered acceptable.[51] If misfitting items are detected, we can further investigate by looking at the distractor analysis (for MCQ and MM items), as the misfit could be due simply to one badly- or unusually-performing distractor (e.g., if a distractor could in fact be considered to be correct).

Before applying the Rasch model, certain conditions must be present, items must be locally independent and the data must be shown to be unidimensional (only the construct of interest is being measured as opposed to any other). Here, some would argue that a construct such as listening ability is multidimensional (see Henning, 1992, for discussion on the multidimensionality of psychological constructs). Similarly, McNamara (1996) makes a distinction between the definitions of 'unidimensionality' in psychology and in measurement, arguing that psychometric unidimensionality does not preclude psychological multidimensionality, which is to say, we can describe a so-called multi-

[51] Large simple sizes can effect the Zstd measures and so Zstd can be ignored in large data sets (Linacre, 2017a), but in the present study they should be examined because the sample is less than 300.

dimensional construct (such as listening comprehension) as a single test score. Candidates can be conceptualised as having more or less of an ability, placed along a continuum of listening ability where more or less ability is required to solve items on the test. Items function together to define the continuum of listening ability, which is expressed not as a raw score but rather as a linear measure. Information about unidimensionality can be provided using *Fit* statistics (Sick, 2010) as well as a further investigation of Rasch residuals. In Rasch measurement Linacre (2017a) advises performing a *Principle Components factor analysis of Rasch Residuals* (PCAR) as the preferred method to identify multi-dimensionality in the data. This analysis identifies structural differences between opposing constructs and can be used to identify a possible secondary component being measured by the test.

In terms of reliability, Winsteps produces summary statistics. The *Person reliability index* "indicates the replicability of person ordering we could expect if this sample of persons were given another parallel set of items measuring the same construct" (Bond and Fox, 2015, p. 40). This is similar to Cronbach's Alpha in CTT. The *Item reliability index* "indicates the replicability of item placements along the same pathway if these same items were given to another sample of the same size that behaved the same way" (Bond and Fox, 2015, p. 41). These indices tell us how reliably test-takers are separated and how reliably test items are separated, regardless of the test-takers who take the test. The *Person separation index* indicates how many statistically distinct levels of proficiency are provided by the data. Linacre (2017a) states that this number must be at least as high as the number of proficiency levels the test is supposed to report. In the present study, that number is two—a pass or fail at CEFR B2.

It can be clearly seen that MTT has many advantages over CTT, especially in the case of criterion-referenced tests. It is this methodology which is now used by all major English test providers (e.g., Cambridge and Pearson) and which will be the methodology of choice for the present study. Despite its advantages, however, Rasch methodology is rarely found in the field of second language education outside the specialism of language testing and while programs such as Winsteps are becoming more and more user-friendly,

specialist knowledge is still required. Certainly, in the educational context of the present study, the lack of expertise presently employed in the development of the *selectividad* exam has been highlighted—it is neither piloted nor analysed in any way whatsoever. Clearly, a move towards the use of industry-accepted research methods such as Rasch analysis would be a massive leap forward for the production of a high-stakes test for national university entrance.

5.2.2 Questionnaires

Questionnaires are common research tools due to the fact that they are "relatively easy to construct, extremely versatile, and uniquely capable of gathering a large amount of information quickly in a form that is readily processable" (Dörnyei, 2007, p.101). Bio data collected from a well-targeted pilot study can be used to cross-reference other aspects of the data collection, for example, to examine possible bias due to sex or age. Self-assessment data can also be collected and correlation studies used to provide evidence for the extrapolation inference. Furthermore, attitudinal information can be collected, which is especially useful for a new test, given that candidates are important stakeholders whose views must be taken into account. Indeed, numerous themes can be covered on a questionnaire and so the first stage in its development is therefore the identification of its purpose (Phakiti, 2013). To this end, those themes considered to be relevant to the present study will be briefly discussed here:

1. Self-assessment. One form of strongly-recommended validity evidence is a measure of the same construct taken from another source. In the present study, it would be impractical, impossible even, to administer a completely validated B2-CEFR related listening test to the same group of participants.[52] Nevertheless, a self-assessment measure can be considered to be a just such a different measure of the same construct. One caveat here, however, is that it has been argued that test takers are not capable of providing accurate self-assessments about their proficiency level. For example, reporting on candidates' self assessment of CEFR 'can dos', O'Sullivan (2008) found that without

[52] I did, however, collect information about any accredited qualifications the participants possessed.

training on self assessment, little useful information was collected. Nevertheless, self-assessment has indeed been used despite these limitations as an alternative criterion measure of the same construct in a number of other studies (e.g., Aryadoust, 2013).

2. Opinions about the test. Opinions on the value of test takers' attitudes to tests has varied greatly over recent years. While researchers such as Bachman (1990, p.285-287) have previously argued that attitudinal evidence about a test from candidates simply represents 'face validity' and should not be part of the test validation process, test takers are currently considered to be important stakeholders. Indeed, many consider them to be perhaps the most important stakeholders of all, as they are the ones most effected by the results from a high stakes test (Brown, 1993; Hamp-Lyons, 2000). Despite such recognition, however, they are often the group whose opinions are least listened to and whose views should consequently be taken further on board—especially as part of any impact study (Hamp-Lyons, 2000). An important link in any VA concerns test consequences and impact; here, a full impact study—including all stakeholders—would obviously be impossible as the test itself has not actually been implemented. However, an attempt can and should be made to discover candidate opinions about the implementation of a CEFR-related listening test. An important concern here is whether or not such a test would indeed be welcomed by the candidates themselves, as it may be argued that a candidate's rejection of a test is a serious rebuttal of any validity argument.

Test takers can also provide invaluable evidence concerning the actual mechanics of the test—especially at the piloting stage. Green (2017) strongly recommends feedback questionnaires for listening tests, suggesting test developers ask test takers about familiarity with topics and test methods, perceptions of interest and difficulty, as well as any aspects which could lead to construct irrelevance, such as the quality of instructions, sufficient time allocations and opinions about test fairness. Once collected, these opinions can be shared with other stakeholders and interested parties. Test-taker feedback questionnaires can therefore be considered to be an important part of the test development cycle. ALTE (2011) recommends their use at the piloting stage and even

provides an example questionnaire in appendix VI, and many exam boards now include their implementation.[53]

3. Opinions about process and strategy use. As previously mentioned, questionnaires have been widely used to collect information about process and strategy use when solving test items. Indeed, Messick (1989) explicitly recommends test taker perceptions as a source of evidence for construct validity. If test takers believe that a test actually does test what it purports to, such a belief will clearly add to the strength of the warrant for the evaluation inference. One previously validated questionnaire, the *Metacognitive Awareness Listening Questionnaire* (Vandergrift et al., 2006), has been used in a number of studies of strategy use in listening comprehension. As the present study uses verbal report methodology to gain detailed insights into test taker processes, it was consequently decided to limit this part of the questionnaire to opinions about the performance of CEFR 'can do' descriptors, as well as perceived meta-cognitive strategy use. If candidates believe they are performing the CEFR descriptors for B2 listening, this provides evidence that the test is indeed testing what it purports to test, and certainly adds to face validity. A similar study was presented by Szabó and Märcz (2012), using the term 'interface validity', whereby opinions on content validity were based on the CEFR 'can do' descriptors at B2.

Once the themes of the questionnaire have been decided upon, we can move on to the planning stage. Here, considerations must be taken into account not only about practicality and ethics (Phakiti, 2013), but also the target population and the sampling techniques to be employed. In the present study, it was decided to administer the questionnaire to the whole sample population directly after test completion and participants were provided with both an explanation of the purpose of the questionnaire and an ethically-motivated consensus box to tick to indicate their willingness to participate (see Dörnyei, 2007 for discussion).

---

[53] For example, Zeng and De Jong (2011) outline candidate reactions to the new Pearson test in order to feedback the results at the test design stage. Although the results are not reported they claim to have received positive reactions to the test and any problems which were highlighted led to test revision.

Questionnaire content needs to be decided upon based both upon the themes that have been identified and on the types of questions which will be included (whether they be open, Likert-scale, rank order, etc.). If the questionnaire needs to be translated, it is essential there are quality control checks put into place to ensure the translation is accurate. It needs to be shown that answers are consistent and reliable, and here the most normal check would be to report Cronbach's Alpha. However, if the questionnaire is divided into sections, with each section intending to measure qualitatively different aspects, then we would not expect the results to have high reliability estimates when measured together and the reliability of each section should be measured separately (Brown, 2001).

Questionnaires, however, do have certain limitations, such as bias and the fact that reliability estimates do not imply a valid questionnaire (Phakiti, 2013). Also, some data may well be missing dues to respondents not completing the whole questionnaire. Despite such shortcomings, however, questionnaires are nevertheless good research instruments for collecting large amounts of information quickly and cost effectively, and in most research contexts provide a useful tool for triangulation purposes.

5.2.3 Verbal Protocols

'Think alouds' and retrospective methods of verbal report can be used in language testing in order to better understand test taking cognitive processes, and have been used extensively for test-taking strategy research (Cohen, 2014). In test validation research, the methodology is therefore extremely useful for answering questions such as, "does the test engage the abilities it intends to assess?" (Xi, 2008, p.186), one of the research questions in the present study. Gass and Mackey (2000, p.13) define verbal protocols as the data one gets "by asking individuals to vocalise what is going through their minds as they are solving a problem or performing a task".

Our listening ability model (presented in chapter 3) shows that before a candidate completes a test task, they are expected to engage in the metacognitive strategy of

planning. Evidence of what processes are being used at this stage can therefore be collected at the pre-listening stage, using think aloud methodology. Concurrent think alouds are classified by Cohen (2000, 2007b) as a 'self-revelation' method; the participant vocalises exactly what they are thinking whilst they are carrying out a language test task.

Listening, however, is an online process, making it difficult to investigate (Vandergrift, 2007). It is impossible to collect think aloud data as a participant is performing a task, we must instead use retrospective methods, such as stimulated recall (Gass & Mackey, 2000). By providing the participant with cues or prompts—such as the test paper, answer sheet and notes—while the participant is finalising their answers, it is hoped that thoughts and processes can be collected whilst they are still in the short term memory. It is a method of introspective 'self observation' (Cohen, 2007b) and has been recommended as a research tool to access the complex cognitive processing test takers engage in when listening (Vandergrift, 2007). Many studies also replay the audio file, stopping at pre-decided points in order to aid recall (T. Brunfaut, personal communication, September 2013). However, in my own study (Shackleton, 2014), it was found that participants were able to understand much more on this third play, often changing a previously incorrect answer to a correct one. These findings were further supported in the present study during the trialling of the methodology, and it was therefore decided to omit this phase when collecting the stimulated recalls.

By tapping into thought processes, this methodology can also be used to investigate construct-irrelevant factors which are not immediately apparent when simply looking at test scores (e.g., Field, 2012b; Yi'an, 1998). This evidence is important, as any interpretation of test scores would not be as valid if some participants are shown to be using irrelevant knowledge, processes or strategies (Xi, 2010). Indeed, the methodology has been successfully applied to examine just what each item on a test is testing; used in conjunction with item statistics, the researcher can subsequently identify reasons for badly performing items (see for example Ancker, 2007; 2011, although she used written retrospective reports). For this reason, Fulcher and Davidson (2007) recommend that the

methodology be used at the test development stage, rather than simply as a method of post priori test validation.

Unlike concurrent 'think alouds', which are not open to reactivity (Ericsson & Fox, 2011), with retrospective 'think alouds' the researcher can ask probing questions in order to encourage the participant to give more useful information. Once the recordings have been collected, they must be transcribed, segmented and coded—a lengthy process. However, a number of computer packages have been developed to aid this coding process, which is typical in many QUAL research methodologies. For example, in *Nvivo* a 'node' is used to represent a categorical label which has been given to a meaningful segment of data. The coding scheme can be developed in a number of ways: for example, following *grounded theory* (see Dornyei, 2007), which is an inductive approach, the patterns emerge from the data. Alternatively, using the approach which will be taken in the present study, we can develop our codes based on a theoretical framework (Gu, 2014). In the present study we wish to demonstrate that the relevant KSAs described in our listening ability model are used to solve test items. It therefore makes sense then to develop a coding scheme based on this theoretical model.

If we want to make generalisations about our results, the approach to coding and analysing the data should be systematic and reliability checks should be carried out, including whether or not another coder would code in the same way. The resulting data can be analysed both qualitatively and quantitatively by using frequency counts. However, other quantitative analysis could be problematic, as the data sets are small and will probably not follow a normal distribution, so non-parametric measures need to be used.

It should be noted here that the method is not without its limitations; for one thing, due to the fact that the data collection is extremely time-consuming, only small samples are typically used (Green, 1998). Furthermore, stimulated recalls rely heavily on memory and reports may be incomplete or inaccurate (Banerjee, 2004). For example, a proficient user may not report a strategy simply because that strategy has become an automated

process (Phakiti, 2003), a finding confirmed by my own study (see Shackleton, 2014). Here, the higher ability participant—the one that scored highest on the test—directly reported a *text level understanding*. Very few strategies were reported and so it could be inferred that this participant had a more automated listening ability. This concurs with the literature; strategies become automated processes as proficiency increases (Saville-Troike, 2005). This has implications for the present study, which will probably find fewer instances of strategy use reported by the quite proficient B2 level participants who took part in the study. Despite these limitations, the methodology is preferable to other research tools which are used to investigate processes when answering test items. Questionnaires have been widely used (e.g., Vandergrift et al., 2006), but it is argued that they are based on those strategies respondents think they have used (Field, 2012b). Expert judgement, another methodology used in the present study, also has drawbacks (see Alderson (1993) for discussion on limitations).

5.2.4 Content standards and expert judgement

In order to demonstrate the match that the test is actually testing what it claims to be testing, some content-related evidence is necessary. In other words, a test must be shown to represent the proposed construct and have good construct coverage. As test tasks do not represent the complete universe of possible tasks for a domain, we must show that a representative sample of tasks has been included if we are to be able to extrapolate performances and generalise about the universe domain; in other words, the observed scores must allow us to provide an accurate estimate of the latent trait being tested (Kane, 2006). Indeed, Xi (2008) states that such evidence is a fundamental contribution to the substantive part of construct validity. Thus far, the present study has provided such evidence by presenting a sound theoretical construct and sampling this construct well through detailed test specifications, which have then been operationalised as the final test tasks.

It is argued, however, that content-related evidence drawn from test providers themselves could have ‘a confirmationist bias’ (Kane, 2006). Typically in language

testing, we have to look to expert opinions to provide supporting judgements about test content. Here, experts who had not been involved in the test development process would be able to make these judgements without bias. Evidence from expert judges that a test does indeed cover the intended construct acts as supporting validity evidence for the observation inference, as a performance standard is not interpretable without a content standard (Green & Inoue, 2016). Subsequently, the results of expert judgements about test items can be compared with test developers' original intentions in order to highlight any similarities and differences between the two groups in terms of the understanding of test content.

Test content and difficulty level should be appropriate for the decisions that are to be taken based on the test results; in the case of the present test, students' proficiency in English expressed as a B2 level based on the CEFR. Test takers need to meet the cut score or pass the test in order to demonstrate that they have reached the required proficiency level and if this cut score is not appropriately set, the results of the test could come into question (Bejar, 2008). The process of setting the cut score should therefore be an integral part of test development (Cizek & Bunch 2007, p.247) rather than a separate phase independent of the development process (Papageorgiou & Tannenbaum, 2016). Here, in 2009 the Council of Europe published *The Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (henceforth *The Manual*), which gives detailed guidelines of how the process of aligning tests to the CEFR should be carried out.

The basis of the CEFR alignment manual is without doubt the use of expert judgements, a methodology which is widely used and recommended in the standard-setting literature. However, methodologies using expert judgement have been reported to be unreliable, arguably due to the fact that the CEFR does not provide sufficient and precise descriptions of proficiency levels (see for example, Alderson et al., 2004, 2006; Fulcher, 2004; Weir, 2005a). Judges in an alignment study may interpret CEFR descriptors differently or have their own internalised idea of just what it means to be at a CEFR level (Eckes, 2012; Harsch & Hartig, 2015; Papageorgiou, 2010). Indeed, North

and Jones state:

> No amount of CEFR familiarisation and standardisation, or estimation of indices of consistency and agreement, will prove that a given group of experts judging the level for a given language are not bringing their own culturally determined interpretation to the task.
>
> (North & Jones, 2009, p.16)

I will now go on to examine the guidelines given by the Council of Europe in order to relate tests to the CEFR in order to incorporate them into the present study.

5.2.4.1 Relating examinations to the CEFR

Since the publication of *The Manual* numerous CEFR-alignment studies have been published (see for example, Brunfaut & Harding, 2014; Figueras & Noijons 2009; Kanistra & Harsch, 2017; Martiyniuk 2010; Tannenbaum & Wylie, 2008). This manual sets out the necessary stages which it states should be followed in order to build an argument for relating test scores to the CEFR. These stages will be followed in the present study and include:

1. Familiarisation. All participants involved in a linking study should be familiar with the different CEFR levels. Training activities should be given and an assessment of the results of this training should be reported. Here, a useful resource is that provided by the *Ceftrain* project.[54] Once judges have carried out familiarisation exercises, it is recommended that they are given CEFR descriptor sorting exercises whereby they allocate CEFR descriptors to levels. Results from such exercises can then be used as evidence that the judges have a sound understanding of CEFR levels. The results can also be used for discussion about salient features of the CEFR levels in order to ensure a common understanding of just what is meant by having a CEFR B2 listening ability.

[54] see http://www.helsinki.fi/project/ceftrain/index.html.

2. Specification. This relates to specifying test content and construct coverage. *The Manual* includes numerous forms and grids for users to describe the test they wish to link to the CEFR. These grids can also be used to communicate important information about the test to stakeholders, which allows for comparison with other tests claiming to test the same level (Harsh, 2014). However, it has been argued that very few test providers pay attention to this stage of CEFR linkage (Green, 2017; Green & Inoue, 2016), and that they instead concentrate on standard-setting procedures. Green and Inoue (2016) report a study for relating speaking exams to the CEFR in which CEFR-alignment as proposed by *The Manual* is followed and special attention is given to the numerous forms which should be completed at the specification stage. In the Trinity CEFR alignment study, however, Papageorgiou (2009) found that judges had difficulties with aligning tasks to CEFR descriptors at the specification stage, mainly due to the wording of the descriptors. The participants also reported that they found the process of filling in all the presented specification forms a tedious endeavour. Similarly, Moe, (2009) reported that there was not enough explanation about what the characteristics of the items should be for each CEFR level. Indeed, the CEFR itself was not developed to provide detailed specifications for assessment purposes; rather, it gives us a starting point—a design pattern (Davidson & Fulcher, 2007), which is supposed to be context free. The context of a particular test and the TLU situation it is supposed to represent should also be taken into consideration during any content analysis study.

The present study uses the CEFR as a framework for test development, specifically for measuring CEFR B2 listening ability. The researcher has had extensive CEFR-related training, which included familiarisation, specification and standardisation similar to that reported by the Pearson linking study (De Jong & Zheng, 2016). As such, this is not a post-hoc linking study but instead forms part of the test development project, the standard having already been explicitly built into the test. Furthermore, the process of standardisation using the 'CEFR content analysis grid for reading and listening' (CoE, 2009, p.154) is a process which resembles a Basket Method standard-setting study, as participants have to estimate at what CEFR level each item is comprehensible. By including a Basket Method type study, it may then be argued that the specification stage

of CEFR alignment is followed to some extent and that a priori intentions can be compared with panelists' opinions.

It was further decided to provide the participant judges with a detailed explanation of the test specifications and research context—that of school leaving/university entrance. A Basket Method study would also be carried out and in this way the test developer's intentions and a priori decisions could be compared and corroborated with expert judges and provide further validity evidence towards the domain modelling and the observation inferences. Once the specification stage of CEFR alignment is completed and the judges are well versed in the content of the test, the standard-setting study to determine cut scores can be carried out.

3. Standardisation. It needs to be shown that the judges are able to relate task difficulty and candidate ability to the CEFR descriptive levels. They also need to be familiar with any standard-setting procedures which will be used. Judges therefore need to partake in training exercises before the standard-setting procedure begins. In the present study, all participant judges had recently taken part in a full standard-setting study in which they related a listening test to the CEFR (see Shackleton, forthcoming) and could consequently be considered to have been well trained in the procedure.

4. Standard Setting. For any test—even those developed to test just one proficiency level—a cut score or score which is needed for passing the test needs to be decided. This score should not be a normative standard e.g., 60%, but should be defensible as a representation of the performance standard (see for example, Wolfe & Smith, 2007a). We would not normally expect candidates to answer every item on the test correctly but we need to decide the minimum score necessary which represents mastery of the proficiency level tested, in this case a CEFR B2.

Numerous methods of standard setting are explained in *The Manual* and many more can be found in the literature. As it has been widely reported that different standard-setting methods yield different cut scores (e.g. Kaftandijeva, 2010), it is therefore

recommended that a combination of methods should be used in order to triangulate results. Although the fact that different standard-setting methods return different results has been well documented, I would argue that repeated standard-setting exercises on the same test should return the same, or at least similar results, if we are to have confidence in this standard. The methods chosen should be the most relevant and feasible for the given context and detailed documentation about the process followed should be provided (CoE, 2009).

Judges essentially assess each item on the test and (in the case of most modern standard-setting methods) use empirical data about candidate performance on the items to inform their judgements. They then discuss and revise their decisions in order to identify the test cut score with the aim being to decide on a final cut score which is both defensible and reproducible. The literature classifies standard-setting techniques into two main types: test-centered and examinee-centered. However, it can be argued that all methods are a combination of the two, as we always take into account information about the student (either real or hypothetical) and information about the construct and test content. The only information we have about the candidates in the present study is a self assessment of CEFR level, mark received on the *selectividad* exam (which does not include a listening section) and whether or not the candidate is in possession of an official CEFR-linked accreditation exam. It was therefore decided that the main study would use a test-centered method but the above examinee information would be used for comparison to give some supporting triangulation evidence and therefore evidence towards external validity.

Of the various test-centered methods found in the literature, it was decided not to use *Angoff-type* probability methods because of problems reported concerning judges inability to understand and correctly articulate conditional probability (Hambleton & Jirka, 2006; Reckase 2010). These methods are also very time consuming. Furthermore, newer methods of standard setting such as the *ROC-curve* method have also been criticised because of the large misplacements of cut scores and standard errors and the complexity of their statistical methods, which makes the communication of results less

acceptable to a wider audience (Kaftandjieva, 2010) and thereby lessens procedural validity.

A valid claim for CEFR linkage must show that the test is reliable and representative of proficiency at a CEFR performance level, as "if an exam is not valid or reliable, it is meaningless to link it to the CEFR" (Alderson, 2012). Consequently, the process is normally carried out late in the test development cycle (Wolfe & Smith, 2007b), i.e., once field studies and trials have been carried out. This part of the study will therefore be carried out once the final test form has been decided upon.

As in Shackleton (forthcoming), the present study used a combination of 1) the *Basket Method* and 2) the *Bookmark Method*, not least because the participant judges had already received training in these two methods. As previously mentioned, the Basket method was used to specify test content in a finely-grained manner. Following the previous study, to my mind, the Bookmark Method (Mitzel, Lewis, Patz, & Green, 2001) most easily reflects the continuous nature of the CEFR scales as well as being an ideal method to use with the Rasch calibrations produced in the present study. I will now go on to explain what these standard-setting methods entail.

The Basket Method, developed by Alderson (2005) during the DIALANG study, is essentially an *item descriptor matching* method where items are classified (or placed in a basket) based on substantive theory, in this case the CEFR model of language proficiency. As mentioned earlier, it is therefore a very similar process to that carried out during the specification stage of CEFR linking. Judges go through test items and answer the question 'at what CEFR level must a test taker be in order to answer this item?' thereby determining the minimum requirement for reaching the standard. The method is considered to be the most simple and practical of all standard-setting methods (Kaftandjieva, 2010) and is one which reflects the importance of the performance level descriptors and places emphasis on test content. However, no information is given about the difficulty of the items and as such one of the main problems with this type of method is the lack of consistency with empirical difficulty measures (Kaftandjieva, 2010).

Furthermore, it has also been reported that the method can yield more lenient standards (CoE, 2009).

The Basket Method was used in Shackleton (forthcoming) as a primer prior to the main study, essentially performing the task of the specification stage of CEFR alignment, due to the fact that the researcher felt the two activities had a high degree of overlap and were essentially repeating each other. For the present study, it was decided to include the method in the main study in order to answer research question six (R6), thereby incorporating part of the specification stage with the standard-setting stage. Two rounds were conducted, allowing for discussion and the presentation of normative data in order to promote discussion and gain a detailed understanding of test content.

The decision to use a second method was made due to the fact that for standard setting to have more meaning in the setting of cut scores (especially for a test development project based on Rasch theory), judges must have access to statistical information about how test items perform. The Basket Method is a solely judgemental method and much has been reported about the inability of judges to decide on item difficulties (e.g., Alderson, 1993). It was therefore decided to also use the Bookmark Method and thus give the judges information from the Rasch analysis of test scores about item difficulty parameters.

The Bookmark Method is a popular standard-setting method whereby emphasis is placed on test content and item difficulty and which lends itself to tests which have been developed using Rasch measurement, where test taker ability and item difficulty are placed on the same scale. Here, the items are placed in an ordered item booklet (OIB), ordered according to the difficulty parameters of the items. Judges must go through this booklet from easiest to most difficult item and place a bookmark at the place where they believe a minimally competent candidate will have less than a specified probability of giving a correct response. In the present study, this minimally competent person would be considered to just have a CEFR B2 listening ability. The candidate is considered to have the ability to master an item in probabilistic terms, known as the response probability

(RP).[55] This probability should be decided in advance, but it is often set at 0.67 (Reckase, 2006), i.e., a 67% or 2/3 chance of getting the item correct. When the Rasch model is used, the ordering of the items in the OIB is exactly the same as ordering the items according to their difficulty parameters (Reckase, 2006). The bookmark is therefore placed between two items on the latent scale represented by the Rasch logit difficulty parameters of the items (β-parameters). In the literature, the use of the test characteristic function to convert latent values to scores is recommended. Therefore, once a cut score has been decided, it is mapped back to the Rasch θ -scale using the test characteristic curve from the set of items. As in other methods of standard setting, more than one round is normally conducted, allowing judges to receive normative and impact data before final decisions are made. Once each judge has reached a cut score decision, the median is taken to give the group standard on the latent variable. Although CoE (2009, p80) recommends taking the lower value of theta for each judge, Reckase (2006) recommends averaging the locations on either side of the bookmark.

Once a final cut score has been decided and the score has been transformed into a candidate ability measure, it can be used as the cut score on future versions of the test—as long as the test has been developed following the same specifications and anchoring, linking and equating techniques are used (see Kolen & Brennan, 2014; North & Jones, 2009; Wright & Stone, 1979). This process, if correctly followed, therefore contributes validity evidence for the 'generalisation inference'.

5. **Empirical validation**. The test should be a valid and reliable measurement instrument and its relationship to the CEFR should be supported by statistical data. It is pointless to try and link an exam which does not have good supporting validity evidence (Alderson 2012, CoE, 2009) and this part of the CEFR linking project will have initially been addressed by previous research questions. With particular regard to the standard-setting process, CoE (2009) outlines the evidence that must be provided in relation to the following three aspects of the process: procedural, internal, and external validity evidence

[55] The notion of RP needs to be well explained in the training phase (CoE, 2009, p.78).

(this follows Kane, 1994). Criteria for evaluating standard-setting studies for these three aspects of validity are given in Hambleton and Pitoniak (2006).

Procedural validity of standardisation and standard setting can be provided by giving evidence of the explicitness of the procedure; ideally the whole procedure should be replicable. The procedure must be practical, providing easily interpretable results. During the implementation of the procedure, evidence must be given as to how the judges were chosen and how the judgement data was handled and at the end the judges should be confident in the resulting cut scores. Much evidence relating to procedural validity can be provided by administering a post standard-setting questionnaire and presenting the results (Cizek, 2012), and it was therefore decided to do so for the present study. The entire procedure must be well documented in order to communicate results so that stakeholders can evaluate those results and the final results must be communicated in a way which can be understood by stakeholders, i.e., they must be interpretable (Wolfe & Smith, 2007a). It is the author's intention that the information documented in the present thesis, which will act as the technical documentation for the initial linking study, serve precisely this purpose.

Internal validity concerns the accuracy and consistency of results and here the quality of the judgements needs to be both determined and reported. Indeed, besides the consideration of results and of how items have been allocated on the latent trait proficiency scale, Cizek (2006) advises that studies should collect evidence of the classification accuracy, that is, evidence to show that the classifications would be replicated using the same standard-setting method. This could include inter-rater and intra-rater reliability studies. Here, simple correlation analyses are not appropriate, as "it is possible to have a perfect correlation of ± 1.00 between two judges with zero-agreement between them about the levels to which descriptors, items, examinees or their performances belong" (Kaftandjieva, CoE, 2004, p. 23). Indeed, just how rater agreement or consistency should be conceptualised is a matter of some debate. Stemler (2004) argues that one single umbrella figure for reliability does not give enough information

about a specific context and that inter-rater reliability is a function of an assessment situation. Ideally, therefore, different measures should be given. Examples include:

1. Consensus estimates: Exact agreement between raters (percent agreement figure), Cohen's kappa.

2. Consistency estimates: e.g., Spearman's rank order correlation coefficient (Spearman's rho), Wilcoxon Signed–Ranks or Cronbach Alpha, which is one form of Intra-Class Correlation coefficient.[56]

3. Measurement estimates: e.g., Factor analysis such as Principle Components Analysis, applying Generalisability theory or using a Many-Facet Rasch measurement model (MFRM).

Indeed, Linacre (2017b) states "there is no generally-agreed index of inter-rater reliability. The choice of method depends on the purpose for which the ratings are being collected, and the philosophy underlying the rating process."

MFRM, an extension of the Rasch model using the statistical program FACETS (Linacre, 2017b), allows us to report both consistency of judges (intra-rater reliability) and agreement among judges (inter-rater reliability). It is useful for looking at scores based on a number of facets, such as in the present study where we have items, judges, rounds and CEFR scales. Linacre (2017b) also gives some outlines as to which measures are preferable depending on the context. In the context of standard setting, I would argue that the judges are acting as independent experts, with each judge bringing something to the table. Indeed, Stemler (2004, p.9) states that "measurement estimates are best used when different levels of the rating scale are intended to represent different levels of an underlying unidimensional construct". This is relevant for the present study, as we have a CEFR scale which represents the unidimensional construct of *listening ability*. As

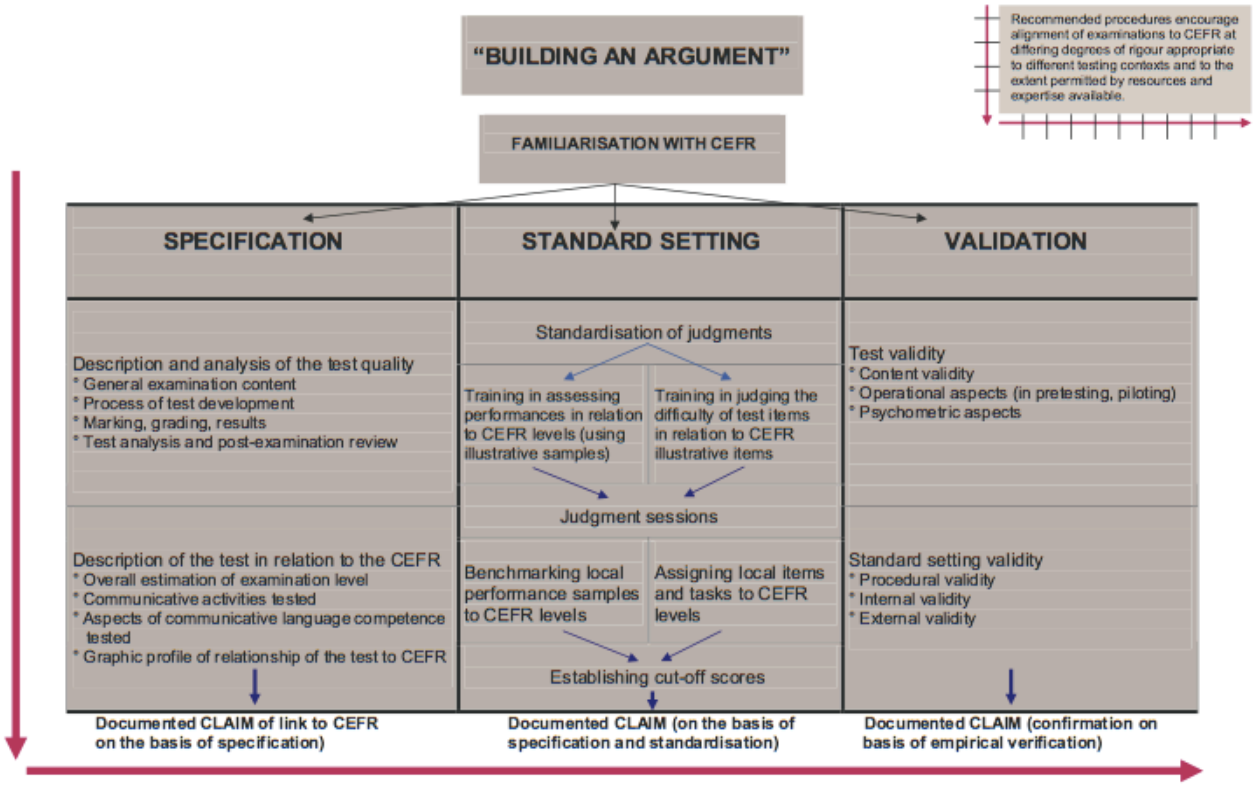
[56] When scores are based on combined ratings coefficient alpha should be reported as a measure of the reliability of the scores (Bachman, 2004).

Engelhard (2011, p.913) explains "the goal within the framework of Rasch measurement is to obtain consistent ratings from the standard-setting panelists rather than perfect agreement on each rating for each item". Other advantages of the methodology outlined by Stemler (2004) include the fact that errors can be taken into account for each judge—leading to a more accurate representation of the construct—and that ratings from multiple judges can easily be used. This methodology is therefore considered superior for evaluating the quality of judgements and has been used in other standard-setting studies (e.g., Engelhard, 2009, 2001). As a result of these observations, it was decided to use the MFRM FACETS program to evaluate the quality of the standard-setting judgements for the Basket Method and then compare results with empirical Rasch difficulty measures in an attempt to create stronger linking with the CEFR.

The Many-Facets Rasch Model (MFRM) is another model in the Rasch family that is used when facets in addition to person ability and item difficulty need to be measured. In the case of expert judgements of test items, the facets would include raters, items and round, while the raters are treated as independent experts and the model gives information about how similar or different their ratings are. The model also provides 'fair average scores' by taking into account the leniency/severity of the raters so final allocations of CEFR level can be based on the opinions of all the judges. This analysis also gives information about both inter- and intra-rater reliability. However, it has been argued that final round judgements rarely evidence much variability because judges are encouraged to converge and come to a consensus decision (Linn, 2003). In the present study, the MFRM analysis is used to report the specification stage in order to answer R6 rather than the final cut score decisions.

External validity refers to the use and comparison of different standard-setting methods, as well as comparisons or triangulation with measures from other studies—both of which will be included in the present study. As well as the two standard-setting methods proposed above, results will further be compared with candidate self-assessments. The manual gives a pictorial representation of how these phases of the linking process are related and is reproduced in Figure 14, below.

Figure 14. Visual representation of procedures to relate examinations to the CEFR (CoE, 2009, p.15).



5.3 Data collection and analysis

Having considered the different methodologies which will be employed in the study, I will now go on to outline the data collection methods, the description of participants and the procedure to be followed for each of them in order to analyse their data.

Data was collected from the following sources:

1. Test scores
2. Questionnaires
3. Verbal reports
4. Expert judgements

The methods of data collection and storage adhered to the ethical research practice guidelines required by Granada University. All students were informed that the study was being conducted purely for research purposes, were given information about the nature of the research project, and were required to tick a box on the test and questionnaire demonstrating their agreement to take part. Each test/questionnaire was numbered directly after administration and as such, the subjects' identities were kept anonymous. Participants' names were, however, collected upon completion of the test and noted against their candidate numbers in order that their the scores could be reported to them by their teachers, as it was felt that this would be a factor which would not only motivate participants to take part in the study, but also to take the test seriously and perform to the best of their ability. Nevertheless, no names were kept on file. Likewise, all subjects who took part in the verbal reports volunteered for the study, were given a consent form and information sheet (see appendix 1), and did not have their names reported. The expert judges were all colleagues from the Centro de Lenguas Modernas (CLM), Granada University, who volunteered to take part after being given an explanation of the purpose of the study.

5.3.1 Test scores and questionnaires

The 33-item test was given to a large sample (N= 153) of first year students studying for a Degree in English Studies at the UGR. The test was administered to these students in October 2016. As such, most of the participants had done the *selectividad* exam for university entrance only three months previously and could be considered to be representative of the target population. The fact that the students had chosen to go to university and study English philology was considered to be indicative of the fact that they were interested in the subject and that they should be either at or around a CEFR B2 proficiency level in English. Directly upon completion of the test, the same participants were asked to complete the questionnaire (see appendix 2). Participant bio-data is shown in Table 5 below, where it can be seen that a large proportion of the participants were female. It is also evident that most of the participants were under 21 and were very much representative of the target population.

Table 5. Participant bio data: Crosstabulation of age and gender showing percentages within each gender.

Age	Gender		Total
	Male	Female	
Under 18	2 (5.1%)	18 (15.8%)	20 (13.1%)
18 - 21	35 (89.7%)	93 (81.6%)	128 (83.7%)
21 - 25	2 (5.1%)	2 (1.8%)	4 (2.6%)
26 - 35	0 (0%)	1 (0.9%)	1 (0.7%)
Total	39 (25.5%)	114 (74.5%)	153 (100%)

5.3.1.1 Analysis of test scores

On completion of the test, participants were required to transfer their answers to an answer sheet (see appendix 3). The researcher marked the answers for Task 4 (the NF task) and the answer sheets were then fed through an optical reader. The results of the test were then entered into IBM SPSS version 20 and an initial analysis was carried out using descriptive statistics and classical item analysis. Due to the aforementioned limitations of CTT, however, the main analysis was carried out using Rasch in version 3.71.0.1 of the Winsteps program (Linacre, 2017a). The Rasch analysis, which includes a distractor analysis, allowed for initial decisions to be made about which items were not functioning correctly and so should be dropped from the test. The badly-functioning items were further analysed through the results of the verbal reports and a final decision was made as to which items should constitute the final 28-item test. The results of the final version of the test were then re-analysed to produce the Rasch results as evidence of the correct functioning of the test in order to answer research question four (R4). A PCAR analysis

was also carried out in order to examine the unidimensionality of the test and answer research question two (R2).

Information was collected through the questionnaire about candidates' self-assessment of their CEFR level. This information could then be used as an external criterion measure to be correlated with test scores. These alternate ability estimates could be compared with total mean score in order to give supporting evidence to the extrapolation inference and answer research question eight (R8). Here, my statistical research hypothesis is:

H1: $\mu C2 > \mu C1 > \mu B2 > \mu B1 > \mu A2$

Null hypothesis: No relationship exists between scores on BFE listening test and listening proficiency level. That is, there will be no difference between the five means.

H0: $\mu A2 = \mu B1 = \mu B2 = \mu C1 = \mu C2$

A similar analysis was carried out with results on the BFE test and reported accreditation exams held by participants, along with results on the *selectividad* test (information which had been collected through the questionnaire).

5.3.1.2 Analysis of questionnaires

The questionnaire (see appendix 2) was delivered in English as it was believed that the level of English of the participants would be sufficient to understand the questions. It was divided into four sections which included bio-data, along with questions covering the themes outlined in the previous section. These themes covered:

1. Opinions about listening instruction at school and whether a listening section on the school leaving exam would be welcomed, information which could be used to answer research question five (R5).

2. Feedback about the test itself in order to discover any construct-irrelevant variance, such as quality of audios, clarity of instructions and time allowed for reading and answering test items. This information can be used as evidence towards the evaluation inference and therefore contributes to answering research question two (R2).

3. Opinions about listening strategy use and whether specific CEFR B2 listening descriptors were elicited by the test. These results can be used as evidence towards the explanation and extrapolation inferences and contribute to answering research question 3 (R3).

The questions included a four-point Likert scale to eliminate the possibility of neutral responses. Response categories contained two positive and two negative responses. The categorical data collected in the questionnaires was coded and reported as descriptive statistics using SPSS. Further analysis was carried out using correlation analysis and comparing responses to questionnaire items and mean score on the test when this was felt to be appropriate. An open-ended question was also included to allow participants to add any extra comments which they felt to be relevant.

5.3.2 Verbal reports

Due to the practical constraints imposed by the time-consuming nature of this methodology, only a very small sample was used. Each session in the present study took about one hour fifteen minutes. Dörnyei's (2007) advice to use 'purposeful' sampling was followed, and my seven respondents were volunteers taken from my own groups at the CLM who I considered to have a B2 level proficiency in listening. The methodology had already been piloted with two students, and this helped me to decide upon my final two-stage design:

1. Concurrent 'think alouds' were collected at the planning stage while students were reading instructions and items. Here, students were asked to simply verbalise their thought patterns. In this way, it was intended to identify any planning and prediction

strategies as well as any problems they may have had with task instructions or information in the items.

2. Retrospective verbal report after finalising answers: subjects did the tasks as if in a test situation (i.e., listening twice with note-taking permitted), followed directly by immediate retrospection whilst finalising answers. Stimulated recall is fairly self explanatory (Xi, 2008). In the present study, subjects were given a short explanation of what was expected, yet no specific training was given for fear of biasing results (Buck, 1991; Gass & Mackey, 2000). Following Buck (1994), they were asked to repeat the content of the section they had just heard in order to assess comprehension, as well as any strategies they had used so as to decide on the answer to the items.

In order to address the issue of time lag between doing the test and reporting, it was decided to do each of the four tasks separately (each sound file is between 3 and 5 minutes long). Banerjee (2004) suggests that respondents should be allowed to answer in L1 as this would allow them to clearly express their thoughts and reduce the cognitive load of reporting. However, although participants were given this possibility, only one of them (Participant 4) chose to report in Spanish and the other six participants preferred to report in English (this may well have been a result of complications and extra cognitive load introduced by working simultaneously in parallel languages). Each report was collected in exactly the same way and the fact that the system is replicable increases the validity of the study (Rost, 2011, p.274). Once the reports had been collected, they were transcribed and entered into the qualitative analysis program *QDA minor lite* ready for coding. The concurrent reports collected at the planning stage prior to the test were analysed qualitatively and examples are given as evidence of any conclusions which were drawn.

As previously mentioned, it was decided that the coding scheme should reflect my listening ability model. The retrospective reports were coded separately for each item on the test and the codes represent the level of understanding which was reached in order to answer the item correctly. This follows Field's (2008a) psychological model which

distinguishes the main stages in developing meaning which vary in depth of cognitive demand on the listener. The following coding scheme was therefore employed:

> **L** – Lexical recognition: The item was answered correctly by only understanding isolated vocabulary from the audio input.
>
> **IU** – Idea unit: A proposition, which could be as little as a noun phrase/an adjective and noun (Buck, 2001, p.27-28), is used to answer the item. This is understanding at a very literal level and includes local factual information.
>
> **MR** – Meaning representation: The listener relates a preposition to the context and draws conclusions which may not be explicitly expressed.
>
> **DR** – Discourse representation: The listener is able to integrate information into a wider picture, including speaker intention.

Following Vandergrift (2003), the results are presented both quantitatively and qualitatively in order to draw conclusions about the listening processes which are necessary to answer the test items. CEFR B2 level listeners are considered to be 'independent' and should have quite automated listening skills. We would therefore expect them to be able to draw on their world knowledge, topic knowledge and the co-text (what has been said before) in order to build meaning from an audio file. Buck (2001, p.21) refers to these processes as 'the cognitive environment' and Field (2013a, p.100-101) includes pragmatic, contextual, semantic and inferential information as part of the meaning construction process. However, Alderson et al. (2004, 2006) specifically mention the lack of level specificity mentioned by the CEFR in relation to process and strategy use. Instead, we are given 'can do' descriptors, which are a taxonomy of behaviours (Alderson et al., 2004, 2006). Here, B2 listeners are expected to understand main ideas and follow conversations and talks. It could therefore be argued that if the audio files are at the appropriate level, they should reach meaning and discourse representations of the input. Indeed, Field states:

> The ability to operate at different levels of processing is a mark of a certain level of expertise: less able L2 listeners are likely to find that their attention is so heavily engaged by processing at word or clause level that wider issues of meaning and discourse structure escape them.
>
> (Field, 2015, p.37)

However, it may also be that, following our listening ability model, listeners employ strategies in order to make sense of an audio file. It was therefore decided to also code any reported strategies and report on these qualitatively. Here, construct-irrelevant strategies would obviously pose a threat to the validity of the test.

In order to provide a meaningful analysis of the behaviours identified in the verbal reports, it needs to be shown that another coder would draw the same conclusions from the data and that the coding system is reliable (Duff, 2006). Mackey and Gass (2005) suggest that a random sample of 25% of the data set should be coded by another coder. As the present study is an individual project, it was decided to re-code one entire protocol six months after the original coding and in this way provide intra-coder reliability. This calculation was effected using both exact percent agreement—useful for nominal data where each category is representative of a qualitatively different idea (Stemler, 2004)—and Cohen's Kappa coefficient, which takes into account agreement by chance.

A specific analysis of problematic items (shown by the statistical analysis of the test results) was also carried out in order to gain insights as to why these items did not function as they should. Incorrect answers were also analysed to gain insight as to why this was so.

5.3.3 Expert judgement and standard setting

Colleagues who had already been trained in standard setting and had been involved in the standard-setting study for the University of Granada B1/B2 bi-level *CertAcles* exam (see Shackleton, forthcoming) were recruited. Unfortunately, as this was a voluntary

study, only eight judges took part. Various recommendations can be found in the literature on the number of participant judges who should be included in a standard-setting study. The recommendation given by the pilot manual for relating examinations to the CEFR is at least 10 judges (CoE, 2003, p. 94), but the final *Manual* (CoE, 2009, p. 38) states that 12 to 15 judges should be considered as the minimum number required. Nevertheless, Livingston and Zieky (1982, p.16) suggest that no less than 5 judges are sufficient.

Table 6 shows information about the participating judges. It can been seen that all the judges are experienced TEFL teachers who have received numerous hours of CEFR-related training and who can therefore be considered 'expert'. This is important, as the participants should have the relevant expertise in both the CEFR framework and hence its performance level descriptors, as well as in the instruction and assessment of the language being tested (Tannenbaum & Cho, 2014).

Table 6. Information about participant judges

Judge	Sex	Qualifications	Number of years teaching TEFL	Number of hours specific training CEFR/Assessment
1	Female	Degree, MA	35	300+
2	Male	Degree	28	80
3	Male	Degree, MA, CELTA	16	70
4	Male	Degree	22	70
5	Female	Degree, MA	36	400+
6	Female	Degree, PGCE, PhD	28	750+
7	Male	Degree, MA, CELTA	18	500+
8	Female	Degree, PGCE, MA	28	750+

This part of the study includes all the steps described in section 5.2.4.1 and results are presented for each stage of the linking process outlined by the Council of Europe (2009). However, the main objectives of the study were twofold:

1. To determine whether or not expert judges believe the test to be an accurate representation of the CEFR B2 BFE listening construct in order to answer research question six (R6). Here the results of the Basket Method are analysed using MFRM. However, as "the CEFR does not describe test properties or item demands and is not

based on a theory of item difficulty" (Harsch & Hartig, 2015, p.334), a further validity check was also carried out to give evidence of the quality of participant judgements by correlating their beliefs about item difficulty with actual item difficulties. This reliability check includes a Kendall's Tau correlation, which is a non-parametric correlation and should be used instead of Spearman's correlation when the data set is small and there are a large number of tied ranks (Field, 2009, p.181).

2. To determine the cut score to be employed on the test as a representation of a minimum CEFR B2 ability level in order to answer research question seven (R7). Here the results of the Bookmark Method are reported. Further analysis in order to give the classification accuracy of the cut scores is reported using the standard error of judgement (SE_J), which is "an estimate of the likelihood of replicating the recommended cut scores" (Tannenbaum & Cho, 2014, p.245). SE_J is given by the formula:

$$SE_J = \frac{SD}{\sqrt{n}-1}$$

According to Cohen, Kane and Crooks (1999, p.364) the SE_J should be ≤ 1⁄2 SEM and so this quality control check was also carried out. Post standard-setting questionnaire results are also reported as evidence towards procedural validity and some external validity evidence is provided by comparing cut score decisions with participants self-assessed proficiency levels.

Chapter 6: Results and Discussion

This section provides a detailed presentation and analysis of my results as they pertain to my research questions.

6.1 R1: What are the statistical properties of the test?

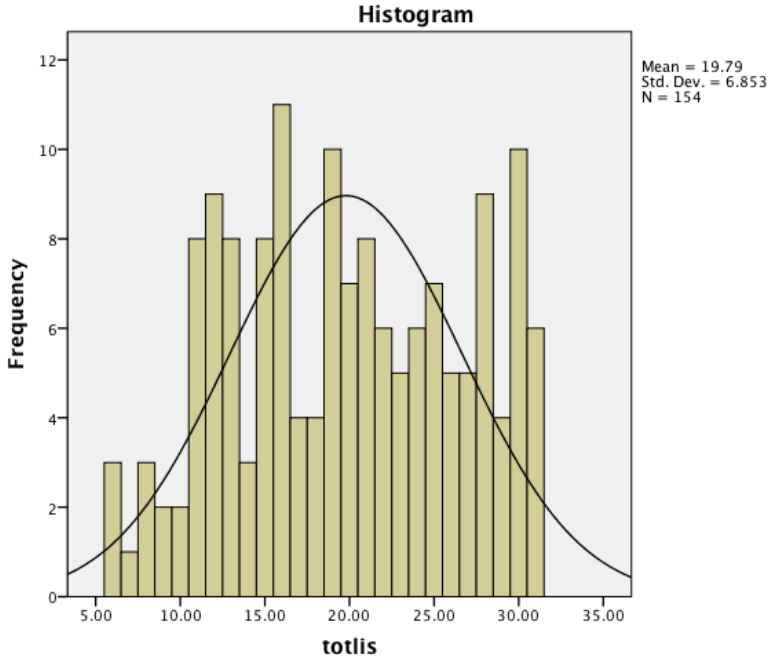
6.1.1 CTT analysis of test results

An initial analysis of the test population distribution using both descriptive statistics (Table 7) and a histogram (Figure 15) shows us that the mean and the median are very similar, with the mode slightly lower at 16. Dispersion of scores shown by standard deviation (SD) and variance shows us that two SDs represents 69% of the total mean score, close to the 68% we would expect in a normal distribution (Green, 2013). Variance is very high (46.96), showing that scores are separating candidates well, also evidenced by the large range. Furthermore, kurtosis is negative (a platykurtic distribution), which indicates a large spread of scores. By dividing skewness by its standard error we obtain the *z*-score of 2.72, which is outside the ± 1.96 range and so points to a non-normal distribution (Field, 2009). The distribution shows a very slight negative skew, which would be expected for a criterion-referenced test which we would expect candidates to pass (Bachman (2004, p.193). The Shapiro-Wilk statistic confirms that the distribution of scores is not normal and so any further analysis with inferential statistics would need to use non-parametric versions.

Table 7. Descriptive Statistics for BFE B2 Listening Test

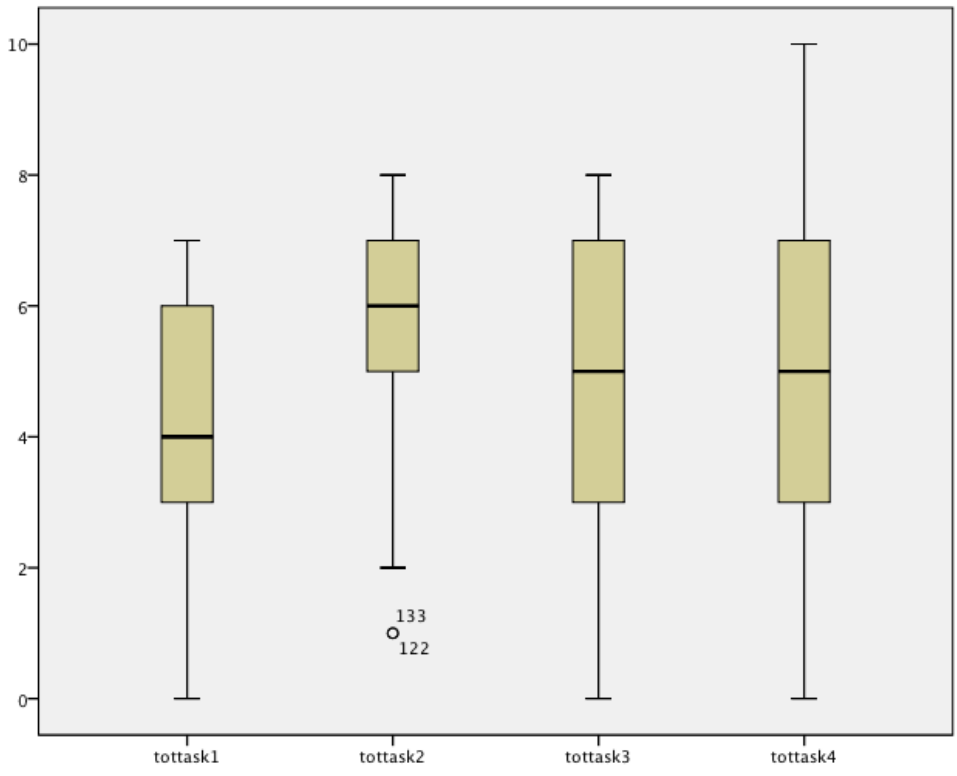
		Total Score
N	Valid	154
	Missing	0
Mean		19.79
Std. Error of Mean		.55
Median		20
Mode		16
Std. Deviation		6.85
Variance		46.96
Skewness		-.04
Std. Error of Skewness		.19
Kurtosis		-1.06
Std. Error of Kurtosis		.39
Range		25
Minimum		6
Maximum		31

Figure 15. Histogram for total listening



Further analysis of the data using box and whisker plots (Figure 16) shows us that the four tasks have similar difficulty (although Task 2 appears to have been slightly easier). There are two outliers for Task 2 and on further investigation it was found that both candidates had low overall scores and only got one item correct on this task. Biographical data shows both candidates claimed their age as under 18 and also seem to be particularly averse to listening. The questionnaire results showed that they believe that listening is not important for learning a language and there should not be a listening section on the *selectividad* exam. Candidate 133 had given extreme negative scores on all other parts of the questionnaire, possibly suggesting that the candidate was not happy with doing the test and had not taken the test seriously.

Figure 16. Box and Whisker plot for the four tasks



The analysis showed that test reliability was good, with a Cronbach Alpha of 0.876. Individual item properties can be seen in Table 8, which shows one difficult item (4.8) with a *Facility Value* (FV) of 16% and three items with low *Discrimination Indices* (DI), (1.3F, 2.8C and 4.10).

Table 8. Item statistics from CTT

	Mean (FV)	Corrected Item-Total Correlation (DI)	Cronbach's Alpha if Item Deleted
Task1.1H	.76	.462	.871
Task1.2B	.60	.314	.875
Task1.3F	.55	.194	.877
Task1.4I	.51	.455	.871
Task1.5C	.63	.504	.870
Task1.6D	.77	.350	.874
Task1.7A	.54	.496	.870
Task2.1B	.89	.315	.874
Task2.2B	.72	.315	.874
Task2.3A	.65	.500	.870
Task2.4C	.64	.488	.871
Task2.5B	.88	.292	.875
Task2.6B	.55	.558	.869
Task2.7C	.69	.303	.875
Task2.8C	.56	.156	.878
Task3.1C	.65	.442	.872
Task3.2C	.77	.492	.871
Task3.3D	.59	.264	.876
Task3.4A	.37	.331	.874
Task3.5A	.35	.380	.873
Task3.6B	.66	.458	.871
Task3.7D	.52	.559	.869
Task3.8B	.83	.305	.874
Task4.1	.38	.463	.871
Task4.2	.50	.530	.870
Task4.3	.53	.437	.872
Task4.4	.37	.471	.871
Task4.5	.44	.439	.872
Task4.6	.70	.313	.874
Task4.7	.88	.353	.874
Task4.8	.16	.461	.872
Task4.9	.51	.439	.872
Task4.10	.62	.157	.878

Table 9. Revised item statistics from CTT

	Mean (FV)	Corrected Item-Total Correlation (DI)	Cronbach's Alpha if Item Deleted
Task1.1H	.76	.452	.873
Task1.2B	.60	.307	.876
Task1.4I	.51	.463	.872
Task1.5C	.63	.494	.871
Task1.6D	.77	.345	.875
Task1.7A	.54	.487	.872
Task2.1B	.89	.313	.876
Task2.2B	.72	.333	.875
Task2.3A	.65	.504	.871
Task2.4C	.64	.502	.871
Task2.5B	.88	.295	.876
Task2.6B	.55	.568	.869
Task2.7C	.69	.319	.876
Task3.1C	.65	.447	.873
Task3.2C	.77	.496	.872
Task3.4A	.37	.331	.876
Task3.5A	.35	.369	.875
Task3.6B	.66	.463	.872
Task3.7D	.52	.554	.870
Task3.8B	.83	.309	.876
Task4.1	.38	.444	.873
Task4.2	.50	.514	.871
Task4.3	.53	.410	.874
Task4.4	.37	.480	.872
Task4.5	.44	.435	.873
Task4.6	.70	.333	.875
Task4.7	.88	.365	.875
Task4.9	.51	.427	.873

Using this information, the 33-item test could be improved by removing the five weakest performing items; the final test form should be a 28-item test following the test specifications. Removing the four items already highlighted along with Item 33D, which had the next weakest DI, gives us a slightly increased Cronbach Alpha of 0.877 and the

individual item properties shown in Table 9. Here, the final 28-item test shows all items to be discriminating well, with DIs ranging from 0.295 to 0.568. The FVs range from 37% to 89%, with the three easy items (higher than 80% FV) still contributing to the reliability of the test.

6.1.2 Rasch analysis of test results

The initial analysis of test scores was carried out using Winsteps version 3.71.0.1 (Linacre, 2017a) for which my place of work owns a license. The data went through six *joint maximum likelihood estimation* iterations before it converged. The standardised residuals were shown to have a mean of 0 and SD of 1, showing that the data fits the Rasch model very well (Green, 2013).

The summary of the Rasch measures is reproduced below in Table 10 and shows person separation to be 2.4; at least two statistically distinct groups can be identified in the test results and the test can successfully separate persons into our two pass/fail groups, with person separation reliability shown to be an acceptable 0.85.[57] The item separation figure is 4.75, showing that the person sample is large enough to be confident about our item difficulty hierarchy (Linacre, 2017a). Item separation reliability is 0.96, which means that we can be fairly confident that the item locations are reliably placed and are reproducible. The person mean ability measure is 0.61—higher than our item mean of 0 set by the model—showing that the participants were comfortable with the test difficulty.

Table 10. Rasch Summary for 33 items

PERSON	154	INPUT	154	MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT		MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	19.8	33.0		.61	.46	1.01	.0	.99	.0
S.D.	6.8	.0		1.23	.10	.15	.8	.38	.8
REAL RMSE	.47	TRUE SD	1.13	SEPARATION	2.40	PERSON RELIABILITY		.85	
ITEM	33	INPUT	33	MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT		MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	92.4	154.0		.00	.21	.99	.0	.99	.0
S.D.	25.5	.0		1.00	.02	.13	1.5	.23	1.4
REAL RMSE	.21	TRUE SD	.98	SEPARATION	4.75	ITEM	RELIABILITY	.96	

[57] Here, Linacre (2017a) states that 0.5 is sufficient for a test that only targets one level.

The real *Root Mean Square standard Error* (RMSE) is close to 1 for the items and the SD shows that real item fit is only very slightly more predictable than expected. Similarly, the mean of the person fit residual is .47 and the SD is 1.13, indicating that persons are more dispersed in ability than would be ideal for the model.

I will now go on to investigate the data in the order recommended by Linacre (2017a). First, we examine item polarity in order to check that we have no negative point-measure correlations, which would indicate that the responses to the item would contradict the latent trait. Such items obviously need to be dropped from any test; because it is the high ability candidates who get the item incorrect while the lower ability candidates answer correctly, such an item clearly does not measure what it is intended to. However, in the present analysis no such items were found and all point-measure correlations were well above the 0.1 recommended by Green (2013). We also need to look at dimensionality, item and person misfits, and item difficulty parameters, which show us whether or not the items are at the correct difficulty level for the population. We must remember that this first analysis is mainly to decide which five items should be dropped from the test in order to produce our final 28-item test. I will therefore first examine item fit statistics, reproduced below in Table 11. During the initial analysis, I deleted four misfitting persons (persons 66, 45, 49 and 15), although this did not in fact have any effect on the item fit statistics.

Individual item statistics are presented below in Table 11. Linacre (2017a) recommends examining outfit statistics first, as the results are often explainable. High outfit MNSQ results could simply indicate carelessness or random responses from low performers. Infit MNSQ values give us the most information as they are less influenced by outliers, they are only influenced by an unexpected pattern of responses near a person's ability estimate. Here I decided to apply the more conservative values of 0.75 to 1.3 (McNamara, 1996). Items with values of less than 0.75 are overfitting as they show less variation than expected by the model while those with values of more than 1.3 are underfitting and are the most problematic as they degrade measurement. These statistics are also reported in their standardised form—the *t*-test significance statistic (Zstd). While

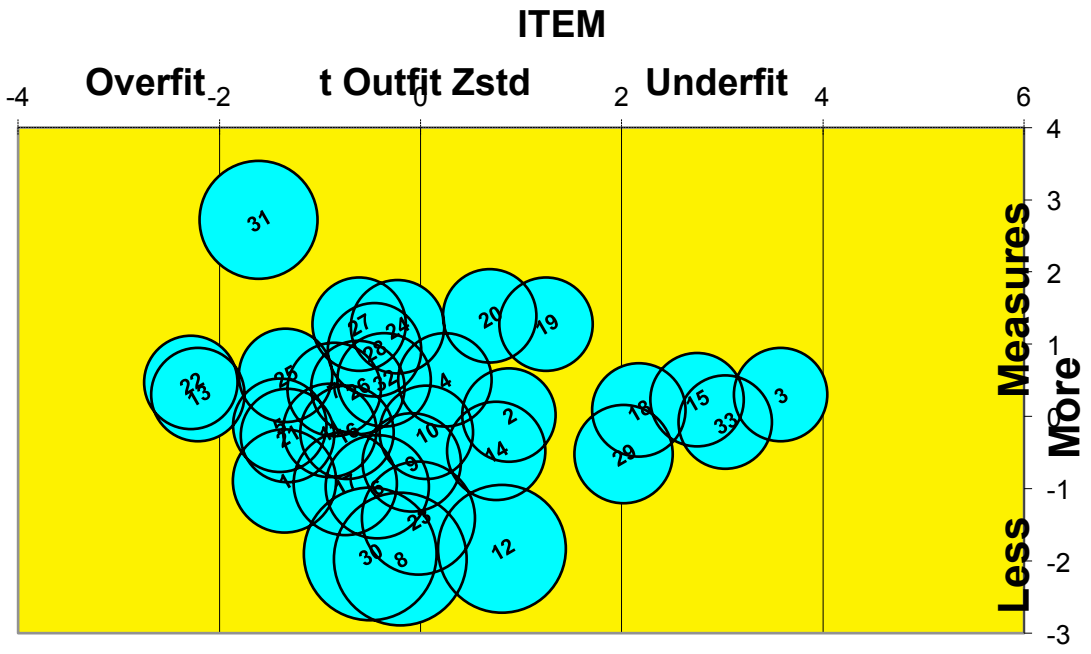
this statistic may be ignored for sample populations larger than 300 (Linacre, 2017a), it is useful when analysing the present population and alerts us to be cautious with items whose values fall outside the ± 2 range. For sample sizes between 30 and 300, values higher than +2 are too erratic to be useful for measurement and values below -2 may conversely be considered too good to be true (Bond & Fox, 2015, p.53).

Table 11. Item Statistics from Rasch Analysis

Entry	Item	Measure	Model	Infit		Outfit	
		in Logits	SE	MNSQ	Zstd	MNSQ	Zstd
1	Task1.1H	-.89	.21	.89	-1.1	.71	-1.3
2	Task1.2B	.02	.19	1.14	1.8	1.15	1.1
3	Task1.3F	.30	.19	1.29	3.5	1.50	3.6
4	Task1.4I	.51	.19	.99	- .1	1.04	.4
5	Task1.5C	-.12	.19	.89	-1.4	.80	-1.4
6	Task1.6D	-.98	.21	1.00	.1	.88	- .4
7	Task1.7A	.37	.19	.92	-1.1	.90	- .8
8	Task2.1B	-1.98	.27	.96	- .2	.87	- .2
9	Task2.2B	- .64	.20	1.10	1.1	.99	.0
10	Task2.3A	- .23	.19	.89	-1.4	1.00	.1
11	Task2.4C	- .19	.19	.90	-1.2	.86	´ -.9
12	Task2.5B	-1.84	.26	.93	- .4	1.28	.8
13	Task2.6B	.30	.19	.85	-2.1	.75	-2.2
14	Task2.7C	- .49	.20	.88	-1.4	.71	-1.7
15	Task2.8C	.23	.19	1.33	3.9	1.39	2.8
16	Task3.1C	- .23	.19	.96	-.5	.91	- .5
17	Task3.2C	- .94	.21	.85	-1.6	.82	- .7
18	Task3.3D	.09	.19	1.19	2.4	1.32	2.2
19	Task3.4A	1.29	.19	1.14	1.5	1.18	1.3
20	Task3.5A	1.40	.20	1.05	.6	1.13	.9
21	Task3.6B	- .26	.19	.95	- .6	.81	-1.2
22	Task3.7D	.48	.19	.85	-2.0	.75	-2.2
23	Task3.8B	-1.42	.23	.98	- .1	.96	.0
24	Task4.1	1.25	.19	.95	- .5	.98	- .1
25	Task4.2	.58	.19	.89	-1.5	.85	-1.3
26	Task4.3	.41	.19	1.00	.1	.95	- .3
27	Task4.4	1.29	.19	.96	- .4	.92	- .6
28	Task4.5	.93	.19	.99	- .1	.96	- .3
29	Task4.6	- .53	.20	1.06	.8	1.47	2.3
30	Task4.7	-1.91	.27	.89	- .6	.77	- .5
31	Task4.8	2.74	.24	.85	-1.0	.57	-1.6
32	Task4.9	.51	.19	1.00	.0	.95	- .4
33	Task4.10	- .08	.19	1.30	3.6	1.57	3.3
	Mean	.00	.20	.99	.0	.99	.0
	SD	1.01	.02	.13	1.6	.24	1.5

We can see at a glance those items which do not fall within the range by examining the bubble chart presented below in Figure 17. The size of the bubble represents the amount of error associated with each entry/item. Item entries 3 (1.3F), 15 (2.8C) and 33 (4.10) show the most underfit, and are therefore the most problematic. This confirms the information given by the CTT analysis. Entries 18 (3.3D) and 29 (4.6) also belong to this group of underfitting items and so should be further examined.

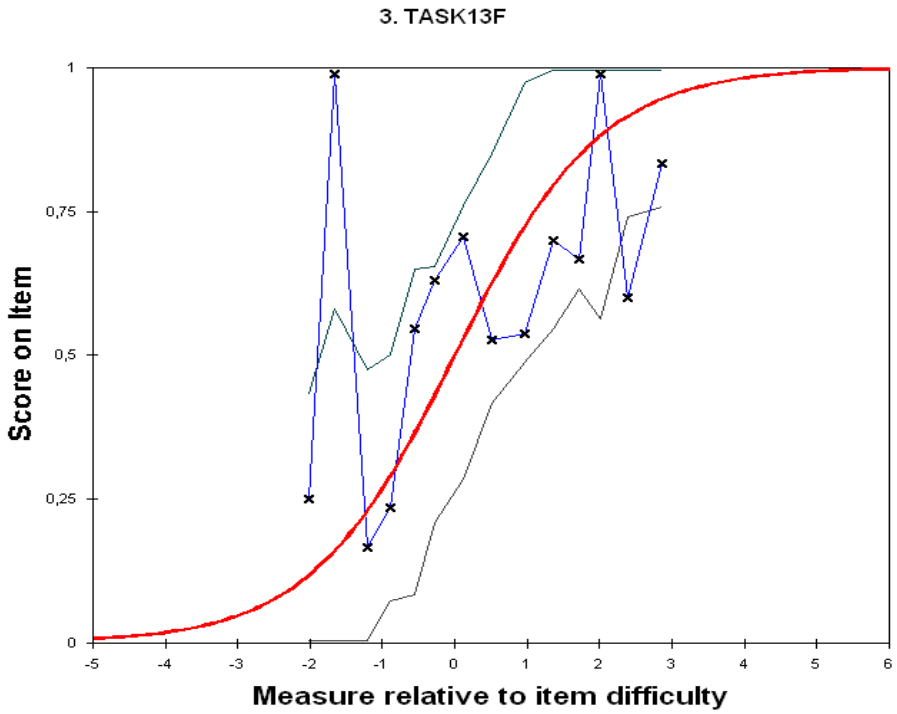
Figure 17. Bubble chart showing outfit Zstd (using item entry number)



Examining the results task by task we can then also look to the distractor analysis to get more information about item functioning. In Task 1, item 1.3F has both large outfit MNSQ and Zstd, suggesting a badly-functioning item. Although the infit MNSQ is just within the accepted parameter (using the more conservative 1.3), the Zstd (significance) is well above the recommended value of 2. Figure 18 below shows the expected and empirical ICC with 95% confidence limits for this item. If an observed data point lies outside the boundaries we may have some un-modelled variance in the observations (Linacre, 2017a). It can be seen that the actual data does not follow the expected curve

very well and a group of low ability candidates have a high probability of getting this item correct.

Figure 18. Expected and empirical ICC curve for item 1.3F



The distractor analysis for this item can be seen in Table 12. Here, the examinee measure for the correct option should be higher than the measure for any single distracter because more able examinees should choose the key, whereas less able examinees should choose the distracters. The measure statistics are accompanied by a standard error estimate. Finally, the measurement correlation is a correlation between the responses (1 for the key and 0 for the distracters) and the person measures. The key should demonstrate positive values, whereas the distracters should demonstrate either negative values or at most very low positive values (Green, 2013). Here we can see that correct answer F was chosen by most candidates (55%) but 20% of candidates chose extra distractor G; what is more, these candidates have a higher average ability than those choosing the correct answer. The item is not functioning well and should be dropped

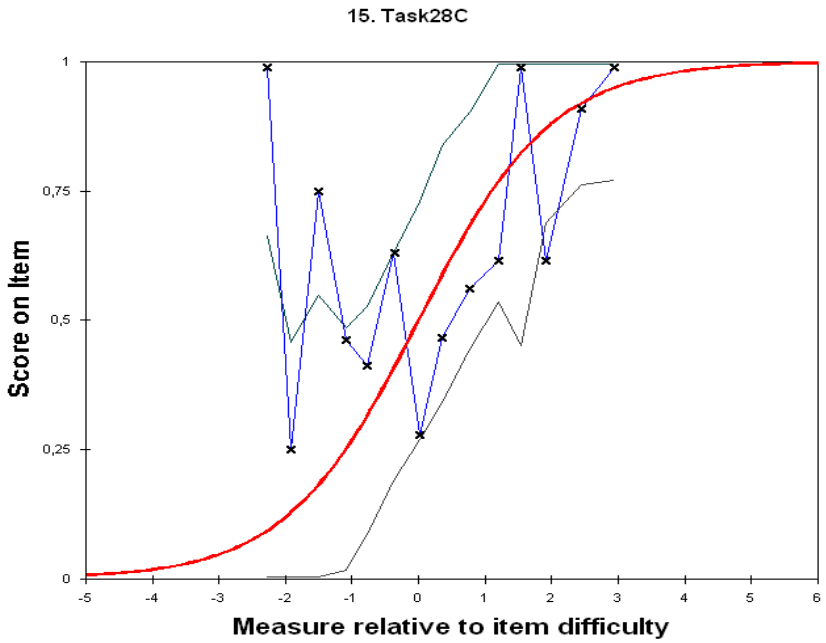
from the test. We may discover further reasons for this after an examination of the verbal reports. The rest of the items in Task 1 appear to be functioning as they should.

Table 12. Distractor analysis Item 1.3F

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE ABILITY	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	ITEM
3	X	0	6	4	-.78	.46	.5	-.22	Task13F
	D	0	1	1	-.67		.3	-.08	
	I	0	14	9	-.49	.18	.5	-.28	
	H	0	4	3	-.08	.35	.7	-.09	
	Y	0	1	1	-.07		.6	-.04	
	B	0	12	8	-.02	.31	1.0	-.15	
	G	0	31	20	.97	.20	2.9	.14	
	F	1	85	55	.91*	.13	1.3	.26	

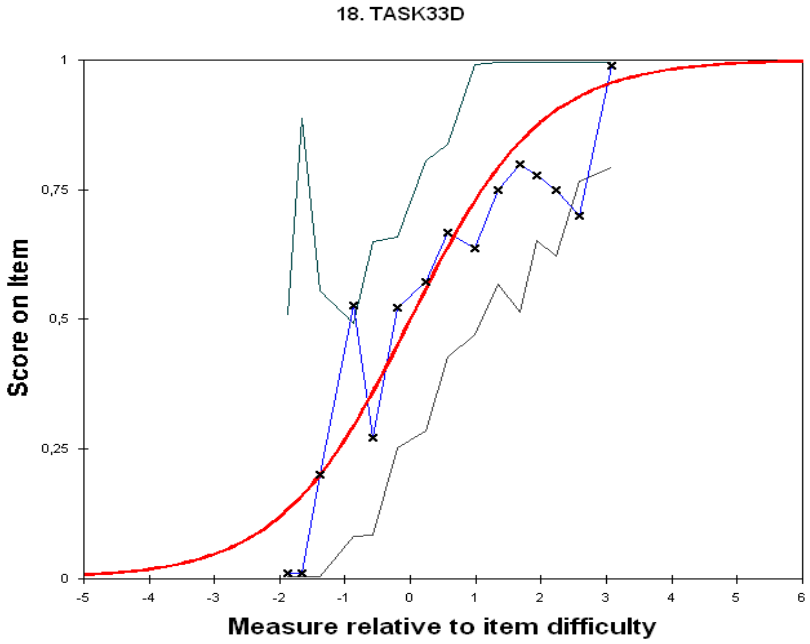
Item 2.8 from Task 2 is not functioning well either in terms of *infit* or *outfit*. A similar investigation of the distractor analysis shows no problems and it is the higher ability candidates who got the item correct. However, quite high ability candidates chose distractor B, which was therefore a strong distractor for high ability candidates. Similarly, the ICC curve in Figure 19 shows that the actual data does not fit model expectations well. This item is easy to remove from the test as it is the last item on this task and so will not cause problems in terms of spread if the soundfile is shortened and faded out.

Figure 19. Expected and empirical ICC curve for item 2.8C



All the items on Task 3 have good infit statistics, though 3.3D is problematic in terms of outfit and Zstd values. The distractors are working well and the higher ability candidates chose the correct answer. The ICC (Figure 20) shows quite a good fit to the model, and the problems could simply be that a few lower-ability candidates are guessing correctly. This item's performance can be further investigated by using the results of the verbal reports.

Figure 20. Expected and empirical ICC curve for item 3.3D

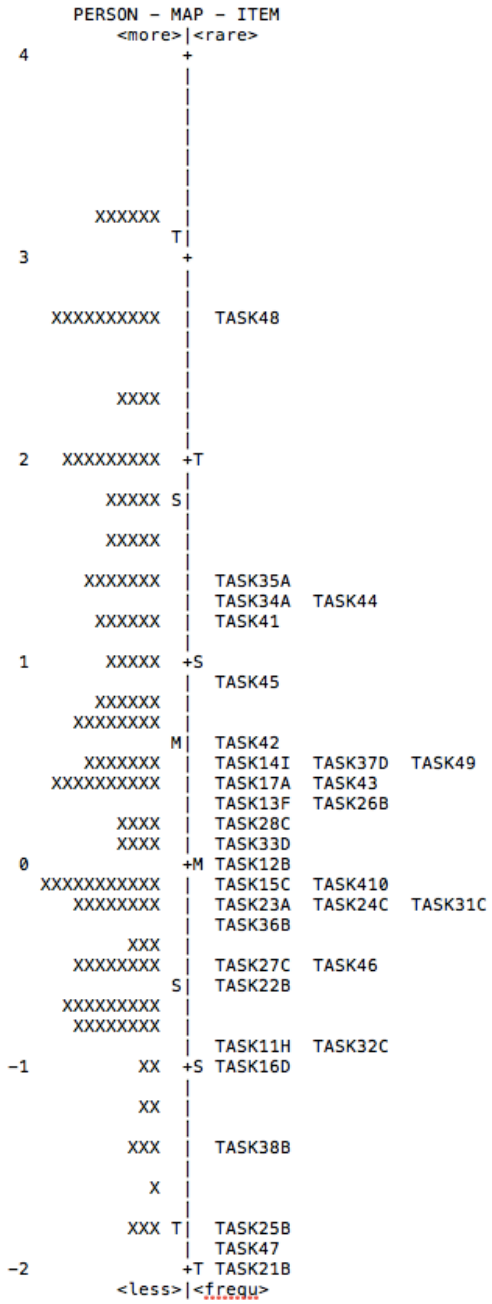


In Task 4 it can be seen that item 4.10 is problematic both in terms of *infit* and *outfit* —it is not functioning well and should be dropped from the test. Again, this is not problematic in terms of spread as the item is the last one on the task. Item 4.6 is also problematic in terms of outfit and deserves further investigation. Our initial investigation, then, shows us we should drop items 1.3, 2.8 and 4.10 as they do not fit the Rasch model and therefore do not function as expected.

The item measure data in Table 11 shows us the difficulty of each item in relation to the population in logits with its associated error (SE). Recommended SEs are no more than 0.3 (Linacre, 2017a). We can see that this is the case for all items on the test with

items 2.1B and 4.7 having the highest SEs, probably because they are very easy items for this population. As there are no candidates at this ability level, there is very little information to model, however, a better understanding of the item difficulties can be gained from studying Figure 21, which shows the variable map for this test.

Figure 21. Item/Person variable map



Here, it can been see that item difficulties range from -2 logits to 2.74 logits, with most of the items falling within the mean item difficulty plus one SD, and although the test appears to have been easy for many candidates (person ability measures do not peak at the mean), there are many candidates matched to item difficulty. The variable map can give evidence of construct under-representation if there are not enough items to measure person ability level (Bond, 2003). We can see that the most difficult item is Item 4.8 and there is a large gap between this item and the next difficult item. This could mean either that the item is too difficult for the population as a whole or that there are not enough items in this difficulty range. McNamara (1996) considers items that are more than 2 SDs from the mean (as is the case) to be potentially misfitting.

Furthermore, as the test was developed following strict CEFR-related B2 specifications, it was decided that this item did not 'fit' with the rest of the items and should therefore be dropped from the test. We can also see that there are three very easy items below the ability of all the candidates, which is not to say that all candidates got the items correct but that all the candidates have a high probability of getting the items correct. These items fall just within two SDs of the mean. It is not uncommon for tests aimed at one proficiency level to contain items both below and above the level, as it is extremely difficult to develop all items at the same proficiency level (R. Green, personal communication, April, 2011).

Our initial Rasch analysis has thus given us four items to be deleted from the test. The other items generally seem to be working well but we have highlighted items 3.3 (the item which was dropped following the CTT analysis) and 4.6 as other possibilities for deletion. It was decided to wait until after the analysis of the verbal reports before reaching the decision about which item to drop.

6.2 Research question 2 (R2): Is the test unidimensional? Do test scores include any construct- irrelevant variance?

In order to answer question two (R2) and provide evidence to support the evaluation inference, results from both a PCAR analysis and relevant results from the questionnaire will now be presented.

6.2.1 Dimensionality (PCAR)

Unidimensionality is an assumption of the Rasch model, though the analysis is carried out once data has been fit to the model in order to generate linearized residuals. In the present study, evidence needs to be provided that the test is measuring one single underlying construct—that of listening proficiency. Once the Rasch dimension has been extracted, PCAR analysis gives the residuals, the differences between model expectations and actual observations (Linacre, 2017a) and shows whether or not the standardised residuals bear a substantive structure of correlations (Aryadoust, 2013). Table 13 shows the PCAR analysis of the data; it can be seen that the Rasch dimension explains 30.5% of the observed raw variance, extremely close to the 30.6% expected by the model. This shows that the Rasch difficulty measures were successfully estimated (Linacre, 2017a). The relatively low percentage of raw variance explained by the measures is not evidence of multidimensionality but shows that person abilities and/or item difficulties have a narrow range (Linacre, 2013), as we would expect on a test measuring one proficiency level.

Table 13. Principle components analysis of Rasch residuals

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)		-- Empirical --		Modeled
Total raw variance in observations	=	41.7	100.0%	100.0%
Raw variance explained by measures	=	12.7	30.5%	30.6%
Raw variance explained by persons	=	5.9	14.2%	14.2%
Raw Variance explained by items	=	6.8	16.3%	16.3%
Raw unexplained variance (total)	=	29.0	69.5% 100.0%	69.4%
Unexplned variance in 1st contrast	=	2.0	4.8% 6.9%	
Unexplned variance in 2nd contrast	=	1.8	4.4% 6.3%	
Unexplned variance in 3rd contrast	=	1.8	4.2% 6.0%	
Unexplned variance in 4th contrast	=	1.5	3.7% 5.3%	

This analysis shows if the test is measuring any other meaningful substantive dimensions. Providing evidence that the test is unidimensional adds to the evaluation inference and supports the construct validity of the test. The analysis looks for patterns in the data which do not follow the expected Rasch measures, it shows groups of items which demonstrate the same patterns of unexpectedness (Linacre, 2017a). The analysis gives a Rasch factor and any other secondary 'contrasts', which could be indicative of another dimension being measured by the instrument. The first PCAR component explains as much of the residual variance in the data as possible. This variance has had the Rasch dimension removed. Consequently it reflects a contrast, not between the Rasch dimension and a secondary dimension, but between two secondary dimensions (Linacre, 2009). Linacre (2008) states that "we expect the first contrast to be somewhere between 1.4 and 2.0". Linacre (2017a) provides a general rule of thumb that in order for a secondary dimension to distort measurement the secondary dimension should have an eigenvalue of at least 2 items to be above the noise level. However, it might be that even if a secondary dimension is shown to be present we do not want to act on this because it may be decided that although a contrast is conceptually different it is still part of the construct under observation. After all, multidimensionality always exists to a greater or lesser extent.

It can be seen that first contrast in the residuals has an of eigenvalue of 2, explaining 4.8% of the variance in the data. This is much less than that explained by the test items (16.3%) or person ability measures (14.2%). The scree plot (Figure 22) shows this information visually, with the secondary contrasts appearing low on the graph. A more detailed analysis of the first contrast can be seen in Figure 23, which shows the loading patterns on this contrast. The horizontal line represents a zero loading and items (represented capital by letters) above the line have loadings above zero, those below the line (represented by lower case letters) have loadings below zero. It can be seen that the items do not form distinguishable clusters, but are distributed in different regions of the map, supporting the unidimensionality assumption (Linacre, 2017a). If no substantive pattern appears in item residuals, items scatter in different regions of the map without

clustering in either its positive or negative loading regions (Aryadoust, Goh, & Kim, 2011).

Figure 22. Standardised residual variance scree plot

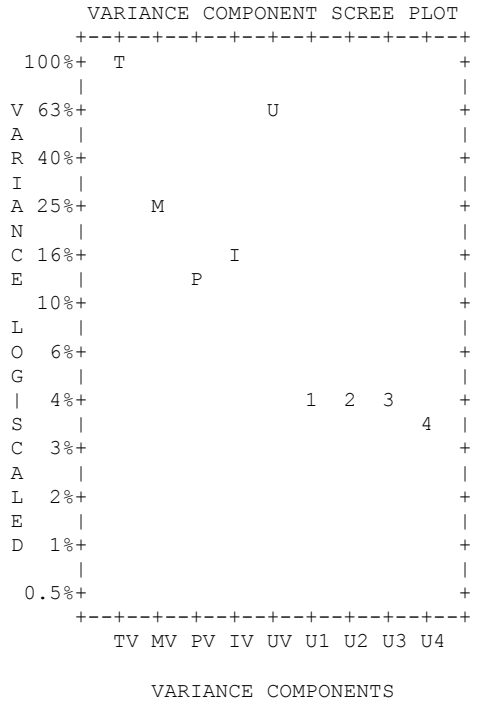
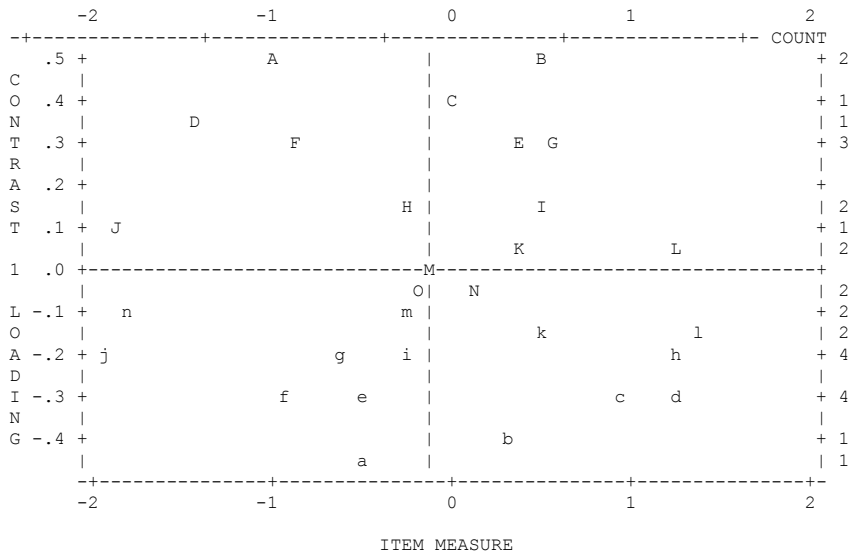


Figure 23. Standardized residual plot for first contrast



The patterns of the loadings are important and the contrast plot should be examined to see if the contrast can be explained. Here it can be seen that, although the eigenvalue on the first contrast is just on the limit (2), there are 3 items (represented by A, B and to a lesser extent C) which lie horizontally at the top of the contrast plot. The loading coefficients for these items can be seen in Table 14. These three items were all from Task 1, which is a gist task in contrast to the other items which are from main ideas tasks. I would therefore argue that the difference is part of the natural variation in language items rather than evidence of a different dimension. Indeed, completely unidimensional data would mean that all items are identical (Reckase, 2009), which would be of no value. It is felt then that the contrast can be explained and the items do not constitute a substantive secondary dimension—that is to say, that the test appears to be a unidimensional test of listening ability. Also, a secondary dimension of only two items out of 33 is not considered to be large. The magnitude of further contrasts did not reach two eigenvalues, which indicates that they may represent statistical patterns without substance (Linacre, 2017a).

Table 14. Standardised residual loadings for first contrast

CON- TRAST	LOADING	INFIT MEASURE	OUTFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	ITEM	LOADING	INFIT MEASURE	OUTFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	ITEM
1	.49	-.97	1.00	.88	A 6	TASK16D	-.43	-.48	1.10	1.13	a 14	TASK27C
1	.48	.51	.98	1.02	B 4	TASK14I	-.38	.30	.84	.75	b 13	TASK26B
1	.40	.02	1.12	1.12	C 2	TASK12B	-.32	.92	.98	.94	c 28	TASK45
1	.34	-1.41	.99	.96	D 23	TASK38B	-.30	1.28	1.13	1.17	d 19	TASK34A
1	.31	.37	.91	.90	E 7	TASK17A	-.29	-.52	1.05	1.41	e 29	TASK46
1	.30	-.89	.88	.71	F 1	TASK11H	-.28	-.93	.85	.82	f 17	TASK32C
1	.29	.57	.88	.85	G 25	TASK42	-.22	-.64	1.08	.97	g 9	TASK22B
1	.15	-.22	.88	1.00	H 10	TASK23A	-.21	1.28	.95	.91	h 27	TASK44
1	.13	.51	.99	.95	I 32	TASK49	-.20	-.26	.94	.80	i 21	TASK36B
1	.11	-1.90	.89	.77	J 30	TASK47	-.18	-1.97	.95	.87	j 8	TASK21B
1	.06	.40	.99	.92	K 26	TASK43	-.16	.47	.84	.75	k 22	TASK37D
1	.03	1.24	.95	.96	L 24	TASK41	-.15	1.39	1.04	1.09	l 20	TASK35A
							-.12	-.22	.95	.88	m 16	TASK31C
							-.11	-1.83	.94	1.28	n 12	TASK25B
							-.03	-.19	.90	.86	O 11	TASK24C
							-.03	.09	1.18	1.31	N 18	TASK33D
							-.02	-.12	.89	.80	M 5	TASK15C

The Rasch model assumes local independence, which means that the probability of answering an item correctly should be independent of the answer to other items. The presence of another dimension may be indicative of a violation of this assumption. Great care was therefore taken to ensure local independence during the item development stage and a correlation matrix of the Rasch model linearized residuals (Table 15) between all

items on the test shows that none of the items have a correlation coefficient greater than 0.5. This may indeed be considered sufficient evidence of local independence (Linacre 2017a).

Table 15. Largest standardized Rasch residuals correlation coefficients

LARGEST STANDARDIZED RESIDUAL CORRELATIONS

USED TO IDENTIFY DEPENDENT ITEM

CORRELATION	ENTRY NUMBER	ITEM	ENTRY NUMBER	ITEM
.31	1	TASK11H	7	TASK17A
.28	28	TASK45	29	TASK46
.26	7	TASK17A	10	TASK23A
.22	11	TASK24C	13	TASK26B
-.25	9	TASK22B	25	TASK42
-.23	2	TASK12B	16	TASK31C
-.22	14	TASK27C	32	TASK49
-.21	6	TASK16D	13	TASK26B
-.21	14	TASK27C	23	TASK38B
-.21	4	TASK14I	21	TASK36B

The fact that the test is shown to be unidimensional is evidence towards construct validity and suggests that no other construct apart from listening ability is being tested. Such evidence can therefore be considered confirmation that construct irrelevant variance is not present.

6.2.2 Questionnaire results for construct irrelevant variance

Further evidence concerning construct-irrelevant variance was collected from the test takers themselves. One of the themes on the questionnaire concerned candidate opinions about test administration and content. Results are shown in Table 16, where a higher number represents higher agreement.

Table 16. Questionnaire results for construct irrelevant variance

Question	Mode	Mean	Standard Deviation (SD)
How authentic did you find the audios used in the tasks?			
1.1 Questions about sport	4	3.36	.888
1.2 Moving to the USA	3	3.23	.646
1.3 Text messaging	3	3.14	.758
1.4 Geography trip	3	3.01	.814
How difficult did you find the listening audios?			
2.1 Questions about sport	3	3.10	.841
2.2 Moving to the USA	2	2.09	.747
2.3 Text messaging	3	2.62	.770
2.4 Geography trip	3	2.96	.760
How difficult did you find the questions?			
3.1 Questions about sport	3	2.57	.960
3.2 Moving to the USA	2	1.89	.705
3.3 Text messaging	2	2.35	.739
3.4 Geography trip	3	2.75	.893
How familiar did you find the topics used in the tasks?			
3.1 Questions about sport	3	2.73	.898
3.2 Moving to the USA	3	2.80	.820
3.3 Text messaging	3	3.00	.866
3.4 Geography trip	2	2.41	.815
How suitable did you find the amount of time to:			
5.1 Read the questions	2	2.14	.694
5.2 Answer the questions	3	2.66	.619
The instructions for the task were clear.			
6.1 Questions about sport	4	3.65	.655
6.2 Moving to the USA	4	3.81	.443
6.3 Text messaging	4	3.78	.502
6.4 Geography trip	4	3.68	.636
The quality of the recording was good.			
7.1 Questions about sport	3	2.75	.928
7.2 Moving to the USA	4	3.55	.608
7.3 Text messaging	4	3.46	.642
7.4 Geography trip	4	3.42	1.025
I recognised the accent of the speaker(s).			
8.1 Questions about sport	2	2.33	.985
8.2 Moving to the USA	3	2.99	.934
8.3 Text messaging	3	2.75	.926
8.4 Geography trip	3	2.64	.942
The speaker spoke at normal speed.			
9.1 Questions about sport	3	2.46	1.031
9.2 Moving to the USA	4	3.28	.752
9.3 Text messaging	3	3.03	.774
9.4 Geography trip	3	3.16	.713

Firstly, results about actual test administration which could cause construct irrelevant variance relate to aspects such as task instructions, quality of the audio and amount of time allowed. Here, students overwhelmingly reported that the instructions for all four tasks were clear, each with a mode of 4 (completely agree). Similar results were reported about the quality of the soundfiles used, with slightly less agreement about the soundfile for Task 1 (though they still agreed that the soundfile was good quality). However, with respect to the amount of time, students reported that while there was enough time to answer the questions, there was not however quite enough to read the questions.

Further investigation comparing mean score with answers to this question showed that the candidates who reported that there was not enough time to read the questions showed a lower ability (mean score = 18.97) when compared to those who believed there was enough time (mean score = 21.52). This aspect could potentially be further investigated by timing a known group of B2 level students.

In terms of test content, the students believed that they were listening to authentic content; this finding is important as authentic audio is a key part of the test construct. Regarding difficulty, students reported that they found the audios quite difficult to understand, with the exception of Task 2 which they found to be not very difficult. A Spearman's rho correlation was run to determine the relationship between opinions about the difficulty of the audio and the total score on each of the four tasks. There were weak to moderate negative correlations between opinions and scores, which were statistically significant: Task 1 r_s = -.486, p = .000); Task 2 r_s = -.453, p = .000); Task 3 r_s = -.243, p = .002); Task 4 r_s = -.280, p = .000). Although it seems obvious that candidates would report more difficulty for tasks on which they received lower scores, previous studies have found that perceptions of difficulty do not follow the psychometric properties of the test. Elder, Iwashita and McNamara (2002), for example, concluded that task difficulty cannot be accurately estimated based on candidate perceptions. In this regard, the present study showed that candidates were able to distinguish task difficulty fairly well based on their perceptions of how difficult they found the audios. The candidates found the questions for Task 2 and 3 (both MCQ tasks) easier to answer than those for Task 1 and

4. Here, it can be seen that candidates believed Task 4 to be the most difficult. However, few candidates reported that the questions were very difficult and again responses were generally related to test score, i.e., the lower scoring candidates reported that the questions were very difficult (see Table 17), though these results did not all have significant correlations. It should be highlighted here that task difficulty is determined through both the items which need to be answered and the soundfile which needs to be processed, i.e., the demands of the task (Field, 2008a).

Table 17. Comparison of mean score and opinions about task difficulty (questions)

	N	**Mean score on test**
Task 1:		
1. Not difficult	22	23.77
2. Not very difficult	50	21.92
3. Quite difficult	51	18.45
4. Very difficult	29	15.76
Task 2:		
1. Not difficult	45	21.87
2. Not very difficult	81	20.04
3. Quite difficult	24	16.08
4. Very difficult	2	12
Task 3:		
1. Not difficult	15	22.4
2. Not very difficult	78	20.6
3. Quite difficult	50	18.72
4. Very difficult	9	15.33
Task 4:		
1. Not difficult	12	20.5
2. Not very difficult	48	20.08
3. Quite difficult	58	21.16
4. Very difficult	34	17.06

Topic familiarity was considered to be an important question, as the exam aims to relate to the TLU domain of school leaving and it would therefore be expected that the students had studied the topics as part of their upper secondary English classes. Here, students reported the topics for Tasks 1, 2 and 3 to be quite familiar. However, Task 4 was found to be less familiar and could be indicative of a certain lack in secondary school textbooks. After all, an important CEFR B2 descriptor is "can understand announcements and messages on concrete and abstract topics spoken in standard dialect at normal speed"

(CEFR, p.67). However, it could also be that the students' definition of 'familiar' meant that they were not familiar with the very specific content of the soundfile—something to be expected.

Table 18. Questionnaire results for familiarity with speakers accent

	N	**Mean score on test**
1. I completely disagree	36	18.50
2. I disagree	49	18.37
3. I quite agree	46	20.11
4. I completely agree	20	25.40

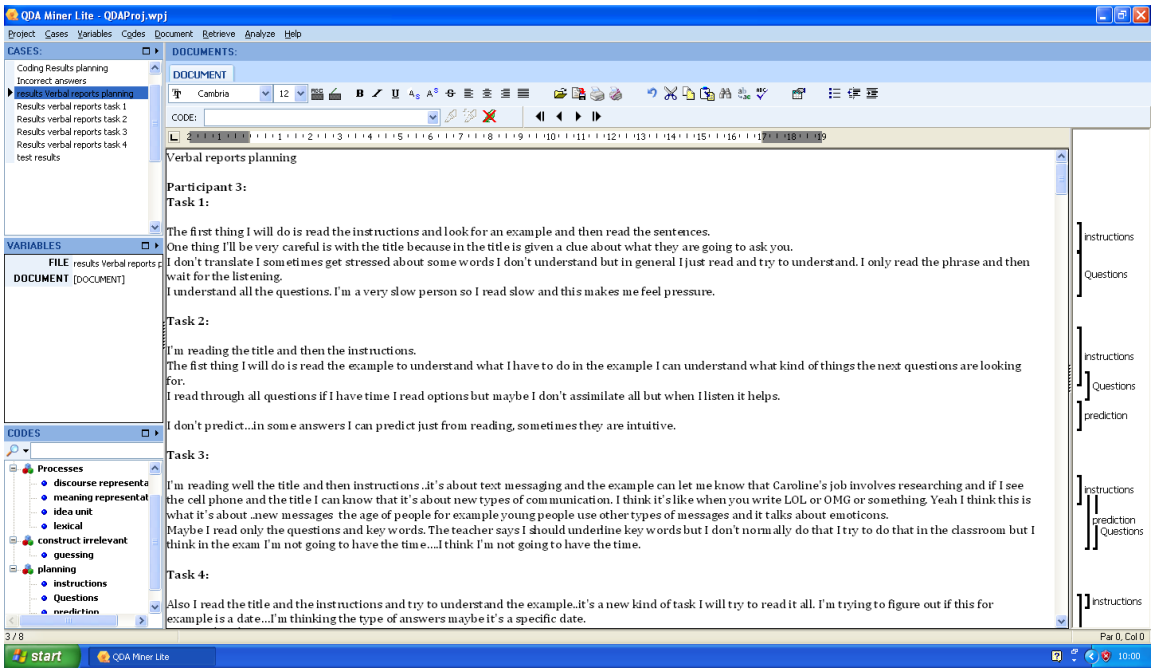
The questions about familiarity with the accent and the speed of delivery of the speakers showed that in general the students believed the speakers spoke at normal speed and they were familiar with their accents, though less so for the speaker on Task 1. Here, responses seem to be related to ability (see Table 18), with the higher scoring candidates stating that they were familiar with the speaker's accent.

In summary, it has been seen that in general students do not believe the test to contain construct-irrelevant variance and most negative perceptions were held by lower ability candidates, possibly in response to the fact that they did not feel that they had performed well on the test.

6.3 R3: Do test takers use the relevant knowledge, skills and abilities to solve test items on the BFE listening test?

The category of planning and goal-setting was analysed qualitatively from the pre-listening concurrent 'think alouds'. The concurrent reports were entered into *QDA Minor Lite* and coded according to the emerging themes which were seen to be reported by the participants—as can be seen in Figure 24. The same process was followed with the retrospective reports and these results will first be presented quantitatively as the level of processing reached by each participant in order to answer the test items.

Figure 24. Screenshot of coding process in *QDA Minor Lite*



The reliability of the coding scheme is evidenced by the following intra-rater agreement results. Exact agreement between both coding sessions was found to be 87% and the intra-coder reliability reported by Cohen's Kappa was 0.782 ($p < .0.001$), 95% CI (0.65, 0.91). According to Landis and Koch (1977) this represents substantial agreement, which can probably be explained by the reduced number of categories which were well defined. Salient points and any meta-cognitive strategy use will be exemplified by presenting example extracts from the reports.

6.3.1 Concurrent verbal reports

Buck (2001, p.104) calls the pre-listening or planning stage "assessing the situation". Here candidates are provided with a context statement and items provide further information about the context as well as a purpose for listening. According to Shohamy and Inbar (1991), this is important as it allows them to activate schemata and generate hypotheses. All participants used planning and prediction strategies, which fell into the following categories:

1. Use of task title and picture to activate relevant schemata.
2. Use of 'key words' in items to be sure of purpose for listening and to activate schemata.
3. Prediction using previous knowledge schemata.

In general, most participants were also seen to read through the task instructions carefully in order to be sure of what was required of them.

Example:

> *First of all I try to understand the task I read the instructions and several questions even the picture of the map catches my attention.*
>
> (**Participant 6, Task 2)**

Here, the example item was shown to be useful as in the following extracts.

Examples:

> *I'm reading the title and then the instructions.*
> *The first thing I will do is read the example to understand what I have to do in the example I can understand what kind of things the next questions are looking for. I read through all questions and if I have time I read options but maybe I don't assimilate all but when I listen it helps.*
>
> (**Participant 3, Task 2)**

> *I think that the example …especially the information…it's good information because if you see he is a doctor or a student, you see him in a certain way…*

he's not a worker so it's key information and you have stereotypes and clues about the situation.

(**Participant 6, Task 2)**

The metacognitive strategy of planning and prediction is an important part of listening comprehension, with previous studies (Vandergrift, 1997; Goh, 2000) reporting that higher ability listeners are more likely to use metacognitive strategies. Here, participants were indeed seen to activate their previous knowledge schemata and to use the context of the situation to help them predict the audio.

Examples:

I'm reading well the title and then instructions …it's about text messaging and the example can let me know that Caroline's job involves researching and if I see the cell phone and the title I can know that it's about new types of communication. I think it's like when you write LOL or OMG or something.

Yeah, I think this is what it's about …new messages, the age of people, for example young people use other types of messages and it talks about emoticons.

Maybe I read only the questions and key words. The teacher says I should underline key words but I don't normally do that, I try to do that in the classroom but I think in the exam I'm not going to have the time.

(**Participant 3, Task 3)**

The majority of the vocabulary is familiar for me. Yes, now I understand deeply the content of the task and now I'm thinking about a Mexican travelling to USA, erm, I'm trying to be empathy with people travelling from Mexico, a less developed country, to the USA, a more developed country. He'll have a cultural shock and I feel that I can do this task better, I am ready.

(**Participant 6, Task 2)**

First I look at questions so I can predict the meaning of the listening—I believe that he's talking about cities and airports, Charlottesville and probably there's a man talking about his travel and the reason for his travel…he's called Jean.

I think the questions aren't so difficult to understand, it will depend on the speed of the listening and also the accent is important for me, I've always had a teacher from the USA and are more accustomed…here, I think it could be Mexican.

(**Participant 7, Task 2)**

I'm gonna hear something about how the research started and problems at the beginning and about the research and what they found.

(**Participant 7, Task 3)**

I'm trying to predict too…it will make sense that free users talk to each other on Facebook. Here she's, she'll talk about how to get information, where it's collected from, what age of participate people will be and what the results are and why people use emoticons and for what they use it. Ok
I have an idea of what she is going to talk and then I'll listen to her and know about it better.

(**Participant 9, Task 2)**

Ok there is a person who is going to move from one city to another…he's going to talk about why he's going to move and what he's going to do.
(Reading example) He's moving to the USA for his doctorate studies so maybe he's going to work.

He wants to find somewhere to live.
Ok it's just like the daily routine or something because it says that after the airport.

Ah he's not going to visit anyone no he's just finished his studies so he's going to work somewhere and he compares the cities and it talks about the advantages and disadvantages and I think he's going to live in the USA because of the last question…it's like a story.

(**Participant 5, Task 2)**

It can be seen that some participants made very strong predictions and built a skeleton story of the audio file from the items, especially for the MCQ type tasks, and here it is clear that topic familiarity has a role to play, as stated by the following participant.

Example:

> *Another question I am thinking that this is the typical topic in the English class related to new technology. I can expect to find this topic in an English exam or course.*
>
> *Obviously it's better if the topic is familiar because you know the vocabulary and you can guess the correct alternative easily if you follow your common sense or knowledge.*
>
> (**Participant 6, Task 3)**

Here, it was also found that participants certainly relate events to their own personal previous knowledge schemata.

Example:

> *So he's talking first about the flat and then when he arrives what he's going to do then what was hard at first in USA. What he had to learn, the differences between Mexico and USA why he felt accepted…I'm thinking about key points.*
>
> *OK I understand.*
>
> *Cos I've been living in USA and I understand the situation.*
>
> *In this one he finds it difficult… for me it will be name or accent, cos that's what happened to me.*
>
> *And he wants Americans to know..for me it would be to know where he's from… people see Mexican people like they are from a village and they don't have culture and internet and things like that.*
>
> (**Participant 9, Task 2)**

Many of the participants made reference to the time element of an exam situation; they feel like they do not have enough time to prepare for a task and this often leads to feelings of anxiety. Again, these comments resonated with the questionnaire results, where many candidates felt that they did not have enough time to read the questions.

Examples:

> *First I'm reading the instructions because I'm used to Cambridge and this might be different. I'm going to underline the key words. I normally underline nouns, especially places.*
>
> *There isn't time to predict, sometimes I don't have time to read the last questions and you are reading and listening at the same time there isn't time to do everything.*
>
> **(Participant 4, Task 3)**

> *…in the exam when I read 45 seconds I feel stressed and I haven't got time I would prefer more time.*
>
> **(Participant 6, Task 3)**

> *I try to read the questions but the first thing I think is that I only have 45 seconds and I need to read quickly, it would be better if I had more time. In this one I have to write the word I don't have to interpret it just write the word.*
>
> **(Participant 4, Task 4)**

Task type was also shown to lead to anxiety for some of the participants, with the less familiar task types being the most problematic.

Examples:

> *Immediately, I recognise that this task is easier than the last one—I only have to understand one of 4 possibilities and I feel more confident with this.*
>
> *The last task gave me anxiety because in the beginning I didn't understand, but here I recognise the task… I feel better.*
>
> **(Participant 6, Task 2)**

These comments would also seem to concur with the questionnaire results, where candidates were least satisfied with Tasks 1 and 4 as a fair measure of their ability (see section 6.5). Any new test then should be well exemplified and students should be familiar with all types of tasks which will appear on the test. Indeed, this is something which is recognised by professional testing bodies: for example, the EALTA codes (2006) asks the question, 'Are test methods/tasks described and exemplified?

In terms of construct-irrelevant variance, some of the participants used guessing strategies before listening or tried to discard unlikely options. This shows the importance of the item writing process; all distractors should be plausible and no answer should be guessable without hearing the audio.

Examples:

> *I look for ridiculous options.*
> *I think it's not easy to predict. I don't find any wild answers, they are all possible.*
>
> **(Participant 8, Task 2)**

> *Here I can guess what it could be in each… for example in the first it could be by cash or by credit card and in the second it could be allergies for example or something like this.*
>
> **(Participant 8, Task 4)**

> *Erm, I think the 9 is to have to compare with another person so maybe with your partner.*
>
> **(Participant 3, Task 4)**

> *Age—I think I can discard D because I think the age is not an appropriate factor …I read something about it from the UN.*
>
> **(Participant 7, Task 3)**

The above examples give evidence of well-written items; no implausible answers were found, wild guessing is incorrect and the last example shows that the candidate discarded the correct answer before listening by applying previous knowledge.

6.3.2 Retrospective verbal reports

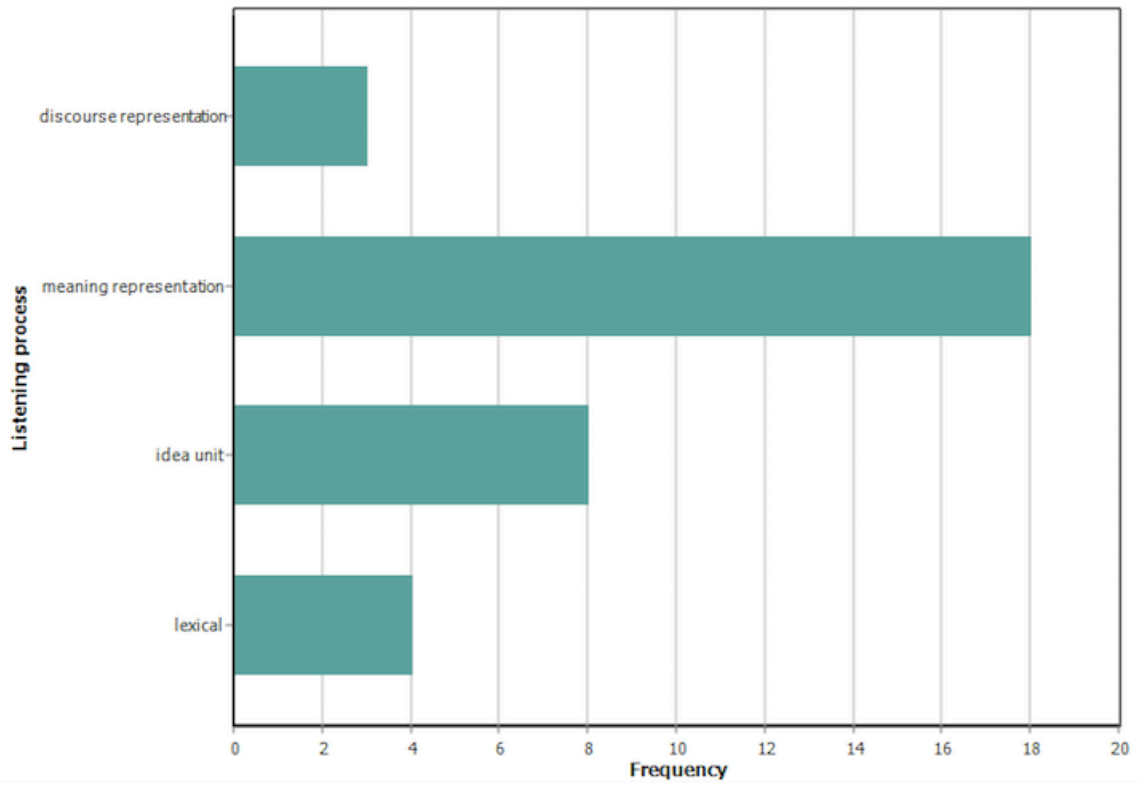
First I will present the level of processing reached for all correct responses to the test items for each of the four tasks, followed by a more in depth qualitative analysis. Table 19 shows correct and incorrect responses to each item on the test for the seven participants who took part in the study.

Table 19. Correct responses to items by verbal report participants

Participant	3	4	5	6	7	8	9	Correct answers
Task 1								
1	✓	✗	✓	✗	✓	✓	✓	5
2	✗	✗	✗	✓	✓	✓	✗	3
3	✓	✗	✓	✗	✓	✓	✗	4
4	✗	✓	✗	✓	✓	✓	✓	5
5	✓	✗	✗	✓	✓	✓	✓	5
6	✓	✓	✓	✓	✓	✓	✓	7
7	✓	✗	✗	✗	✓	✓	✓	4
Task 2								
1	✓	✓	✓	✗	✓	✓	✓	6
2	✓	✓	✓	✓	✓	✓	✓	7
3	✓	✓	✗	✓	✓	✓	✓	6
4	✓	✓	✓	✗	✓	✗	✓	5
5	✓	✓	✓	✓	✓	✓	✓	7
6	✗	✓	✗	✗	✓	✓	✓	4
7	✓	✓	✓	✓	✓	✓	✓	7
8	✓	✓	✗	✗	✓	✓	✓	5
Task 3								
1	✓	✓	✓	✓	✓	✓	✓	7
2	✓	✓	✓	✓	✓	✓	✓	7
3	✗	✓	✓	✓	✓	✓	✓	6
4	✗	✓	✗	✗	✗	✓	✓	3
5	✓	✓	✗	✗	✗	✗	✗	2
6	✓	✓	✗	✗	✓	✓	✓	5
7	✓	✓	✗	✓	✓	✓	✓	6
8	✓	✓	✗	✓	✓	✓	✓	6
Task 4								
1	✓	✓	✓	✗	✗	✗	✓	4
2	✓	✓	✓	✓	✓	✓	✓	7
3	✗	✓	✓	✗	✓	✗	✗	3
4	✓	✗	✗	✗	✓	✗	✗	2
5	✓	✓	✓	✗	✗	✗	✗	3
6	✓	✓	✓	✓	✓	✓	✓	7
7	✗	✓	✓	✓	✓	✓	✓	6
8	✓	✓	✗	✓	✗	✓	✓	5
9	✓	✓	✓	✓	✓	✗	✗	5
10	✓	✓	✓	✗	✗	✓	✓	5
Total	27	27	19	18	27	26	26	

Task 1 is a gist task. Scores on this task ranged from 3 to 7 out of a possible score of 7. Figure 25 below shows the highest process in the listening ability model which was reached in order to answer the items correctly.

Figure 25. Highest level of processing used to answer Task 1



It can be seen that most items were answered from a meaning representation of the soundfile (18), and only three instances of a full discourse representation were reached. Here, it was the three higher scoring candidates who showed full understanding. A total of 12 items were answered correctly by participants only understanding isolated vocabulary or idea units. Here the use of the metacognitive strategy of 'inference' was highly evident and participants were able to use the cognitive environment along with pragmatic knowledge to answer items when the audio was only partially understood.

Example:

> *I heard 'I completely agree' and so I think it's this because you can't disagree about women doing sport as it is not accepted. And I heard about people fighting against each other, this was another key piece of information for me…people fighting. This is a boxing competition.*
>
> **(Participant 6, Q1.5)**

Here the participant has only understood a couple of idea units, but by using pragmatic knowledge and inference was able to choose the correct answer.

Example:

> *He was talking about advancing so I said it was the drugs. So here I think its clear that it's the drugs one cos he says that it's a difficult question, and that's true, and then he says it's hard to deny that in every sport there is some kind of advance that they take… and at the end he says something about to improve their performance so I think it's this one.*
>
> **(Participant 9, Q1.7)**

Here the participant has understood a number of idea units and by using inference has managed to build a meaning representation of the audio. Similarly, inference was very much at play for those items which were answered correctly from solely understanding isolated vocabulary.

Example:

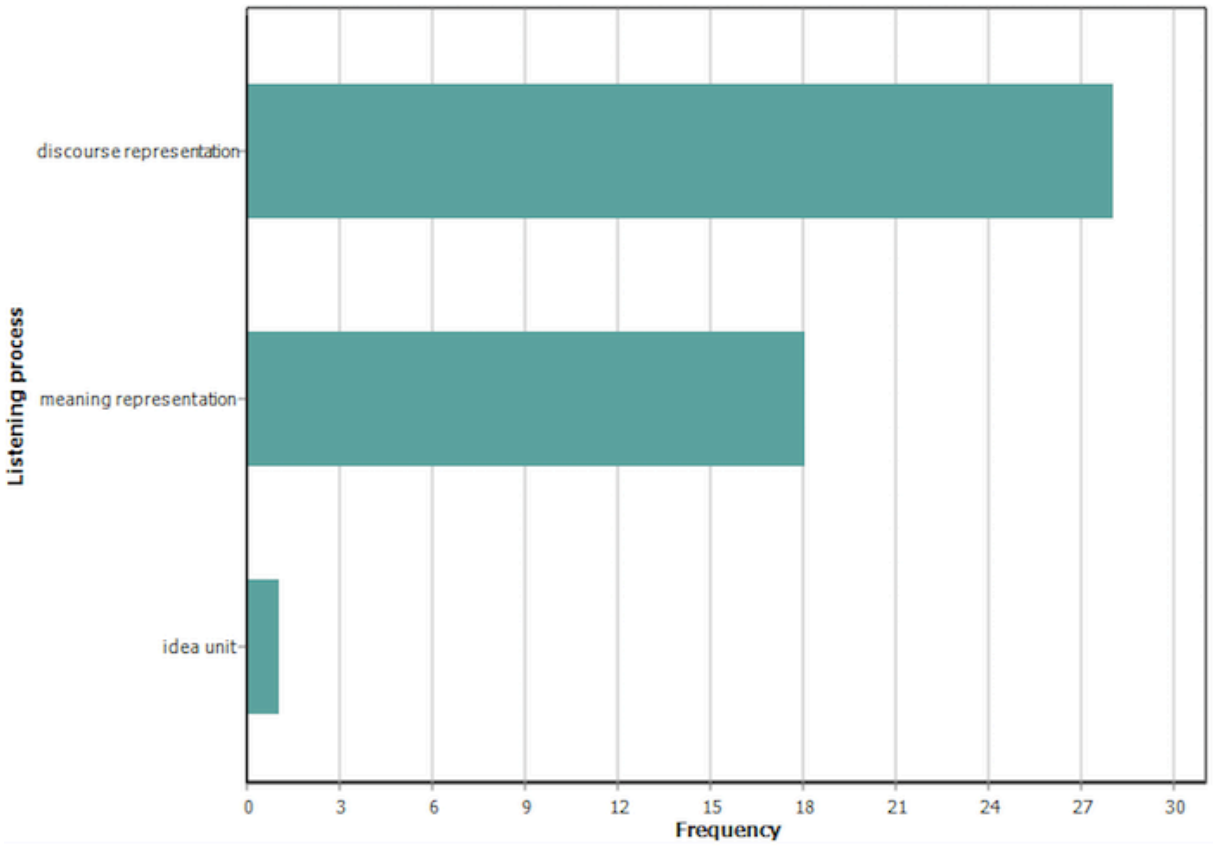
> *H…it's talking about the players have more money doing something… so I thought about publicity…I don't know. Yeah …he said 'soda' …He was talking about best players and also about soda so I imagine publicity.*
>
> **(Participant 5, Q1.1)**

It would seem then that Task 1 follows the intention of the test developer, as for a gist-type task candidates are not expected to fully understand the audios. The task has been designed to elicit the metacognitive strategies of inference and monitoring along

with the application of prior world knowledge (pragmatic, discourse and cultural) as presented in the BFE listening ability model.

Task 2 is a MISD/IPM task and the task type is MCQ; as such, we would expect the candidates to understand and follow a larger proportion of the audio. Figure 26 shows this to be the case, where scores were between 4 and 7 out of a possible 8.

Figure 26. Highest level of processing used to answer Task 2



It can be seen that no items were answered by simply understanding isolated vocabulary. One item was answered by only understanding an idea unit and this participant answered by discarding distractors and relying on the correctly understood idea unit.

Example:

> *yeah the last two questions was more difficult for me... I'm not sure. His country's history… He didn't say anything about this… He said something like he show his passport. I didn't get all the words so this was a kind of choice for discarding all the other answers.*
>
> **(Participant 7, Q2.8)**

In fact, the discarding of distractors was a common strategy on this task, which shows that the metacognitive strategy of monitoring was taking place.

Examples:

> *He has to learn about how to get around because he says you have to learn how to get on the bus and other people have to understand what you need or what you want…the public transport was one of my clues. I listened for the education system because it was possible… but he only said about the bus.*
>
> **(Participant 3, Q2.4)**

> *He talks about groceries and the shops to buy them and that the water can't be drunk in Mexico but in the USA yes. I understood better. I try to dismiss distractors, sometimes they give false clues for example here they talk about food and water.*
>
> **(Participant 4, Q2.3)**

> *Yeah, that was quite difficult for me… I don't remember what was 'work out' the translation… but he said that the American people, they try to guess where he's come from and the people say France, Poland. I'm 60% sure that it is 'work out his accent'. 'Try to get to know him' I think no. I discard, 'understand his accent', no because he can communicate they can understand fluently…yeah he said something about his name but this was more about the physical aspect, he don't look like the typical Mexican guy, it's not about say his name… that wasn't the meaning.*
>
> **(Participant 7, Q2.7)**

As can be seen most of the items on this task were answered correctly by reaching a discourse representation of the audio. In many cases, this involved inference relying on previous knowledge and even using clues from the speakers intonation and underlying intentions, therefore following the intention of the test developer.

Examples:

> *To find where to live, he just went looking on internet… he saw an announcement so he only has to write to the house of the person who was renting the room and he said 'that's it'… the way he said it… it was easy.*
>
> (**Participant 3, Q2.1)**

> *He wants people to know he is from Mexico because Americans think it is a low country, like they don't know they are like them they think they are under them and he wants to show them a new experience… a new chance. He's explaining how people get shocked when he says he's from Mexico because they have a stereotype about how they look like and how they are behind them and things like that. That's why he says that it's important to him for people to know that he's from Mexico.*
>
> (**Participant 9, Q2.9)**

Those answers which were correct from building meaning representations from the audio also relied on prior knowledge and inference.

Example:

> *He had to learn about… he didn't know how to get around, yeah that was immediately after this question he said that he had to learn about the neighbourhood or something and also for me that was familiar… the first thing you have to do when you go to another country is that you have to learn about how to get around… by foot or… I don't know. The first time I looked at other options but second time I was totally sure that was the answer.*
>
> (**Participant 7, Q2.4)**

A couple of the participants also made reference to the speaker and his accent, which they considered to be familiar and therefore helpful.

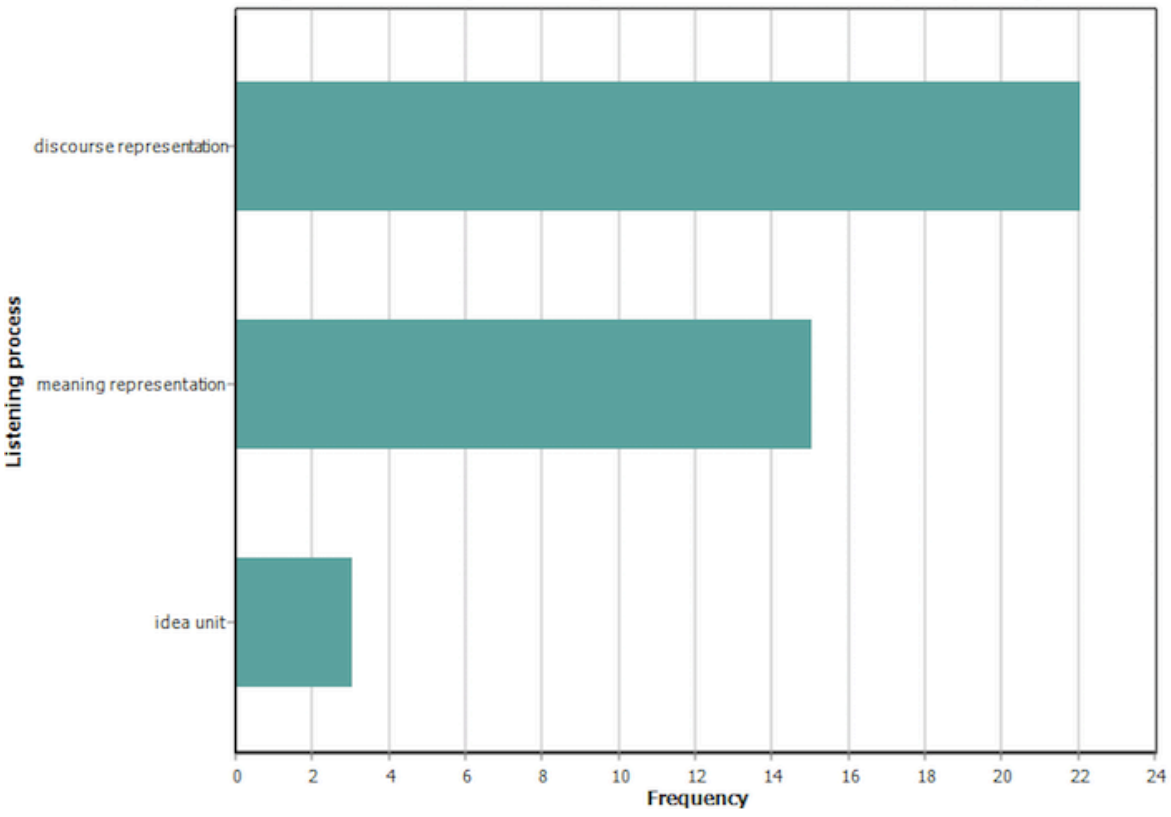
Examples:

> *I think it was easier. Yeah I understand the context cos I'm from Mexico. I understand the problems about that you can't drink water from the tap. and many people think Mexicans look erm have dark type skin, they have stereotypes. I can relate to the story. I recognise his accent. Mostly I understand because I can relate to the context, I have some experience in most of the things he was talking about. Also accent… the accent, I'm accustomed to this accent.*
>
> **(Participant 3)**

> *Well I listened to him like he was talking to me… it was really easy because he speaks like us, like Spanish people. He vocalised better. And well closer to me… it was comfortable to listen to.*
>
> **(Participant 9)**

Similar results were found for Task 3 which is also a MISD/IPM task but includes a dialogue rather than a monologue. Figure 27 shows the level of processing reached in order to answer items correctly for this task.

Figure 27. Highest level of processing used to answer Task 3



Again most of the items were solved by a mix of both bottom up and top down processing. Accurate decoding of idea units were expanded upon by using contextual clues to expand mental models and build meaning. As Goh (2002) and Vandergrift

(2003) put it, listening is a problem-solving process which includes a combination of strategies used in an orchestrated way. Certainly, in the instances where the item was answered correctly by simply understanding an idea unit, meaning was created by using inference, monitoring and contextual clues.

Example:

> *Then Caroline says people use emoticons to change the tone… I chose this one by discarding the others. The last two I didn't hear anything. The first two she said something about the feelings and so… but I didn't hear about the tone… but feelings is more related with the tone than the look of the message.*
>
> **(Participant 7, Q3.8)**

One of the problems which was mentioned about this task was the speed of delivery, participants believed the speakers to be speaking very quickly. It should, however, be noted here that research findings have shown that this is indeed the case for conversations (Tauroza & Allison, as cited in Buck, 2001). People speak much more quickly when having a conversation than when giving a talk or presentation and so this is an accurate representation of the TLU.

Examples:

> *It has been more difficult than I thought, in one part I have lost the connection, the link between the listening and the questions, this is terrible because the cost is you have to leave them blank… but I have found it hard for me this task. The speed of the speaker was very fast.*
>
> **(Participant 6)**

> *Ok they speak really fast… so it was hard for me to follow all the questions But as I was remembering what she was saying, I was answering the questions. But she spoke fast. I had to read again and maybe it wasn't that easy. For example this one I didn't answer, cos I didn't remember what she said.*
>
> **(Participant 9)**

We should also remember here that construct-irrelevant variance can be introduced by the extra cognitive load of reading the items and options for MCQ items. Reading is not part of the listening construct and if possible should be kept to the minimum by

having short questions and options. Other construct-irrelevant processes were also evident in this task, for example, guessing based on previous knowledge schemata.

Example:

> *Most of her data was collected… maybe over a long time because from existing corpus is not possible by her research group, I don't think so… I think that maybe an alternative with the most sense could be D.*
>
> **(Participant 6, Q3.3)**

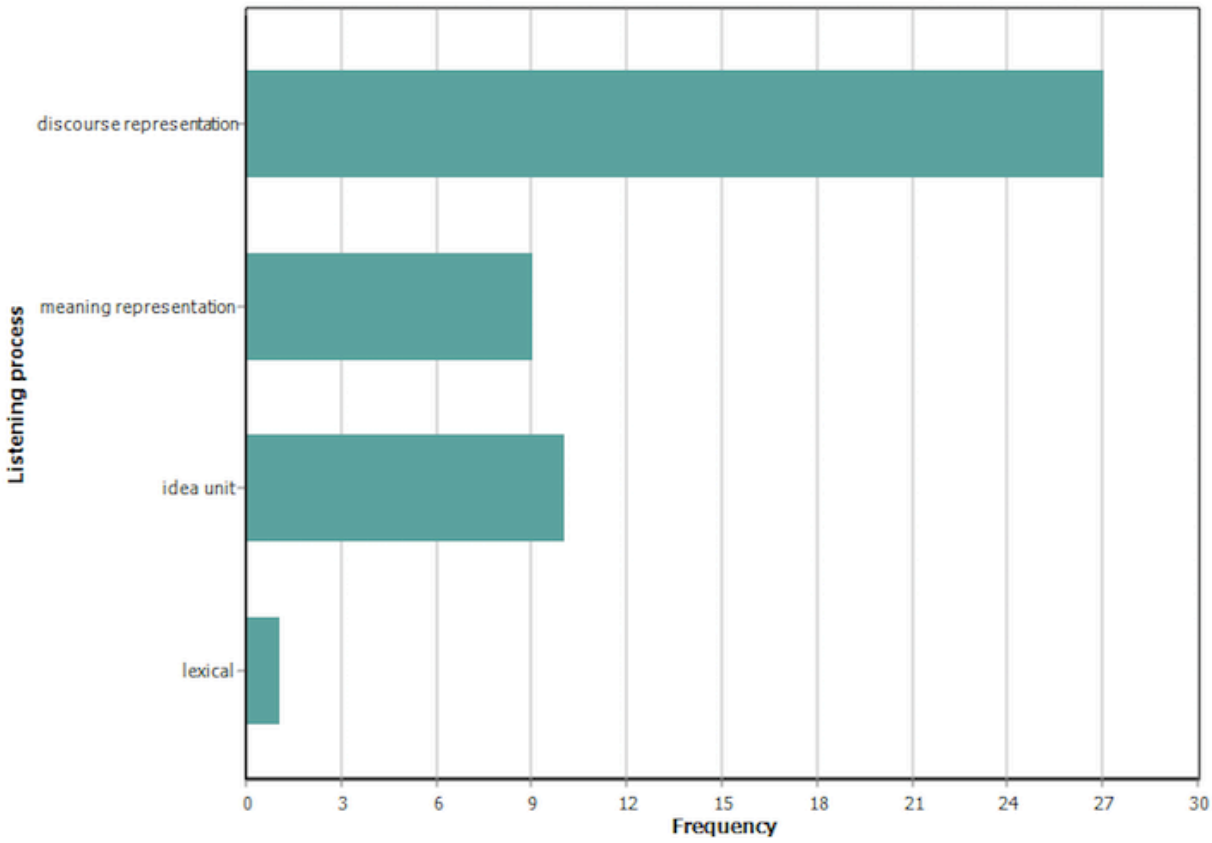
If we remember this was one of the items which was shown to be problematic in terms of Outfit and Zstd values and this could very well be due to this guessing factor. This participant (6), who was the lowest scoring, did not understand as much of the audio as the other participants and used guessing-type strategies on another item. This means that two of the items were answered correctly on this task without the proper recourse to the listening ability model. Indeed, guessing is a problem with MCQ type items and has been reported in a number of other studies (Barta, 2010; Yi'an, 1998). It does seem to be, however, that guessing is based on partial understanding and is informed by the co-text and what has been said before. These two instances were the only ones on the whole test where guessing led to the choice of the correct answer.

Example:

> *In general the majority of the answers come from guessing strategy because I haven't heard literally the answer… maybe D, change the tone… maybe I have heard something and it was my intuitive way…*
>
> **(Participant 6, Q3.7)**

Task 4 is a NF task and the intention is to test search listening for specific information and important details. Figure 28 shows the level of processing reached to answer the items on this task.

Figure 28. Highest level of processing used to answer Task 4



Here one thing which can be noted is the fact that none of the items were answered correctly by guessing. Most of the answers were correct due to a discourse representation of the text. However, it should be noted here that due to the nature of the input audio—a teacher giving a class quite factual instructions about an upcoming geography trip—the discourse was quite straightforward and contained mainly local factual information. The participants commented on the context and believed it to be representative of the TLU domain.

Example:

> *So you have to write three words… what she's trying to explain about a trip … ah ok I have to look for the words. (Reads through the questions). What they have to tell the company 'Information pack'… I'll listen for that word so there will be control when you go out and come in and more rules for students…*

> *where they go, the material they are going to collect, what they will compare it with. It will be like going to class… I'll imagine I'm in class and I have to understand what she tells us.*
>
> **(Participant 9)**

In general this task required the application of linguistic knowledge. Strategy use was limited to predication and monitoring and no instances of using contextual clues to infer meaning were reported, although intonation patterns were used as clues that signalled important information.

Examples:

> *Here, I understood something about medical records and she emphasised this as if it was something really important she wanted to say.*
>
> **(Participant 4, Q4.2)**

> *They have to check information pack for I think it's clothing and weather because she said that because of the weather they have to chose one clothing or another one. 'Check information pack' ….before I put clothing and weather now I think it's only clothing (crosses out weather).*
>
> **(Participant 5, Q4.3)**

> *She says 7th of May. In two weeks before they set off. At first I thought she was going to talk about money or say 'before we depart'. At first I thought it was going to be money. Here I understood everything.*
>
> **(Participant 4, Q4.1)**

To summarise, it can be seen that the participants demonstrated that they were using the knowledge and skills proposed by the BFE listening ability model in order to solve test items. Consequently, this part of the study has provided strong construct validity evidence supporting the explanation and extrapolation inference of the overall validity

argument. As expected the B2 level participants demonstrated a high level of automated listening ability and were able to build meaning from the audios in order to answer test items. With very few exceptions, the processing which led to choosing the key is 'passage dependent', as it should be (Buck, 2001, p.126), and completely reliant on recourse to the audio input.

Indeed, Field (2016) argues that students who have had extensive listening practice begin to recognise chunks of language and more basic operations become automatic. This means that due to the reduced demands on the working memory they are able to perform more higher order processes. These processes were necessary in order to solve the test items and very few participants managed to solve test items by simply relying on the successful decoding of isolated words. This observation can be seen in more detail in Table 20, which gives the level of processing reached on an item by item basis. It can be seen from Table 20 that the most difficult items on the test for this small group of participants were Q3.5 and Q4.4, which were two of the most difficult items shown by the Rasch analysis of test scores. However, the most difficult item (Q4.8) was answered correctly by five of the participants. It can be seen, however, that their correct answers were mainly based on low level processes; they were unable to fully understand the idea in the audio and the correct answer was shown to be uncertain.

Examples:

> *River maybe? I heard it the first time.*
>
> **(Participant 6)**

> *Will be collected from… I don't know I'm not sure, I didn't hear it… could be local rivers… it makes sense, it matches.*
>
> **(Participant 9)**

Table 20. Frequencies of level of listening process reached for each correct item

	Lexical recognition	Idea unit	Meaning representation	Discourse representation	Number of correct responses (N=7)
Q1.1	1	0	2	2	5
Q1.2	0	0	3	0	3
Q1.3	0	2	2	0	4
Q1.4	0	1	4	0	5
Q1.5	1	2	2	0	5
Q1.6	2	2	3	0	7
Q1.7	0	1	2	1	4
Q2.1	0	0	2	4	6
Q2.2	0	0	2	5	7
Q2.3	0	0	2	4	6
Q2.4	0	0	2	3	5
Q2.5	0	0	3	4	7
Q2.6	0	0	2	2	4
Q2.7	0	0	3	4	7
Q2.8	0	1	2	2	5
Q3.1	0	0	2	5	7
Q3.2	0	0	5	2	7
Q3.3	0	0	2	3	6 (1)
Q3.4	0	1	0	2	3
Q3.5	0	1	1	0	2
Q3.6	0	0	1	4	5
Q3.7	0	0	2	3	6 (1)
Q3.8	0	1	2	3	6
Q4.1	0	2	1	1	4
Q4.2	0	1	2	4	7
Q4.3	0	0	2	1	3
Q4.4	0	1	0	1	2
Q4.5	0	0	1	2	3
Q4.6	0	1	1	5	7
Q4.7	0	0	0	6	6
Q4.8	1	2	1	1	5
Q4.9	0	1	1	3	5
Q4.10	0	2	0	3	5

I will now go on to analyse the verbal reports for the other items which were shown to be badly functioning by the Rasch analysis to see if reasons for this can be explained.

Item 1.3F was answered correctly by four of the participants, who all used inference and to some extent a process of elimination in order to arrive at the answer.

Example:

> *The third one I heard something about competition and maybe TV… I needed to check… I understand only a little about the money in sports and TV and football. I think it could be TV channels compete to buy the most popular sports… I'm not sure but comparing it with the other possible answers… it's the only one.*
>
> **(Participant 8, Q1.3)**

The participants who got the item incorrect understood the ideas of investing money and sports events becoming more expensive but they did not make the inferential link necessary to arrive at the correct answer. Consequently it seems that the problem with this item is that the audio is too vague and not enough contextual information is given.

Item 2.8C was answered correctly by five participants, all of whom did not seem to have a problem with the item. The participants who got the item incorrect understood some of the content but were not convinced by the answer, which would explain the item statistics. That is to say, some candidates who should have arrived at the correct answer did not do so.

Example:

> *Maybe A or B. Mexican history maybe I'm not sure… sometimes I think in order to be usual Cambridge don't include controversial problems and I think about this here… countries problems… this topic in Cambridge no, so maybe history as this is more neutral… this is my strategy. He said that it's important for American people to know why he has a Mexican passport or something similar or maybe it's related to history but maybe B.*
>
> **(Participant 6, Q2.8)**

Item 4.10 was answered correctly by five participants but the item was shown to be problematic even for those participants who got it correct.

Example:

> *Should include different methods and digital photos… Include different methods, for example digital photos. (Reads stem). 'Might' —so digital photos?*
>
> **(Participant 4, Q4.10)**

Indeed the problem here seems to be with the stem and the word 'might' was not acted upon to represent things which *could* be included. All the participants who got this item incorrect put 'different/mixed methods' even if they had understood that digital photos could be included; this explains the statistical properties of the item.

The two other items which need to be examined as possibilities to be dropped are 3.3 and 4.6. Both items were easy for the seven participants (all seven got both items correct). It has already been seen that 3.3 was answered correctly in one instance by a reliance on guessing strategies. However, 4.6 does not seem to be problematic and the participants answered correctly due to an understanding of the audio. It was therefore decided that the final version of the test would not include Item 3.3. In the main, all the items answered incorrectly by the participants were done so because they missed the information or were unable to decode sufficient input.

In conclusion, it has been seen that the BFE listening ability model was very much evidenced. Listening is a cognitively demanding task which involves decoding and understanding a message and then using that information to complete a task. Here, working memory plays an important role (Juffs & Harrington, 2011). These real-life cognitive processing demands need to be incorporated into the testing situation (Wagner, 2013a). Using authentic sound files with purposeful items, and basing the listening activity on expert behaviour is a good way of doing this. The present study incorporated *text mapping* into the test development process in an attempt to mirror the cognitive processes used by expert listeners. The verbal reports certainly seem to show that the participants in this study were using the real life cognitive process demands proposed by the BFE listening ability model. The research methodology has shown to be invaluable as

a tool for investigating the psycholinguistic validity of item response patterns. The data from this part of the study provides strong triangulation supporting the probabilistic quantitative Rasch analysis and gives information about just how items were answered correctly.

6.4 R4: Are scores on the final test form reliable?

The piloting of items before a live test administration is an essential part of any test development cycle; in order to have confidence in the test as a reliable measurement instrument only items which have been shown to function correctly should be included in the final test form. The previous research questions have identified the items which are not working well and have provided the information necessary to construct the final 28-item test form. A detailed Rasch analysis of this final BFE test follows. The final test can be seen in Appendix 4 and the audio file is included as a CD in Appendix 5.

The summary statistics of persons shown in Table 21 shows that the candidates have an *Infit MNSQ* of 1 and a *SD* of .15, giving evidence that their behavior follows that expected by the Rasch model quite well. Person separation is 2.37 and person separation reliability is .85. This gives evidence that the test can distinguish between at least two performance levels in the sample reliably, that is, that the test contains a sufficient number of items to distinguish between high and low performers. The Rasch average standard error of measurement (SEM) is 0.11, which shows that the ability levels of the candidates reported by the Rasch analysis are very precise. Cronbach Alpha is also reported as 0.88 (the test SEM using CTT is 2.15), slightly higher than the original test form, giving evidence from CTT that the test is a reliable measurement instrument. *Person raw score-to-measure correlation* is the Pearson correlation between raw scores and measures, and is reported to be .98, close to the expected value of near 1.0.

Table 21. Rasch summary of 154 measured person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	17.1	28.0	.73	.51	1.00	.1	1.01	.1
S.D.	6.3	.0	1.41	.15	.15	.7	.56	.8
MAX.	27.0	28.0	3.69	1.03	1.41	2.2	5.33	2.6
MIN.	5.0	28.0	-1.81	.42	.63	-1.9	.33	-1.9

REAL RMSE	.55	TRUE SD	1.30	SEPARATION	2.37	PERSON RELIABILITY .85
MODEL RMSE	.53	TRUE SD	1.31	SEPARATION	2.44	PERSON RELIABILITY .86
S.E. OF PERSON MEAN = .11						

PERSON RAW SCORE-TO-MEASURE CORRELATION = .98

CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .88

In addition to the reliability evidence presented above a more detailed analysis of person behavior can be carried out. Winsteps, *Table 6* was examined in order to identify any misfitting persons. This table shows that there were no misfitting persons (using the 0.5 to 1.5 parameters) in terms of *Infit,* which gives evidence that the test produces a reliable ability measure for all candidates.[58]

Similarly, the summary statistics of items is shown in Table 22. Here, it can be seen that item separation is 4.58 and item separation reliability is 0.95. These results give evidence that the sample size used in the study was big enough to give stable item estimates, especially if we consider that the test does not have a very wide range of item difficulties (it only intends to test one CEFR ability level). The *item raw score-to-measure correlation* (Pearson correlation between raw scores and measures, including extreme scores) is -1.0, confirming that a higher measure implies a lower probability of success on an item (Linacre, 2017a). It can therefore be argued that item difficulties are reliable and a similar group of candidates would produce the same results. These results show test reliability and also contribute to evidence for the construct validity of the test.

Table 22. Rasch summary of 28 measured item

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	94.0	154.0	.00	.21	1.00	.0	1.01	.0
S.D.	24.3	.0	.98	.02	.10	1.0	.21	.9
MAX.	135.0	154.0	1.59	.26	1.21	2.1	1.48	2.0
MIN.	53.0	154.0	-1.83	.19	.84	-1.9	.59	-1.9

REAL RMSE	.21	TRUE SD	.96	SEPARATION	4.58	ITEM	RELIABILITY	.95
MODEL RMSE	.21	TRUE SD	.96	SEPARATION	4.66	ITEM	RELIABILITY	.96
S.E. OF ITEM MEAN = .19								

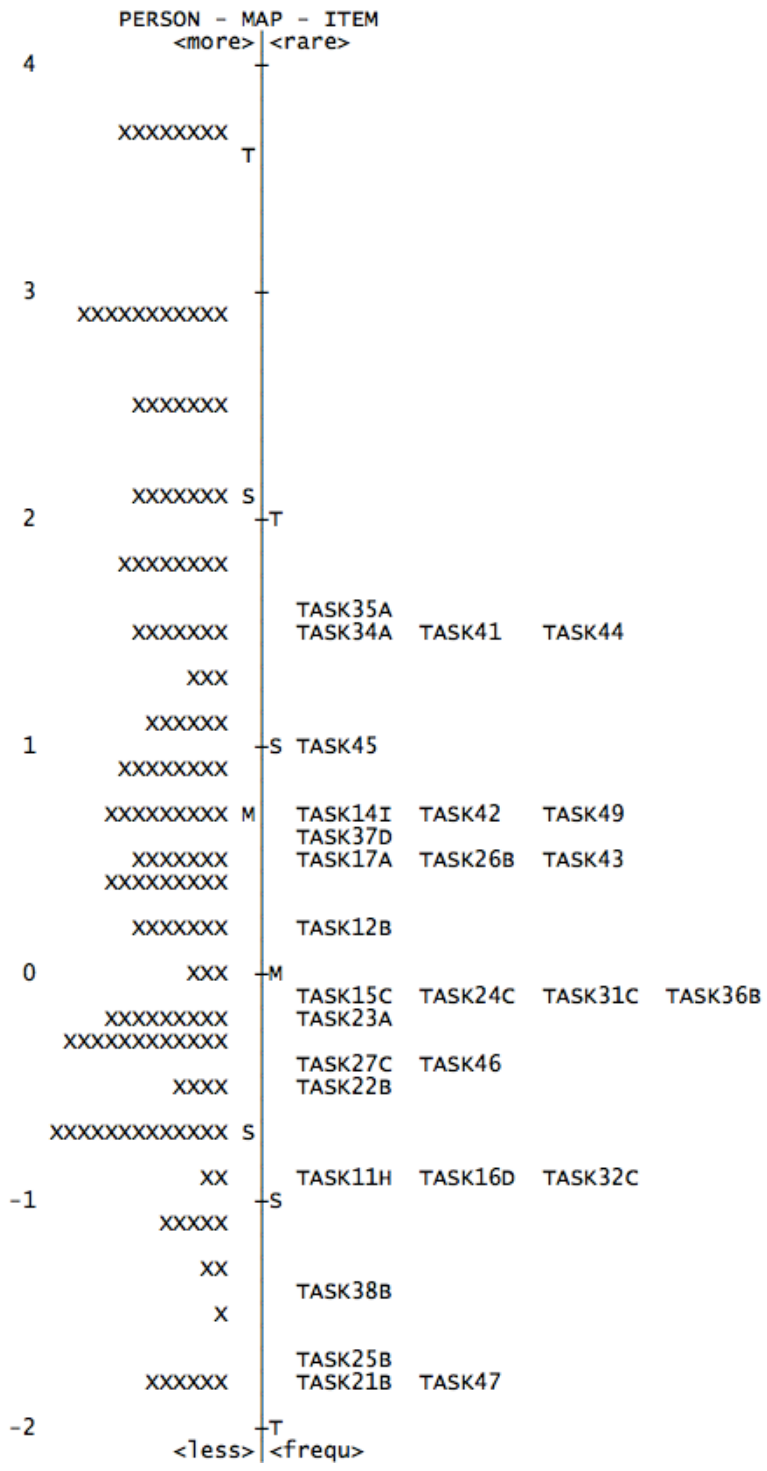
ITEM RAW SCORE-TO-MEASURE CORRELATION = -1.00

The final item difficulty hierarchy can be seen in Figure 29, which shows the item variable map having items spread along the difficulty scale, all falling within 2 SDs of the mean. Candidate abilities have a mean which is slightly higher than item difficulty, showing that the candidates had an ability level slightly higher than test difficulty. However, anecdotal evidence from the candidates' teachers suggests that many of the

[58] Here, McNamara (1996) argues that person misfits should not exceed 2% of the population.

participants have a CEFR B2 level and this would therefore be as expected.

Figure 29. Item variable map for final 28 item test



The Rasch individual item analysis of the final 28-item test can be seen below in Table 23. The items have an *Infit MNSQ* of between the acceptable values of 0.84 and 1.21, confirming that all the items on the test are productive for measurement. High *Outfit MNSQs* are less of a threat to measurement and are usually caused by a few random responses by low ability candidates or carelessness by higher ability candidates. However, for the present test all *Outfit MNSQs* are less than 1.5 and as such are acceptable.

Table 23. Item Statistics from Rasch Analysis for final 28 item test

Entry	Item	Measure in	Model	Infit		Outfit	
		Logits	SE	MNSQ	Zstd	MNSQ	Zstd
1	Task1.1H	-.86	.21	.95	- .5	.79	- .7
2	Task1.2B	.17	.19	1.18	2.1	1.24	1.4
3	Task1.4I	.67	.19	1.02	.2	1.01	.2
4	Task1.5C	-.05	.19	.98	- .2	.91	- .5
5	Task1.6D	-.86	.21	1.02	.3	.93	- .2
6	Task1.7A	.46	.19	.96	- .4	.91	- .6
7	Task2.1B	-1.83	.26	.94	- .3	1.26	.7
8	Task2.2B	- .52	.20	1.09	1.0	.97	- .1
9	Task2.3A	- .17	.19	.90	-1.3	.96	- .2
10	Task2.4C	- .09	.19	.89	-1.4	.87	- .7
11	Task2.5B	-1.70	.25	.91	- .5	1.44	1.0
12	Task2.6B	.46	.19	.85	-1.9	.74	-1.9
13	Task2.7C	- .40	.20	1.15	1.7	1.19	.9
14	Task3.1C	- .09	.19	.98	-.3	.88	- .6
15	Task3.2C	- .86	.21	.86	-1.4	.86	- .4
16	Task3.4A	1.47	.20	1.21	2.0	1.29	1.7
17	Task3.5A	1.59	.20	1.14	1.3	1.25	1.4
18	Task3.6B	- .13	.19	.96	- .4	.81	-1.0
19	Task3.7D	.60	.19	.90	-1.1	.84	-1.1
20	Task3.8B	-1.40	.24	1.04	.4	1.18	.6
21	Task4.1	1.47	.20	1.01	.2	1.11	.7
22	Task4.2	.67	.19	.94	- .7	.93	- .4
23	Task4.3	.46	.19	1.06	.7	.97	- .1
24	Task4.4	1.51	.20	.99	- .1	.95	- .2
25	Task4.5	1.04	.19	1.01	.1	.97	- .1
26	Task4.6	- .44	.20	1.05	.6	1.48	2.0
27	Task4.7	-1.83	.26	.84	- .9	.59	- .9
28	Task4.9	.67	.19	1.06	.7	1.01	.1
	Mean	.00	.21	1.0	.0	1 .01	.0
	SD	.98	.02	.1	1.0	.21	.9

The t-test significance (Zstd) values are mostly acceptable (within the ± 2.0 range) for sample sizes between 30 and 300 (Bond & Fox, 2015, p.53). There are two items which have a *Infit Zstd* higher than the recommended value of ± 2. However, both items have MNSQs near 1.0 and therefore this indicates little distortion of the measurement system, regardless of the Zstd value (Linacre, 2017a); if mean-squares are acceptable, then Zstd can be ignored, (Green, 2013).

Point-measure correlations are all good (between 0.33 and 0.62) indicating that all items on the test are discriminating between higher and lower abilities. This measure indicates item polarity and as long as there are no negative or very low values (less than 0.1), it can be concluded that all items are measuring the construct being tested (Green, 2013). Values which are much higher than expected values show overly predictable responses whilst values which are much lower than expected values show unmodelled variation or noise in the data (Aryadoust, 2013). The model standard error (SE) is less than the recommended 0.3 (Linacre, 2017a) for all the items.

In sum, the Rasch analysis shows the 28-item test to be more reliable than the original 32-item test. All the items fit the Rasch model and the sample used in the study was large enough to provide precise difficulty estimations of the items. It can therefore be concluded that the candidates can meaningfully be compared on the measurement scale and that a higher score on the test does indeed mean a higher listening ability. This information can therefore be used to carry out a standard-setting study to determine the appropriate cut score which should be used to represent CEFR-B2 proficiency in listening.

6.5 R5: What are candidates' opinions of the BFE listening test? Do candidates believe that the test will have positive washback?

Candidate opinions about possible construct-irrelevant variance have already been reported in answer to research question two (R2). Here further candidate opinions will be reported in relation to:

1) Opinions about how the test reflects classroom practices for listening.
2) Opinions about listening processes and strategy use.
3) Opinions about the test as a fair measure of listening ability.
4) Opinions about present listening instruction and possible washback effects of including listening in the final baccalaureate test.

Table 24 shows the questionnaire results for the test as a reflection of present classroom practices where a higher score represents more agreement.

Table 24. Questionnaire results for test representing present classroom practices

Question	Mode	Mean	Standard deviation (SD)
1. The topic was typical of those I studied at school			
1.1 Questions about sport	4	2.89	1.06
1.2 Moving to the USA	3	2.88	.92
1.3 Text messaging	3	2.95	.95
1.4 Geography trip	3	2.64	.96
2. The audio was similar to those I studied at school			
2.1 Questions about sport	1	1.72	.87
2.2 Moving to the USA	3	2.41	1.01
2.3 Text messaging	3	2.42	1.05
2.4 Geography trip	2	2.18	.97

It can be seen that in general the topics which were included on the test were representative of those studied at school during baccalaureate—especially for Task 1 (sport), which has a mode of 4 ('completely agree'). There were, however, a number of participants who disagreed, possibly because they were thinking about the task as a whole rather than the topic of sport, which was indeed covered in all the text books analysed. Also, for this task the opposite was found to be true for opinions about the soundfile, which has a mode of 1 ('completely disagree'). The soundfile for Task 1 was extremely authentic, with the speaker answering questions in a very conversational style. It would seem, then, that this type of authentic listening is not being used in the classroom. Certainly, the textbooks which were examined during the present study did not contain any such soundfiles. I would argue that this is something which needs to be addressed and there needs to be a shift in listening instruction towards dealing with authentic discourse. By introducing such audios on the test classroom practices would have to change. The audios used in Tasks 2 and 3 were considered to be more similar to those presently used, even though both these audios were also authentic tracks taken from the internet.

Opinions about processes and strategies used in order to answer test items can be seen in Table 25, where again a higher score means higher agreement. The strategy of using the context and co-text to guess unknown words has a mode of 4, completely agree. Planning and monitoring by comparing understanding with previous knowledge of the topic also showed high agreement and were both important strategies shown to be used by the participants in the verbal reports. These three questions about strategy use show that the candidates believed that they were using important metacognitive strategies which are included in the BFE listening ability model in order to answer test items. This gives evidence that the test can be considered to be a representation of the proposed listening ability model. Indeed, the first three strategies are included in the MALQ questionnaire (Vandergrift et al., 2006) as they are considered to be important metacognitive strategies necessary for successful listening.

Table 25. Questionnaire results for opinions about process and strategy use

Question	Mode	Mean	Standard deviation (SD)
1. I planned how I would listen	3	2.78	.83
2. I used known words to guess unknown words from context	4	3.31	.76
3. I compared my understanding with previous knowledge of the topic	3	2.91	.83
4. I guessed the answer			
4.1 Questions about sport	2	2.56	.90
4.2 Moving to the USA	2	2.21	.91
4.3 Text messaging	2	2.36	.81
4.4 Geography trip	2	2.61	.79
5. I understood the main ideas			
5.1 Questions about sport	3	2.75	1.00
5.2 Moving to the USA	3	3.47	.65
5.3 Text messaging	3	3.22	.68
5.4 Geography trip	3	3.13	.77
6. I could follow the audio			
6.1 Questions about sport	3	2.68	.99
6.2 Moving to the USA	4	3.51	.64
6.3 Text messaging	3	3.14	.79
6.4 Geography trip	3	2.90	.89

The construct irrelevant strategy of guessing the answer shows a mode of 2, disagree. This replicates the results of the verbal reports, which showed that the participants were generally not using guessing strategies. The participants believed that on the whole they understood the main ideas and could follow the audio. This is especially true for Task 2, which was seen to be the task which they felt was the easiest (see section 6.2.2). Here it should be noted that the CEFR listening descriptors for B2 include "can use a variety of strategies to achieve comprehension including listening for main points checking comprehension by using contextual clues" (CoE, 2001 p.72). The fact that the participants believed they were using these strategies gives evidence that the tasks elicited CEFR B2 strategies and that the participants were, in general, using CEFR B2 listening 'can dos'.

Table 26. Questionnaire results for opinions about the test as a fair measure

Question	Mode	Mean	Standard deviation (SD)
How do you feel about the test as a fair measure of your English listening ability?			
1 Questions about sport	2	2.27	.976
2 Moving to the USA	3	2.97	.784
3 Text messaging	3	2.59	.784
4 Geography trip	2	2.26	.851

The results of whether or not the candidates believe the test to be a fair measure of their ability can be seen in Table 26. Students' perceptions of the test as a fair measure of their ability is an ethical aspect which should be taken into account by test developers. However, it should be noted here that candidate views about 'fairness' have been reported to be 'complex and varied' because candidates have different definitions of the construct (Harding, 2008). In the present study, the candidates believed that both Tasks 2 and 3 were generally a fair representation of their listening ability (both MCQ tasks) but they were not as satisfied with Tasks 1 and 4 as a fair measure of their ability. One possible explanation could be that they were more familiar with the MCQ task format and if the test was in fact introduced they would obviously have had more practice with both the MM and NF task format.

Further analysis can be seen in Table 27 which shows that opinions about fairness are generally related to success on the test, using a Spearman's rho correlation there is a medium association which is statistically significant between total score on each task and opinions about each task as a fair measure of ability; Task 1 $r_s = .406$, $p = .000$); Task 2 $r_s = .498$, $p = .000$); Task 3 $r_s = .473$, $p = .002$); Task 4 $r_s = .319$, $p = .000$). This mirrors the results found by Iwashita and Elder (1997).

Table 27. Comparison of mean score and opinions about test fairness

	N	Mean score on test
Task 1:		
1. Not satisfied	39	16.30
2. Not very satisfied	51	17.76
3. Quite satisfied	44	22.41
4. Very satisfied	18	27.33
Task 2:		
1. Not satisfied	7	12.71
2. Not very satisfied	28	15.21
3. Quite satisfied	80	20.41
4. Very satisfied	37	23.38
Task 3:		
1. Not satisfied	12	16.17
2. Not very satisfied	55	16.84
3. Quite satisfied	69	21.58
4. Very satisfied	16	25.25
Task 4:		
1. Not satisfied	30	16.23
2. Not very satisfied	62	20.02
3. Quite satisfied	50	22.20
4. Very satisfied	10	19.82

Table 28 shows the questionnaire results for students previous experience and general opinions about listening, as well as whether or not they believe a listening section should be included on the final baccalaureate test. It can clearly be seen that the participants believe that they did not do enough listening practice at school. On further analysis it can be seen in Table 29 that those who did actually believe that they did sufficient listening at school scored higher on the test.

Furthermore the participants seem to believe that the reason for this is because currently listening is not tested on the *selectividad* exam. Those participants who said they did not practise listening in class because listening is not on the *selectividad* exam were shown to receive a lower score on the test (candidates who agreed had a mean score of 19.8 and candidates who disagreed had a mean score of 23.2).

Table 28. Questionnaire results for previous experience and opinions about listening

Question	Mode	Mean	Standard deviation (SD)
1. At school we did enough listening practice in class	2	2.06	.83
2. At school we didn't do much listening practice because listening is not on the '*selectividad*' exam	4	3.16	.99
3. Listening is important for learning a language	4	3.93	.34
4. I practised listening outside school	4	3.51	.69
5. A listening section should be included on the *selectividad* exam	4	3.34	.79
6. I wish we has used authentic audios at school	4	3.55	.79
7. Listening strategy training would be useful	4	3.76	.47
8. My listening ability would be better if we had practised at school	4	3.77	.58

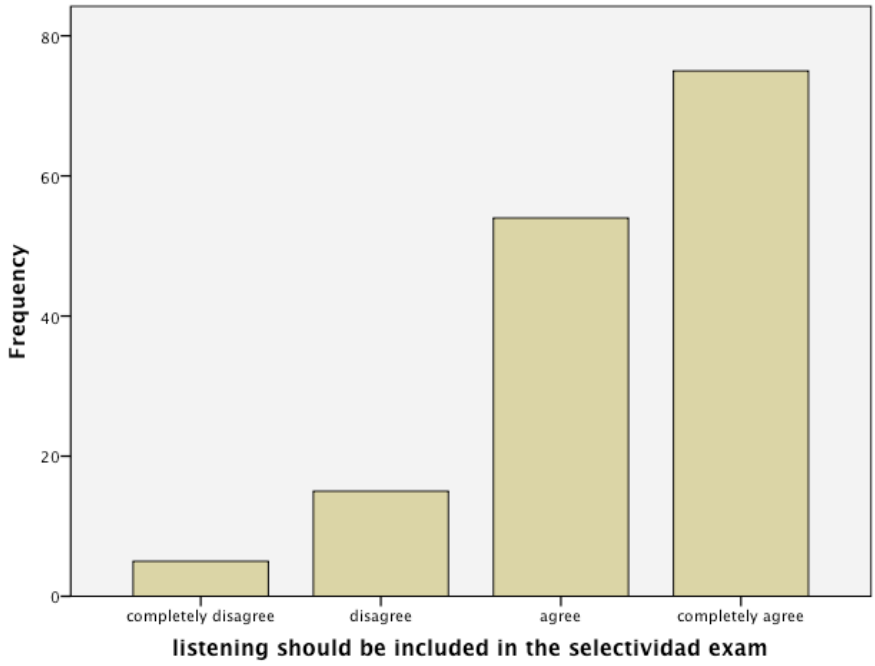
Table 29. Comparison of mean score and opinions about sufficient listening at school

	N	Mean score on test
1. Completely disagree	40	19.45
2. Disagree	68	19.32
3. Quite agree	35	20.69
4. Completely agree	7	23.71

Indeed, all the questions relating to introducing a listening section and therefore spending class time developing listening ability have a mode of 4; complete agreement. Finally, there was overwhelming agreement that a listening section should be included in the school leaving test. This can clearly be seen in Figure 30 where the majority of participants were in complete agreement. The participants are important stakeholders in the school-leaving exam and their views should be taken on board. They obviously believe that listening is important for language learning, and they would like listening

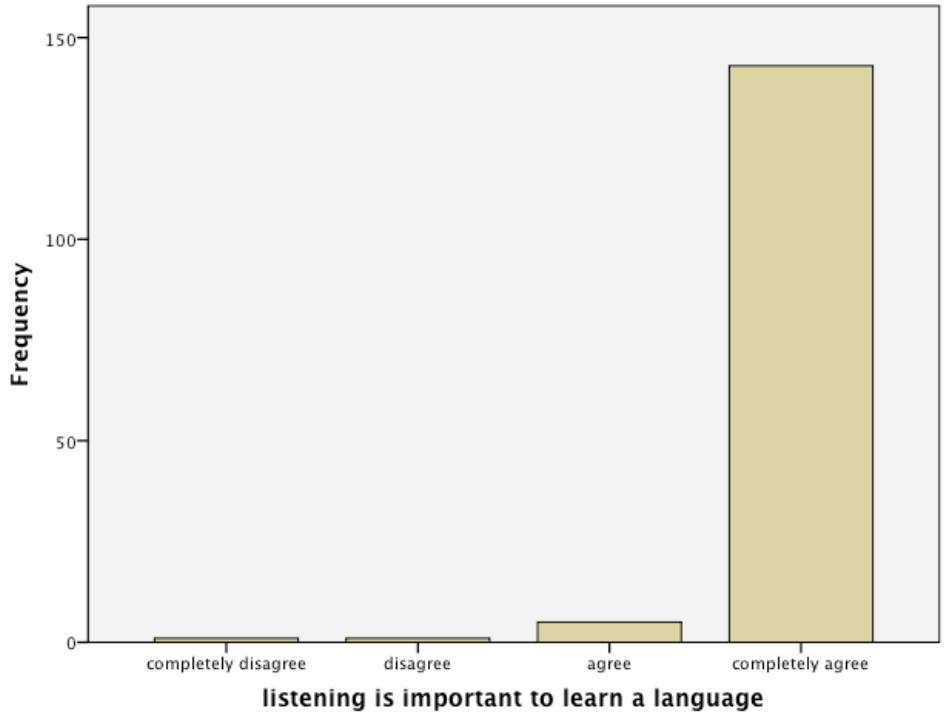
ability to be taught in class, they believe that this would lead to better understanding and that a measure of listening ability would therefore be desirable as part of a measure of their overall English language ability. Here, using *Spearman's Rho,* there was no correlation between views about the inclusion of a listening section and total score on the test. That is to say, all ability levels believed a listening section should be included on the test and opinions were not influenced by listening proficiency.

Figure 30. Histogram of opinions on whether a listening section should be included in the final school leaving exam



In terms of potential washback it would therefore seem that the students themselves are in agreement with the notion that by including a listening section on the school leaving/university entrance exam positive washback would be achieved. It is overwhelmingly believed that listening is an important activity which leads to language learning. This can be seen in Figure 31, where only two participants disagreed.

Figure 31. Histogram of opinions about listening being important to learn a language



The open-ended question, which requested any extra comments from the participants was only answered by eight people. Three of them commented on the test itself, stating that topics should be more common, that Task 1 was difficult as the speaker mumbled, and that there was not enough time to complete Task 4, respectively. The other five comments referred to opinions about listening instruction at school and can be seen below:

1. *It is difficult for me to understand audios because in school teachers didn't give it importance.*

2. *I was never satisfied with English at school so went to academy. Teaching is poor (just book). I don't think there should be listening in selectividad because I think audio quality will be bad.*

3. *Not only listening but reading and talking should be a lot more interesting at school.*

4. *It's important to hear authentic English accents, it's something we should have done before.*

5. *Please teach school teachers how to teach English properly.*

It can be seen that all the comments show dissatisfaction with the present baccalaureate curriculum and its teaching.

6.6 R6: Do expert judges believe the test tasks to be an accurate representation of the CEFR-B2 listening construct?

As explained in the methodology section, this part of the study is in effect a primer study for the standard-setting cut score study. All the judges who took part in the study were considered to have a thorough knowledge of the CEFR. This can partly be evidenced at the familiarisation stage of standard setting using activities such as CEFR descriptor matching exercises. The present study used a mix of listening descriptors taken from the CEFR scales and the judges were asked to allocate a CEFR level to each descriptor. The results gave a Cronbach's Alpha of 0.98. The correlation between CEFR descriptors and participant's judgement using a non-parametric Spearman's rank correlation coefficient can be seen in Table 30. All the correlations were strong, evidencing the judges understanding of the CEFR levels. These results were presented to the participants and further discussion took place in order to reach consensus on salient features of CEFR levels.

Table 30. Correlation between CEFR descriptors and judges allocations

Judge	1	2	3	4	5	6	7	8
	.810	.825	.929	.947	.903	.896	.980	.986

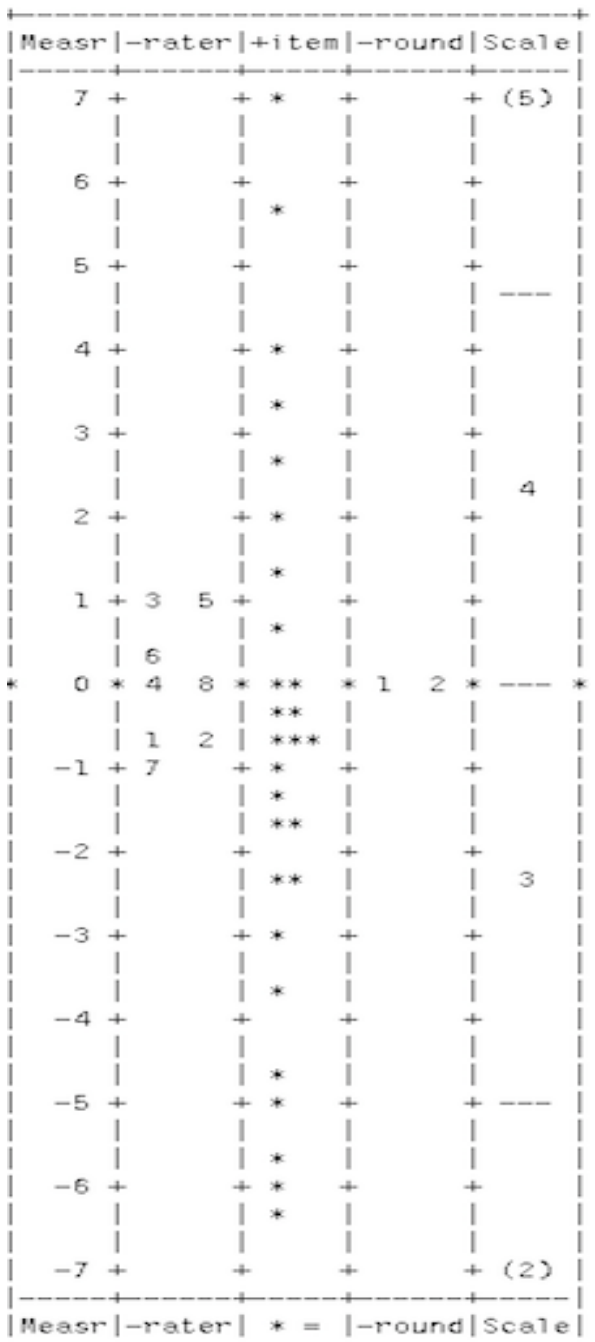
(*Note: all correlations were significant at* $p \leq .01$)

The test specifications and the test tasks themselves were explained to the participants in detail, including an explanation of which CEFR descriptors the tasks were intended to test and the listening skills they were intended to elicit. The participant judges then worked through the tasks on the test and following principles of the Basket Method answered the question 'at what CEFR level must test taker be in order to answer this item? After the first listening, round 1 results were collected and discussion was encouraged, judges were then permitted to change their CEFR level allocations and round 2 results were collected.

The results of this part of the study were then analysed using a Many-Facet Rasch Measurement model (MFRM) in the programme FACETS (Linacre, 2017b). Figure 32

shows the vertical ruler, which is a descriptive summary of the MFRM analysis of the results from all judges for each item on the test on both rounds. The analysis allows for multiple aspects of the judgements to be taken into account, and calibrates items, judges, rounds and the rating scale onto the same equal-interval scale.

Figure 32. Vertical ruler from FACETS analysis for Basket method results



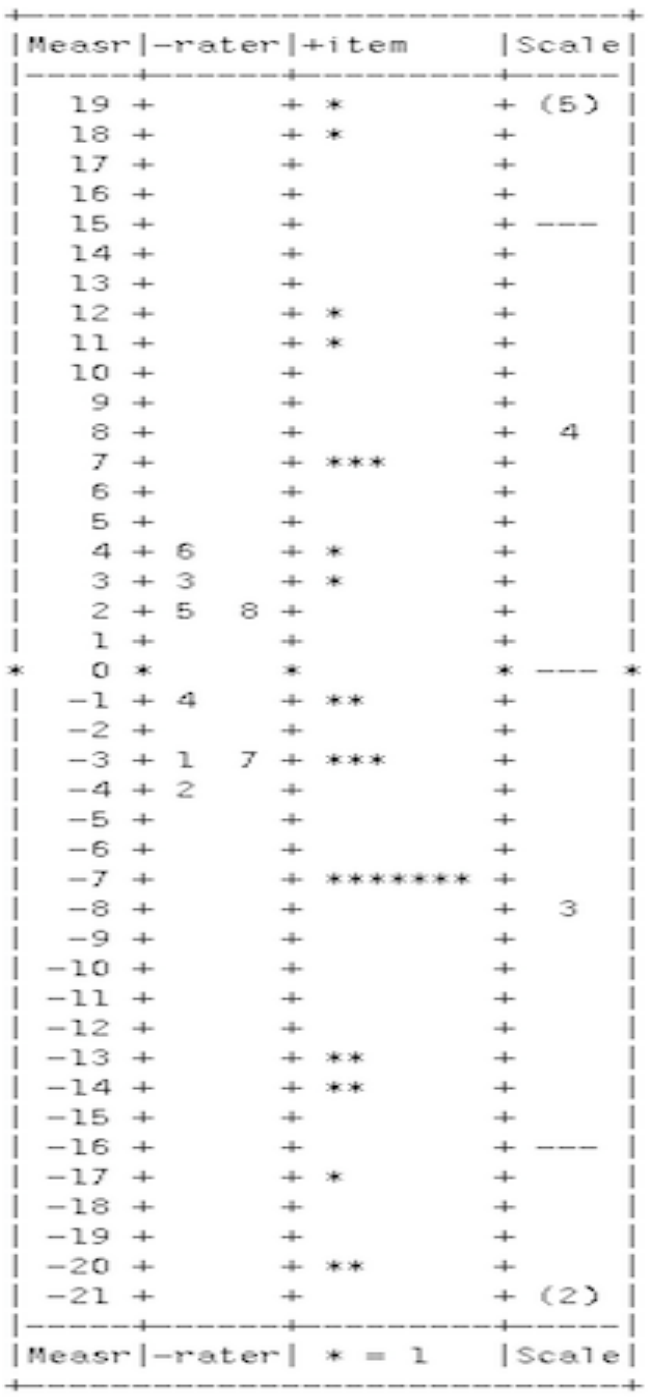
The scale used to represent the minimum CEFR level that a candidate should have in order to answer each item is as follows:

1= minimum B1
2= strong B1
3= minimum B2
4= strong B2
5= C1 and above

The results from the two rounds differed only minimally, though round 1 showed some *underfit*. However, the decisive results are those from round 2 following the discussion and so I re-ran the data for this round, the results of which can be seen in Figure 33. Here, there were no unexpected responses and it can therefore be concluded that the analysis did not produce any high standardised residuals, as it showed no large differences between the observed and expected responses.

It can be clearly seen that the judges believed the content of the items to be spread along the ability scale from CEFR B1 (3 items) to CEFR C1 (2 items). Most of the items were placed within the CEFR B2 category, with 16 items thought to be answerable by a minimum CEFR B2 candidate and 7 items thought to be answerable by a strong CEFR B2 candidate. This gives evidence that the judges believed the test to be representative of CEFR B2 listening.

Figure 33. Vertical ruler from FACETS analysis for round 2 Basket method results



However, it can be seen that the judges are separated and they had different severity measures, that is, some judges believed that items were more difficult than other judges and vice versa. Here, the most severe judge is judge 6, who believed the items to be

easier than the other judges. Nevertheless, this kind of variance is taken into account by the model and the position of item difficulty results from all judges' estimations after having taken into account their patterns of severity. In other words, the model treats judges as independent experts, each of whom brings something to the table. Table 30 shows the judge measurement report from FACETS output. Here it can be seen that the judges had 784 opportunities for agreement on ratings, yet exact agreements about the level of each item were only 585 (74.6%), very close to the 592.3 (75.6%) expected by the model. The model does not expect severe judges to agree with lenient judges. The separation index is 3.87 and separation reliability is .94, showing that the judges are rating differently and that this is not due to chance. This finding is also confirmed by the Chi-square statistic, which is significant at p = .00.

Table 31. Judge measurement report for round 2 of the Basket method

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	N rater
87	28	3.11	3.00	3.65	.80	1.20	.5	.41	1.1	.91	.91	.93	71.9	74.1	6 6
88	28	3.14	3.00	3.04	.77	.72	-.9	.19	.6	1.37	.96	.93	76.0	76.3	3 3
89	28	3.18	3.00	2.45	.77	.67	-1.0	.18	.5	1.41	.96	.93	78.1	78.2	5 5
89	28	3.18	3.00	2.45	.77	.96	.0	.26	.5	1.15	.92	.93	77.0	78.2	8 8
93	28	3.32	3.03	-.65	.90	.57	-.6	.14	.9	1.30	.93	.93	81.1	80.8	4 4
98	28	3.50	3.35	-3.49	.70	.83	-.4	.42	.5	1.17	.93	.91	71.9	73.1	1 1
98	28	3.50	3.35	-3.49	.70	1.16	.5	.41	.5	.97	.89	.91	70.9	73.1	7 7
99	28	3.54	3.46	-3.98	.71	.75	-.6	.26	.5	1.29	.93	.91	69.9	70.7	2 2
92.6	28.0	3.31	3.15	.00	.77	.86	-.3	.29	.7		.93				Mean (Count: 8)
4.7	.0	.17	.19	3.07	.06	.21	.6	.11	.2		.02				S.D. (Population)
5.0	.0	.18	.20	3.28	.07	.23	.6	.11	.2		.02				S.D. (Sample)

Model, Populn: RMSE .77 Adj (True) S.D. 2.97 Separation 3.87 Strata 5.49 Reliability (not inter-rater) .94
Model, Sample: RMSE .77 Adj (True) S.D. 3.19 Separation 4.15 Strata 5.87 Reliability (not inter-rater) .95
Model, Fixed (all same) chi-square: 137.1 d.f.: 7 significance (probability): .00
Model, Random (normal) chi-square: 6.9 d.f.: 6 significance (probability): .33
Inter-Rater agreement opportunities: 784 Exact agreements: 585 = 74.6% Expected: 592.3 = 75.6%

The judges appear to be acting as independent experts. The estimated discrimination values are between .91 and 1.41, showing a reasonable fit to the Rasch model (Linacre, 2017b). The internal consistency of the judges (or intra-rater reliability) is shown by Infit MnSq. For raters or judges in a situation where agreement is encouraged, recommended Infit Mnsq values are > 0.4 < 1.2 (Wright & Linacre, 1994), thus all values fall within the acceptable range. The Infit MnSq values along with their corresponding Zstd values give evidence that the judges are internally consistent and are applying the CEFR descriptor scale in a reliable manner. Here, there are many overfitting judges and all of the Outfit MnSq values are overfitting. This is to be expected in the present study as overfitting values are typical when central tendency is shown (Eckes, 2011). The full range of the

rating scale was not expected to be used with the same frequency, after all it is hoped that the judges believe that the items are mainly suitable for CEFR B2 level. It can therefore be concluded that all the ratings can be used to produce reliable difficulty results for each item on the test.

Table 32. Item measurement report from Basket method

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu	item
40	8	5.00	5.00	(20.79	1.91)	Maximum					.00	.00	3	T1Q4
38	8	4.75	4.95	18.18	1.01	.94	-.1	.48	1.8	1.26	.57	.61	12	T2Q6
34	8	4.25	4.04	12.15	1.14	.79	-.3	.33	1.5	1.42	.70	.70	6	T1Q7
33	8	4.13	4.01	10.83	1.22	.80	-.2	.30	2.1	1.37	.49	.49	16	T3Q4
32	8	4.00	4.00	7.44	2.97	.02	.1	.01	1.0	1.25	.00	.15	2	T1Q2
32	8	4.00	4.00	7.44	2.97	.02	.1	.01	1.0	1.25	.00	.15	14	T3Q1
32	8	4.00	4.00	7.44	2.97	.02	.1	.01	1.0	1.25	.00	.15	20	T3Q8
31	8	3.88	3.98	4.41	1.17	.67	-.4	.28	1.9	1.47	.45	.42	9	T2Q3
30	8	3.75	3.95	3.27	1.01	.94	-.1	.48	1.6	1.25	.57	.61	28	T4Q8
27	8	3.38	3.17	-1.20	1.35	.22	-.9	.10	.2	1.49	.92	.86	10	T2Q4
27	8	3.38	3.17	-1.20	1.35	.22	-.9	.10	.2	1.49	.92	.86	17	T3Q5
26	8	3.25	3.04	-2.72	1.14	.79	-.3	.33	1.5	1.41	.70	.71	21	T4Q1
26	8	3.25	3.04	-2.72	1.14	1.08	.3	.49	1.6	.95	.66	.71	23	T4Q3
26	8	3.25	3.04	-2.72	1.14	.79	-.3	.33	1.5	1.41	.70	.71	25	T4Q5
24	8	3.00	3.00	-7.49	3.78	.01	.6	.01	1.5	1.22	.00	.12	4	T1Q5
24	8	3.00	3.00	-7.49	3.78	.01	.6	.01	1.5	1.22	.00	.12	7	T2Q1
24	8	3.00	3.00	-7.49	3.78	.01	.6	.01	1.5	1.22	.00	.12	8	T2Q2
24	8	3.00	3.00	-7.49	3.78	.01	.6	.01	1.5	1.22	.00	.12	18	T3Q6
24	8	3.00	3.00	-7.49	3.78	.01	.6	.01	1.5	1.22	.00	.12	19	T3Q7
24	8	3.00	3.00	-7.49	3.78	.01	.6	.01	1.5	1.22	.00	.12	22	T4Q2
24	8	3.00	3.00	-7.49	3.78	.01	.6	.01	1.5	1.22	.00	.12	24	T4Q4
22	8	2.75	2.95	-12.66	1.01	.64	-1.3	.33	1.7	2.02	.63	.60	5	T1Q6
22	8	2.75	2.95	-12.66	1.01	1.24	.8	.65	1.8	.45	.52	.60	11	T2Q5
21	8	2.63	2.87	-13.73	1.08	1.24	.5	.92	1.1	.64	.67	.76	13	T2Q7
21	8	2.63	2.87	-13.73	1.08	.73	-.3	.36	.9	1.40	.77	.76	26	T4Q6
19	8	2.38	2.18	-17.14	1.35	2.35	1.3	1.23	.9	.29	.68	.86	1	T1Q1
17	8	2.13	2.01	-20.00	1.21	1.10	.3	.44	3.3	.88	.43	.50	15	T3Q2
17	8	2.13	2.01	-20.00	1.21	1.10	.3	.44	3.3	.88	.43	.50	27	T4Q7
26.5	8.0	3.31	3.29	-2.89	2.03	.59	.1	.29	1.5		.39		Mean (Count: 28)	
5.7	.0	.72	.73	10.69	1.15	.57	.6	.30	.7		.33		S.D. (Population)	
5.8	.0	.73	.75	10.89	1.18	.58	.6	.31	.7		.34		S.D. (Sample)	

With extremes, Model, Populn: RMSE 2.34 Adj (True) S.D. 10.43 Separation 4.46 Strata 6.28 Reliability .95
With extremes, Model, Sample: RMSE 2.34 Adj (True) S.D. 10.63 Separation 4.55 Strata 6.40 Reliability .95
Without extremes, Model, Populn: RMSE 2.35 Adj (True) S.D. 9.56 Separation 4.06 Strata 5.75 Reliability .94
Without extremes, Model, Sample: RMSE 2.35 Adj (True) S.D. 9.76 Separation 4.15 Strata 5.86 Reliability .95
With extremes, Model, Fixed (all same) chi-square: 1905.1 d.f.: 27 significance (probability): .00
With extremes, Model, Random (normal) chi-square: 34.9 d.f.: 26 significance (probability): .11

Table 32 shows the item measurement report and here we can see that a fair average for each item difficulty is given which has been adjusted by judge severity/leniency patterns. The Infit MnSq shows that there is only one underfitting item, the first item on the test (T1Q1), with a value of 2.35, though the Zstd is acceptable. There are quite a few overfitting items (Infit MnSq < 0.5), which means that the judgements were overly predictable, something that could be explained by the fact that the judges were in perfect agreement about the difficulty level of the item. Such a response is not surprising, as judges are encouraged to agree about item difficulty after discussion. Similarly, the Outfit MnSq overfit values are typical of overly predictable ratings with central tendency.

The overall resulting measures of the 28 items on the test can also be seen in Table 32. The average level of item difficulty is 3.31 (B2), and the average of the estimated item difficulty calculated by Rasch model is 3.29 (B2); the model estimate is extremely close to actual item difficulties given by the judges. Separation represents the differences in item difficulty, and larger values indicate greater differences. A separation index value of more than 2 is desirable, which means that there exists a significant difference between the individual item difficulties, Here, the actual value is 4.46 and the reliability of the separation index is .95. The Chi-square value which is significant at p=.00, shows that the differences between the item difficulties are significant and have not occurred by chance. Here, it can be seen that the judges believed that the easiest item on the test was T4Q7, which they believed to be a B1 item, and the most difficult item was T1Q4 which was judged to be a C1 or higher item. The item variable map (Figure 29) did indeed show T4Q7 to be one of the easiest items on the test (along with T2Q1) and T1Q4 to be one of the most difficult. Here, the quality of the expert judgements may be examined by comparing the Rasch difficulties of the items from the actual test with the expert judgement study to see how well they correlate. The Spearman's Rho correlation was r_s = .522, p = .004, showing a moderate positive linear correlation between judges' estimates and actual item difficulty, a finding which shows that the judges were able to predict item difficulty with some success.

The rating scale structure from the analysis can be seen in Table 33. The rating scale was composed of five categories (minimum B1 to C1 and above), yet it can be seen that category 1 was not used. It should be pointed out that the test items were developed to target only one level, CEFR B2, and so ideally, extreme categories would not be used. The judges believed the items could be answered by strong B1 (13%), minimum B2 (53%), strong B2 (30%) and C1 and above candidates (4%). In total, this would mean that they considered 83% of the items to be CEFR B2 level items. Here, the distribution of observations across categories is not uniform, but as Linacre (2002, p. 7) states "when investigating highly skewed phenomena …the long tails of the observation distribution may capture the very information that is the goal of the investigation". Indeed, it is to be hoped that the judges do not give many items a score of B1 or C1.

Table 33. Rating scale category structure

DATA					QUALITY CONTROL		RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts			Cum.		Avge	Exp. OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
2	29	29	13%	13%	-18.73	-18.75 .3			(-16.69)		low	low	100%
3	114	114	53%	66%	-7.54	-7.49 .4	-15.62	.42	-4.24	-15.60	-15.62	-15.62	99%
4	64	64	30%	96%	6.67	6.57 .2	.35	.42	4.97	.34	.35	.34	99%
5	17	9	4%	100%	17.98	17.78 .2	15.27	.63	(16.35)	15.26	15.27	15.26	100%
									(Mean)		(Modal)	(Median)	

The average measures increase in line with the categories and the Rasch Andrich thresholds showed that judges placed items into distinct categories which increase in difficulty level and are very close to the measures expected by the model, a finding further confirmed by the Outfit Mnsq, which should not exceed 2. Here, the low overfitting values indicate small variations in scores or 'muting'.

Figure 34. The Rasch-Andrich Rating Scale Model

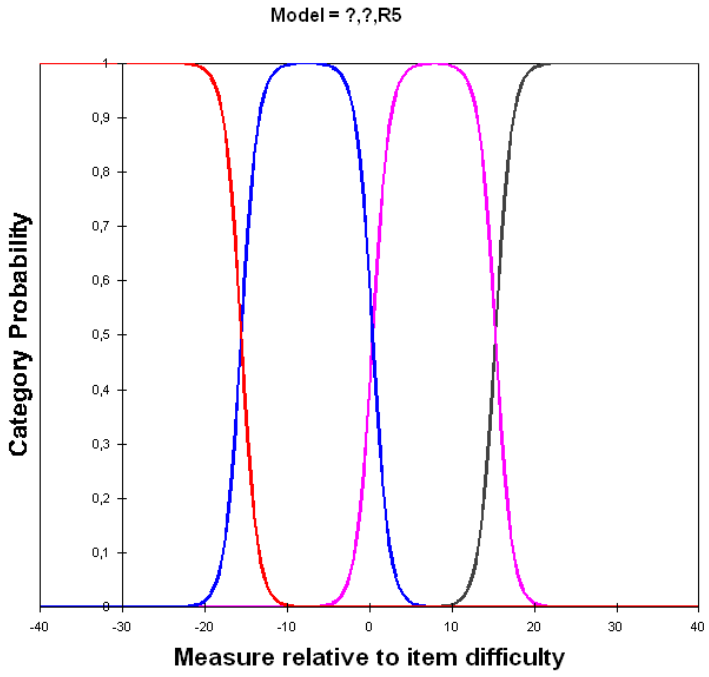
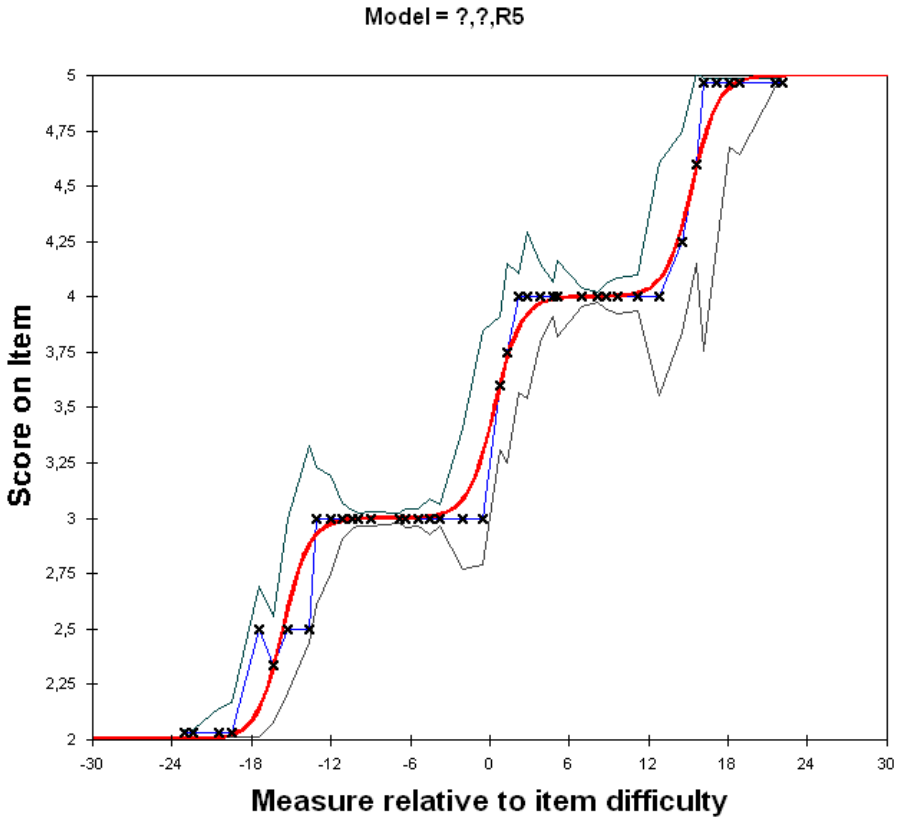


Figure 34 gives a graphical representation of the rating scale structure. Here, the probability curves show extremely nice 'hills' and represent category thresholds that advance monotonically, which indicates that the meaning of the rating scale was very clear for the judges involved in the study. However, it should be pointed out here that

when judges do not use the full range on the rating scale (central tendency/narrow range), rating scale category thresholds will be widely dispersed and separate distinct peaks will be shown for each rating scale category (Myford & Wolf, 2004). Also, data which contains very little randomness and is too predictable will expand the measurement system (Linacre, 2002) and so would make the item difficulties appear to be more different. This is exactly what is seen in the present data. Here, too much randomness or 'noise' in the data would be a bigger threat to measurement. However, the categories represent a wide range of item difficulties and this appears as a 'dead zone' in the middle of the category, where measurement precision is lost, as can be seen in the ICC curve in Figure 35. For good discrimination between items, a steep ICC curve would be expected rather than the flat steps showing the 'dead zone'.

Figure 35. Expected and Empirical ICC curve for Basket method results



Such discrimination often happens in situations which have forced consensus Guttman-style response sets (Linacre, 2002), where the judges are acting not as locally-independent experts but as 'rating machines'. In the present study this should not necessarily be considered a problem as consensus was encouraged through discussion and the CEFR B2 band can be considered to be a wide band. Instead of rating items on a continuous CEFR scale, a dichotomous yes/no study might have been more appropriate, but this would not have been obvious before the results had been analysed. Alternatively, in order to complete the inappropriate gaps in measurement, a future study could perhaps introduce a more finely grained rating scale such as 'minimum B2, 'low B2', 'mid B2' and 'high B2'. That being said, these results suggest that the judges showed high degrees of exact agreement and rated most items within the B2 band, which gives confidence in the reliability of CEFR-level estimations given for the items.

The above information gives evidence that the judges believed the exam to be representative of the CEFR B2 listening construct and can be used to suggest a preliminary cut score based on judges opinions. If the cut score is set to include all the items which judges believe could be answered by a minimally competent CEFR-B2 candidate, it should be set at 19 (out of 28). However, it should be remembered that this part of the study attempted to answer R6, as well as familiarise the judges with the test items in preparation for the main cut score study, where judges would be given empirical item difficulties and would therefore have more information on which to base their judgements.

6.7 R7: What should the cut score be on the test in order to provide an accurate evaluation of a CEFR B2 candidate? (Can parallel test forms be produced?)

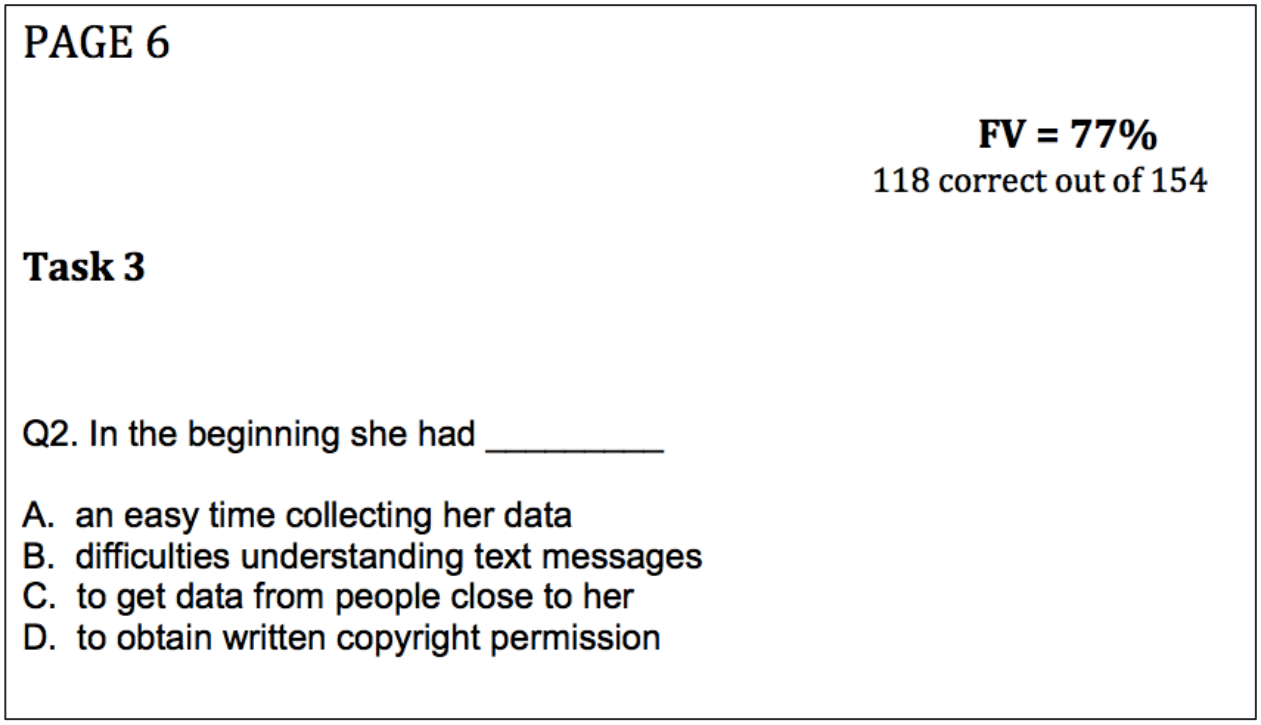
6.7.1 Bookmark cut score study

Once the Basket Method study had been carried out and judges were familiar with test content, participants were given feedback before moving onto the main Bookmark study. Impact data was presented using the cut score of 19 determined by the Basket method. The results were 67 (44%) pass and 87 (54%) fail. Unfortunately, the only other information available about the candidates is their self assessments. By comparing the cut score of 19 with self assessments, where n=151, 64.7% of candidates who believe they have a CEFR B2 listening level would pass, yet the caveat should be made that this self-assessment data could be incorrect due to the previously mentioned problems of candidates being unable to accurately give a self-assessment without prior training. Furthermore, it will be seen in section 6.8 that correlations between self-assessments and total score were not strong and consequently judges were advised to only take into account test content.

The Bookmark Method was explained to the judges and particular care was taken to ensure that they understood the concept of *Response Probability* (RP). In the present study this was set at .67 or a 2/3 chance of getting the item correct. The judges were asked to have in mind a minimally competent CEFR-B2 student and imagine that this student would have a 2/3 chance of getting the item correct. The judges were then presented with the ordered item booklet (OIB). The information about item difficulty presented in section 6.4 was used to develop the OIB, where each item was placed on one page and the booklet was ordered following item difficulty. For a listening test, this method used in isolation could be confusing for participants because items are not placed in the original test order. Indeed, CoE (2009) recommends that each participant should be provided with a computer so they can move between items in a non-linear fashion. However, in the present study the primer Basket Method study has provided a far more pragmatic solution to this problem and judges were already well familiarised with test

content. An example page of the OIB is shown in Figure 36, it includes the item and information about item difficulty in the form of the Facility Value (FV).[59]

Figure 36. A page from the OIB



The judges then worked through the OIB and placed an initial bookmark at the place where they believed that the minimally competent candidate would have a 2/3 chance of getting the next item correct. The results of round 1 were then collected. This was followed by in depth discussion and the replaying of a number of parts of the test audio, where necessary. Judges were then allowed to change the position of their bookmark and final results for round 2 were collected. The results of both rounds can be seen in Table 34.

[59] It was decided not to include logit difficulty measures as the judges are not familiar with these measures and could find them confusing.

Table 34. Cut score results using Bookmark method

Round 1	Page number (number of judges)	Mean (SD)	Median	Ability (θ) after adjusting for RP of .67	Final raw cut score
	13/14 (1) 14/15 (1) 15/16 (3) 16/17 (2) 17/18 (1)	15.13 (1.73)	15/16	0.64	17
Round 2					
	14/15 (2) 15/16 (6)	14.75 (0.46)	15/16	0.64	17

The median of the second set of bookmarks for all panellists is used as the group standard, and the lower value of theta should be used in order to set the cut score (CoE, 2009). The final median score was the same on both rounds, showing that the judges were confident in their decisions. The theta value for a score of 15 is -.05 logits (this is the ability level at a 50% probability of correctly answering the item). By adding 0.69 in order to adjust for a RP of 0.67 and mapping this back to the test characteristic curve, we get the necessary beta value (item difficulty) of 0.64 logits, which corresponds to a raw score of 17. This final cut score would give the following impact data: pass at CEFR B2 = 83 (54%), which, by comparing with self-assessments, is 72.5% of those who believe they have a B2 level or above. A final discussion then took place, which also took impact data into account. Participant judges were seen to be happy with the final decisions and in agreement that this should be used as the final cut score for the test.

Procedural validity evidence about the validity and reliability of the standard-setting process is presented in Table 35, which shows the results of the post standard-setting questionnaire. For each question, participants answered a 4-point Likert scale in which 1 represents complete agreement. Overall a very high level of agreement can be seen. The judges felt confident about both the procedure and their final decisions. Most notably, all judges felt that the final cut score decisions were a fair and accurate representation of CEFR B2 for listening.

Table 35. Post standard setting questionnaire results

	Mean	SD
I felt I had a sound understanding of CEFR levels for listening after the familiarisation sessions.	1.25	.463
The explanation of the test construct and specifications helped me.	1.25	.463
I understood the Basket method standard setting procedure.	1.13	.354
I felt confident about my decisions answering the question 'At what CEFR level must test taker be in order to answer this item?'	1.38	.518
I understood the concept of 'minimally competent candidate'.	1.00	.00
The Basket method standard setting exercise helped me to understand test content.	1.13	.354
I understood the Bookmark method standard setting procedure.	1.00	.00
I understood the concept of RP and what the 67% probability means.	1.13	.354
I felt confident about the placing of my bookmark on round 1.	1.75	.463
I felt confident about the placing of my bookmark on round 2.	1.00	.00
The final recommended cut scores of the group are a fair representation of CEFR B1 and B2 levels for listening.	1.00	.00

The concept of a minimally competent candidate was important for both the Basket Method study (R6) and the Bookmark Method study. Indeed, failure to understand this key concept could result in cut scores which are too high or too low (Pagageorgiou, 2010). Here, the questionnaire results show that the judges felt confident with this concept. Indeed, the high level of confidence in both the procedure and the final results could be put down to the fact that the judges had recently taken part in a standard-setting study that followed exactly the same procedure (Shackleton, forthcoming) and so they were well-practiced in the complete process.

Internal validity of the standard-setting study has already been evidenced by the results of the Basket Method. The standard error of judgement results of the Bookmark study are now presented in order to give evidence of classification accuracy. Cohen, Kane and Crooks (1999, p.364) argue that the SE_J should be $\leq$ 1⁄2 SEM. These results are presented in Table 36, where it can be seen that the SE_J was always much smaller than one-half of SEM, and the cut scores fulfil the quality criterion well.

Table 36. Results for classification accuracy

	Round 1	Round 2
Standard deviation of mean cut score (SD)	1.73	0.46
Standard error of test (SEM)	2.15	2.15
Standard error of judgement (SE_J)	.14	0.01
$\frac{SE_J}{SEM}$	0.065	0.001

External validity has been provided by using two different standard-setting studies. The two procedures provided different cut scores, 19 and 17. However, the Rasch score to measure table shows that a cut score of 19 would fall within one *standard error* of the final cut score of 17 taken from the Bookmark study, which gives evidence that the two standard-setting studies give similar results. Future validation studies would be necessary to give stronger evidence of external validity. The results presented in section 6.8, do however give some external triangulation evidence towards the standard-setting study. This analysis looks at the correlation between test scores and a self-assessment of candidates CEFR proficiency level.

6.8 R8: Do test scores correlate with similar measures of the same construct?

As previously stated, a self-assessment measure has often been used to correlate results on a test with a similar measure of the same construct. Table 37 shows the mean, mode and median scores on the test for each self-assessed CEFR level. It can clearly be seen that scores increase as self-assessed CEFR levels increase.

Table 37. Scores obtained on test compared to self-assessed CEFR levels

Self assessed CEFR level	N	Median score	Mode score	Mean score (SD)
CEFR B1	71	13	12	13.95 (5.06)
CEFR B2	66	20.5	23	19.79 (5.32)
CEFR C1 and above	14	25.5	26	24.43 (2 .7)

An initial correlation analysis was carried out using a non-parametric Kendall's tau between total score on the test and a self-assessment of CEFR level in listening. Here, three candidates had not provided an estimated CEFR level, so for this analysis N= 151. The results showed that there was a moderate positive correlation between the total score on the test and candidates' self assessment of CEFR level which was statistically significant; $\tau = .487$, $p = .000$.

Further analysis was carried out, where the independent variable is the self-assessed CEFR level and the dependent variable is total listening score. A Kruskal Wallis test, which is the non-parametric counterpart to one-way independent ANOVA, revealed that there was a statistically significant difference in median score on the four tasks and total listening across the three ability levels identified by candidates' self-assessment; for Task 1, $\chi 2$ (d.f. 2, n = 151) = 25.068, $p \leq .001$; for Task 2, $\chi 2$ (d.f. 2, n = 151) = 41.886, $p \leq .001$; for Task 3, $\chi 2$ (d.f. 2, n = 151) = 52.079, $p \leq .001$; Task 4, $\chi 2$ (d.f. 2, n = 151) = 31.050, $p \leq .001$ and for total listening, $\chi 2$ (d.f. 2, n = 151) = 51.667, $p \leq .001$. The test does not intend to separate B2 and C1 students, therefore post-hoc Mann-Whitney U tests

were carried out between B1 students and B2 and above students in order to find the effect size. For Task 1, Task 2, Task 3, Task 4 and total listening, r = 0.40, 0.49, 0.55, 0.41 and 0.55, respectively. These all have a medium to large effect size (Cohen, 1988; cited in Pallant, 2007). The effect size for total score on the test is large and the null hypothesis (there is no difference in the test scores according to self-assessed CEFR level) can therefore be rejected. This is supporting evidence that the test successfully separates CEFR B2 and above candidates from CEFR B1 candidates and providing a correct cut score has been set B1 level candidates would be unable to pass the test. By applying the final cut score of 17, it can be seen that this score is much higher than the mean, median and mode of the self-assessed B1 candidates. This suggests that they would be unable to pass the test.

A chi-square test of independence using a 2x2 contingency table comparing the self-assessed categories of B1 and B2 and above with the dichotomous categories of pass or fail (using 17 as a cut score) was also carried out. The Pearson's chi-square test examines whether there is an association between two categorical variables. There were no expected values of below 5. The cross- tabulation results in Table 38 show that 63 candidates who thought they should pass did so, but that 17 candidates failed who had predicted they would pass. Similarly, 51 candidates who thought they did not have a B2 level failed but another 20 candidates passed. Standardised residuals show that self-assessed B2 candidates are under-represented in the fail category (-3.2) and self-assessed B1 students are under-represented in the pass category (-3.0).

Table 38. Cross-Tabulation of Self-assessed CEFR level Vs. Pass on test

		Self-assessed CEFR level	
		B1	**B2 and above**
Test results	**Fail**	**51**	**17**
(pass mark is 17)		(3.4)	*(-3.2)*
	Pass	**20**	**63**
		(-3.0)	*(2.9)*

$\chi 2$=38.88 on 1 D.F. - P-value < 0.00 - Phi/Cramer's V = 0.507.

A chi-squared test of independence indicated a highly significant association between self-assessed level and performance on the test, $\chi 2$ (d.f. 1, n = 151) = 38.88, p = .000, Phi/Cramer's V = .507, giving evidence that the null hypothesis should be rejected. This gives further evidence that self-assessed B2 and above students are more likely to pass the test than those students who believe they are not a CEFR B2 level. The overall effect size was 0.507, which according to Pallant (2007) is a large association between the variables. However, it should be noted here that without specific training in self-assessments, the participants may not have been capable of correctly assessing their CEFR-related proficiency level in listening (see O'Sullivan, 2008).

Similarly, the relationship between test score and accreditation certificates reported to be held by the participants was examined. Such certificates included those awarded by Cambridge, ACLES, Trinity and EOI. Here, not all candidates reported holding such accreditations and N = 70. The correlation analysis using Kendall's tau between total score on the test and level of accreditation held is $\tau = .490$, p =.000, which shows that there is a statistically significant correlation between scores on the test and an external evaluation. Comparing two groups, B1 and below level accreditation and B2 and above level accreditation for the independent variable, a Mann-Whitney U test was carried out and it was found that test scores from candidates holding B1 or below certification (*Mdn* = 19) differed significantly from those holding B2 or above certification (*Mdn* = 26), $U = 34.5$, $z = -2.92$, $p < .001$, $r = -.35$.

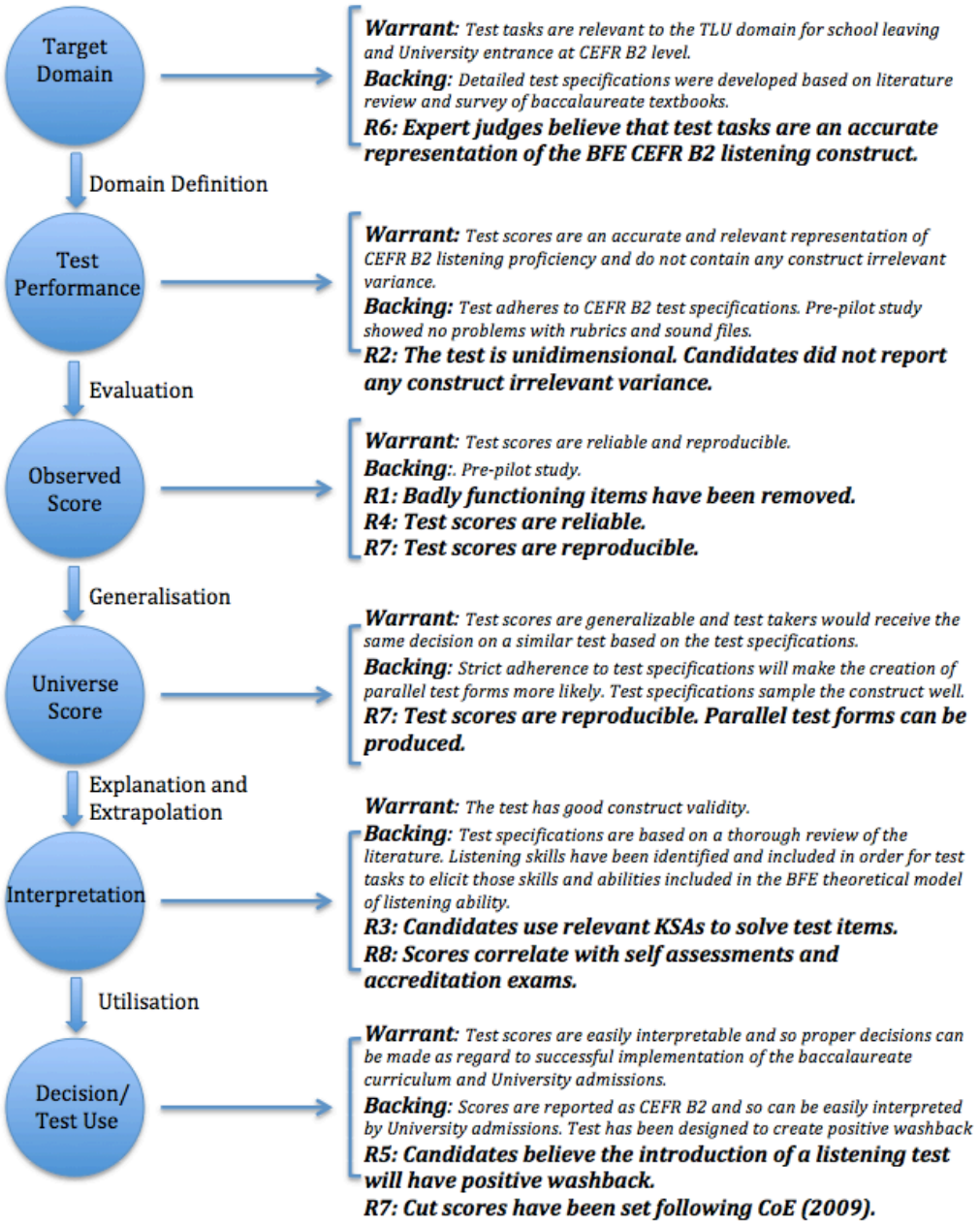
A final analysis was carried out comparing test scores with the candidates' reported score on the *selectividad* exam. A total of 127 students had given their *selectividad* score and a Spearman's rho correlation for this analysis was $r_s = .494$, $p = .000$, giving some evidence that actual marks received on the present school leaving exam have a weak to moderate positive linear relationship with the scores obtained in the present study. However, as there is no listening section on the present *selectividad* exam the results of both tests cannot be considered to be measures of the same construct and these results should not be presented as evidence in the final BFE validity argument. I would also note that due to the aforementioned problems of using a candidate self-assessment measure

that perhaps a more convincing alternate measure would have been a teacher assessment of the candidates. However, as this study was carried out during the first month of their university course the teachers did not yet feel qualified to give an accurate assessment of their students.

Chapter 7. Baccalaureate final exam (BFE) validity argument

This chapter will synthesise the results of the study and present the BFE final validity argument for test score interpretation and use. An attempt has been made to address all the inferences and their warrants—an extremely difficult feat within the confines of a single project. It may be that by examining the validity argument further, weak links will emerge which can be used to provide the necessary guidance for further investigation; after all, a validity argument is only as strong as its weakest link (Kane et al., 1999, p.15). While an accumulation of studies would be necessary to provide backing for each warrant and disprove all possible rebuttals, the present study has nevertheless provided some backing for each of the warrants in the BFE IA, and the final validity argument can be seen in Figure 37.

Figure 37. Final BFE CEFR B2 Validity Argument



The domain definition is based on the warrant that test tasks are a good representation of the TLU. The test specifications were largely based on an extensive review of the literature and so are founded on substantive theory. The detailed specifications provided include all the critical listening skills and cover the CEFR descriptors for CEFR B2; in so

doing, they place a strong emphasis on authentic listening behaviours. If correctly followed, every test version would have good construct coverage. As Hughes (2003, p.27) stresses, a test in which major areas identified in the specification are under-represented is likely to be inaccurate and to therefore have a harmful washback effect, since areas that are not tested are likely to become areas which are ignored in teaching and learning (as is the case in the present *selectividad*). Further backing was added to this warrant by R6 (and R7). Both the Basket Method and Bookmark Method procedures required extensive discussion about matching individual item difficulties to the CEFR levels. Consequently, the test's substantive and construct validity were placed under scrutiny and CEFR-aligned content was confirmed. In the Trinity ISE linking study, Kanistra and Harsch (2017) report similar results and conclude that such discussions contribute to identifying a minimally competent candidate and, in addition, an explicit relationship to the CEFR descriptor scales is articulated. Such an approach to defining a just-qualified candidate is certainly made easier by the fact that the test was specifically developed in order to operationalise the description of CEFR B2 proficiency. The fact that the standard has already been built into the test implies more meaning is given to the cut scores (Tannenbaum & Wylie, 2008, p.29-30). A test that had not been developed based on the CEFR would have to pay particular attention to the specification stage of content alignment as, if content is not aligned to the CEFR "there is little justification for conducting a standard-setting study" (Tannenbaum & Cho, 2014, p. 237).

Furthermore, R7 provided the relevant cut score to be used on the test. By associating a test score to a CEFR level meaning is added for test users (Kane, 2012). For example, if the score reports B2 proficiency in listening, the score user is informed about the kind of activities a person receiving this score is able to perform (Tannenbaum & Cho, 2014). Scores are easily interpretable, allowing score users to make relevant decisions and so this part of the study has contributed to the utilisation inference. Within any argument-based approach "validity evidence of the cut score is an essential component" (Papageorgiou & Tannenbaum, 2016, p.110) and "the crucial point in the process of linking an examination to the CEFR is the establishment of a decision rule to allocate students to one of the CEFR levels on the basis of their performance in the examination"

(CoE, 2009, p.11). This part of the study has certainly provided a stronger claim for the interpretation and use of test scores as being representative of CEFR B2 level listening proficiency. The standard-setting process has been described in detail as the core part of the linking process and, in doing so, it has been documented that a reasonable and systematic process was followed in order to reach the final standard. As Linacre (2002, p.858) states "it is the accumulation of information, not the ratings themselves, that is decisive". Good validity evidence has been provided to support the setting of recommended cut scores in which the standard-setting process may be considered a "blend of judgement, psychometrics, and practicality" (Hambleton & Pitonoak, 2006, p.435), and where "the question is not whether the cut score is correct but whether decisions based on the cut scores are reasonable, broadly acceptable and have mostly positive consequences" (Kane, 2017, p.11).

The domain definition warrant is therefore well supported. The Basket Method standard-setting study provided good internal validity evidence, shown by the fact that the judges estimated item difficulty level well and had good intra-rater reliability. Also, exact and expected agreements were extremely close and the group of judges appeared to be acting as 'rating machines', making a final decision that nearly all item were representative of CEFR-B2 level listening. Procedural validity evidence was also excellent and was improved by the fact that all the participants had already had experience in carrying out such a study. Perhaps the weakest part of the standard-setting study was the external validity evidence. The only available data was student self-assessment, information about other accreditation certificates, and scores on *selectividad* (for a limited number of candidates).[60] As previously stated, none of these external measures can be considered to be completely accurate. Future studies would therefore need to be carried out using more relevant external measures, and here a good external comparison could be made by using teacher judgements of candidates' CEFR level in listening. Once larger data sets have been collected, a prototype group method similar to the one reported by Eckes (2012) could be carried out.

[60] Though it should be highlighted here that the pre-pilot study was administered to a known group of test takers and appears to be targeting the correct level.

The evaluation inference in the IA relies on the warrant that the test is unidimensional (ie., that it is a test of listening) and does not contain any construct irrelevant variance. A PCAR analysis showed this to be the case and the fit statistics along with these results give evidence that the test has psychometric unidimensionality—that is, it tests one latent trait or construct (Sick, 2010). Construct-irrelevant variance was also investigated by one section of the feedback questionnaire. Here, the results showed that in general there were no problems with the test and its administration. However, the amount of time for reading the questions was highlighted as possibly being too short, although it was in fact lower ability students who reported this. I would therefore recommend that this issue be further investigated by specifically timing students with the correct level in order to assess task and instruction time adequacy. This would be easy to do and if more time was needed it would be easy to incorporate into the test. A further noteworthy comment is that participants believed the MCQ tasks to be the easiest, perhaps because they were more familiar with this test format. Certainly, example tests would have to be made available to ensure that all candidates are familiar with test content and format (EALTA, 2006). I would also add here that the results for R3 also provide support for this warrant. The participants did not use construct-irrelevant test wiseness strategies, which would have been a serious rebuttal of the evaluation inference had they been evidenced.

The generalisation inference is supported by the backing of the pre-pilot study along with the results to R1 and R4. The final test form contains items which show internal consistency. The scores are reliable, and therefore reproducible, and this is especially true because of the Rasch measurement model used in the study. The items fit the Rasch model well and suggest that we can be confident about the difficulty parameters produced by this analysis. The results suggest that a sufficient number of tasks are included on the test to reliably estimate the listening ability of candidates. The well-defined test specifications mean that comparable tests can be produced over administrations. A serious rebuttal of this inference would be that test forms are not parallel in difficulty. While such a question is beyond the scope of this study, as only one test form is being examined, in future we would need to make sure that appropriate equating and scaling procedures for test scores are used, which is easily achievable using Rasch measurement.

The explanation and extrapolation inference is based on the warrant that the test has good construct validity. Construct validity is important if we want a meaningful interpretation of the scores, unless scores are a reflective measure of the construct it would be impossible to generalise. The present study provides what Kane (2001) has called a 'strong form' of construct validity in that the test has been planned based on a theory of language use for oral comprehension—it is theory driven. Differences in test scores should be interpretable, a lower score should be representative of a lower ability and vice versa. The verbal reports give evidence that this is so and higher scorers understood more of the audio files. In this way, the evidence presented here follows the data-driven view suggested by Zumbo (2009). The warrant is also to some extent supported by results for the backing for previous warrants. For example, the results to R6 showed that expert judges believe the test to be representative of the CEFR B2 listening construct. Here the results from the verbal reports in R3 provide good evidence to show that candidates use the relevant KSAs to solve test items. Results showed certain variation in the level of processing necessary across the four tasks. This suggests that different listening skills were indeed being tested. In general, no construct-irrelevant test taking strategies were employed and candidates needed to understand the audio input in order to answer the items correctly. In this particular circumstance, as in many other QUAL investigations, care needs to be exercised because of the inevitable small samples used in QUAL designs (Dornyei, 2007). However, the questionnaire results, which provide some triangulation from a much larger sample, showed that participants believed they were using important metacognitive strategies in order to solve test items. Furthermore, using both self-assessments and reported accreditation exams as an alternate measure of the same skill, a correlation was found to exist. However, only a small number of participants reported such an accreditation and we cannot be sure that such measures can be relied upon. The accreditation may have been given years previously and participants may in fact have a much higher ability than accredited. Conversely, candidates may have a lower ability than their accreditation result if ability has been lost over time. Once again, stronger external validity evidence is necessary and a future study could consider the possibility of administering another validated test intended to measure the same construct and compare and correlate results.

The utilisation inference is largely supported by the statement of the problem (chapter 2), whereby the present *selectividad* exam has received years of academic criticism. This criticism is so strong—especially regards the lack of any listening component on the exam—that the introduction of a listening section would add to this inference whatever its form. Such an argument, however, is an oversimplification; I certainly would not argue that introducing a listening component with no construct definition or proper test development cycle—as proposed by some authors—would be sufficient.

In the context of the present study, educational reform proposed by the new LOMCE is based on the premise that the new assessments would result in positive outcomes for all students. Indeed, policy-makers the world over use tests to bring about changes in educational systems (Cheng, Sun, & Ma, 2015). Test consequences have been conceptualised as 'impact'—macro-level effects of a test on education and society—and 'washback'—the micro-level impact of a test on teaching and learning (McNamara, 2000). Positive washback would result from a test that accurately represents the aims of the curriculum (Wall, 2013). Indeed, Messick (1996) recommends that by minimising construct-irrelevant variance and construct under-representation during test design, then positive washback will be promoted. It can therefore be argued that the evidence to support all the previous links in the VA contribute to the final inference and provide evidence that positive washback will be achieved. As Kane (2013) argues, decisions based on accurate test score interpretations (something which is well supported by the present study) generally should induce intended positive outcomes.

A number of washback studies have indeed reported this to be the case. For example Andrews, Fullilove and Wong (2002) found that the introduction of a use of English speaking test for university admission in Hong Kong led to improvements in students speaking proficiency. Similarly, Hirai and Koizumi (2009) found a new speaking test in Japan had a positive impact on students motivation and learning. However, many washback studies have shown that the concept is more complex than this and new tests designed to promote positive change do not always have the desired outcome (Cheng, 2013). One of the major factors which has been reported to influence test washback is

that of the opinions and attitudes of teachers. Indeed, it is their attitudes and beliefs that will affect their teaching methodology rather than the design of the test itself (Watanabe, 1996). This is an important consideration and the challenge of implementing a new test may be difficult if a top-down approach is taken. Teachers—as well as other stakeholders—will need to be considered before a new test can be implemented.

Following the argument that stakeholders should be involved in the test development process (Fulcher & Davidson, 2007), this study also provides evidence to support the utilisation inference by canvassing the opinions of the test takers themselves—arguably the most important stakeholders. Here, it was reported that the topics were representative of those studied in the final two years of secondary school. However, there were mixed results concerning their beliefs that the present test was a fair measure of their ability, which seemed to depend on how well the participants had performed on the test. Also, two of the audios were not considered to be familiar; these two audios were the ones developed from prompts in order to get samples of authentic spoken discourse. Nevertheless, the participants also indicated that they strongly believed that they should have been given authentic listening practice at school.

It is important then to recognise that the utilisation inference is probably one of the weakest parts of the validity argument. The whole reasoning behind the introduction of the test is to lead to positive washback. While the participants do initially seem to be in agreement with this, future studies would be necessary to see if this was in fact the case. A large-scale baseline study is therefore recommended so that the introduction of a new test could be monitored and compared to this study in order to discover if positive washback in teaching and learning is actually achieved through the introduction of a new test.

The participants do not believe that they did enough listening in their language classroom, and they believe the reason for this is because there is no listening section on the *selectividad* exam. There was an overwhelming belief that listening is important for learning and should be included as part of the school leaving exam. If we listen to the

students themselves, then there certainly does not seem to be opposition to the introduction of a listening component. However, the caveat should be made here that the sample population was taken from a group of English language students who will probably be in favour of English language instruction. A more thorough study of student opinions would be possible if the sample included participants from a wider range of disciplines. Of course, there are many other stakeholders who should also be consulted, such as teachers, parents and educational policy makers.

Test use is an important consideration. Backing has been given to support the interpretation of the test score, i.e., that it is a valid and reliable representation of CEFR B2 listening proficiency. Value has been added to the score, which is therefore easily interpretable by university admissions officers. This is an important consideration as if the test were to be used for university entrance, it would be the university departments themselves which would have the responsibility of deciding and justifying the proficiency level which they deemed necessary for successful course completion.

Following the validity argument approach put forward by Cizek (2016), the justification of test use is something which would require a completely separate validity argument. Here, we must investigate many other factors, such as possible negative effects of introducing the test, a cost-benefit analysis, and an evaluation of alternate forms of testing methods.[61] Cizek argues that, in terms of validity, 'intended score meaning' and 'intended test use' cannot form part of the same argument—they cannot be combined. Following this conception of validity, the present study has indeed provided support for the intended score interpretation. Justification for test use for school leaving is to some extent supported by the argument concerning positive washback. Nevertheless, backing, which would definitively prove positive test consequences, has not yet been provided and this aspect would require further study.

---

[61] See for example Cheng (2013), who argues that better formative assessment systems conducted along with high stakes proficiency tests would lead to improved test washback by combining assessment for learning with assessment of learning.

Chapter 8. Conclusion

This project has developed a CEFR-related B2 listening exam for school leaving and university entrance. By following a validity argument approach, possible rebuttals have been identified and data taken from a number of sources have been investigated using a mix of statistical methods, as well as more qualitative techniques, in order to establish and support the inferences which can be made from test scores. The premise has been to provide a test of listening ability which is suitable for its purpose and has the aim of producing positive washback within the Spanish education system.

Here, the construct definition was of the upmost importance, as a clear definition of a construct to be measured is a fundamental first step in any validity argument in order to support the domain definition inference. Following a substantial literature review, the construct has been defined by following a process representation of listening ability. This definition acknowledges that listeners already understand the processing in their L1 and are able to transfer this understanding to listening in the L2 (Field, 2008a). The approach is well-founded in research and is based on the notion of 'expert listeners'. Such an approach has been recommended as a model for teaching the listening skill (Siegel, 2015), and consequently the test supports learning, as long as teachers have a sound understanding of the processes involved in listening. The fact that only authentic soundfiles have been used on the test, a practice which follows numerous calls in the

literature, is considered to be a major improvement on most listening tests. The original sound files provide a richer, more contextualised representation of the communicative event (Lynch, 2010) and thus provide good construct representation.

The results of the various parts of the study have provided backing to support the underlying assumptions in the IA in order to produce the final VA. The complete argument reaching the final inference, that the test scores can be used to make relevant decisions, has been largely supported by the use of Rasch measurement. As previously stated, a Rasch measurement framework helps to interpret the meaning of our measures, and a standardised score scale can be determined which can be applied to every version of the test. We can determine a location on this score scale which will represent our standard, CEFR B2 listening ability. Here, the cut score is the θ value on the Rasch scale which gives the probability of a correct response to the items for the minimally competent candidate. As such, relating *Basket Method* calibrations to Rasch item measures and using a *Bookmark Method* of standard setting on the Rasch logit scale makes these standard-setting methods useful, and the interpretation of test scores becomes a simpler endeavour. Furthermore, there is no need to repeat the standard-setting procedure each time a new version of the test is administered, as linking and equating techniques can be used to ensure that every test version is the same difficulty level.

However, validation is an ongoing process and any implementation of a new test must refine the validity argument. Future research should be drawn from any other potential rebuttals, and additional rebuttals could be voiced by any one of the stakeholders involved in the test. For example, any decision to administer the test through computers would entail a whole new set of possible rebuttals, which would have to be thoroughly investigated prior to implementation. Here, test consequences have already been highlighted as a weakness in the validity argument. Impact on teaching and learning should be investigated through washback studies (McNamara, 2006), and if an AUA is being followed (Bachman & Palmer, 2010), the starting point would necessarily have to be an examination of the beneficial consequences which we want to encourage.

The introduction of a new listening section—particularly one with supporting validation evidence—is both required by law and furthermore essential if new curriculum objectives drawing on the CEFR are to be covered. Clearly therefore, national tests must be CEFR-related (CoE, 2008), and following the experiences of other European countries, efforts to align school curriculums to the CEFR need to be accompanied by changes in evaluation procedures. It obviously follows that in the future, Spain cannot continue using the English section of the *selectividad* exam for school leaving and university entrance in any meaningful way.

At present, this necessity seems to be catered to through the use of external international exams—indeed, many school-leavers are entering university already in possession of such accreditations. Here, in the context of bi-lingual schools in Madrid, Griffiths (2017) has reported improved learning outcomes due to the implementation of external English proficiency exams which include a listening component. However, despite the obvious issues of fairness regarding the access to the resources necessary to obtain these accreditations, there exists the further issue of whether or not such international exams are indeed relevant for the present context. As Weir has put it:

> In comparing international tests with locally-developed ones, it would be wrong to assume that the former, even though developed by native speakers of English, are always superior [...] Global, multi-national, generic language tests taken by people around the world are unlikely to be particularly sensitive [...] to the needs of people within a particular society. In contrast, domestic tests can be more easily tailored to the local educational system and the needs of learners within a country.
>
> (Weir, 2013, cited in Wu, 2014)

As has already been discussed, in 2008 the European Commission set the objective that all citizens of the EU should achieve proficiency in two languages as well as their mother tongue. In order to provide a smooth transition from school to university or the

labour market, plurilingual policies need to be both instigated and supported throughout the course of the obligatory school system. Students should not only be provided with the opportunity to receive a CEFR-related accreditation, but also given the means by which to achieve the demands placed on them. Indeed, such plurilingual policies are becoming the norm, especially in the context of higher education. For example, the Ministry of Education, Culture and Sport published a document regarding the internationalisation of Spanish universities (2014), which has placed key priority on the internationalisation of higher education. Presently, English is seen as the language of higher education and many universities have introduced EMI and CLIL initiatives on their courses (Fortanet-Gómez, 2013). Such initiatives need to consider language proficiency entry levels, the pre-requisites for helping ensure success on these courses.

If the final baccalaureate test were to be used for university entrance in these particular circumstances, the present *selectividad* would not give admissions departments much information about the language competencies of prospective students. Again, we are reminded that test use is of upmost importance. Policy decisions need to be made about the language proficiency necessary to follow such courses. It has been reported that the threshold language level most associated with academic language proficiency in Europe is B2 (Degeyers et al., 2017). Indeed, Carlsen (2018) reported that students whose proficiency level was lower than B2 when entering university lacked those language skills needed for success on their course in a Norwegian context. Furthermore, there is still debate about whether a higher level might be more appropriate (Taylor & Geranpayeh, 2011). Certainly, if a new test were to be used for university admissions, it should be clear just what competencies are assessed, and a CEFR-related test can give test users a description of these competencies. Furthermore, listening might well be considered the most important competence for success on these courses (Vandergrift, 2004).

If we are indeed to encourage positive washback, then the new test construct and methodologies must be known and understood by both students and their teachers (Hughes, 2003). Here, teachers will play a central role, and will consequently need

extensive training in the new content and methodology. As Wang (2010, p. i) argues "it appears that for fundamental changes in teacher practice to occur, they must be accompanied by other changes in teachers' knowledge, beliefs, attitudes and thinking that inform such practice". I would even consider that teachers be involved in the test development process itself, as happened in Austria (see Froetscher, 2016). Such involvement develops teachers' ownership and understanding of the process of assessment (Harlen, 2005), and it becomes the teacher who shapes students' perception of the test, and who thus becomes the driving force for change, rather than the assessment process itself (Cheng et al*.,* 2015).

For listening, teachers will need to understand and teach both top-down and bottom-up processes and strategies (Field, 2008a; Richards, 2008). As well as teacher training, appropriate teaching and testing materials will need to be made freely available. These should be representative of the types of materials which appear on the test and should include unscripted, authentic spoken recordings "to expose the listener to the natural cadences of the target language and to train the learner in the unfamiliar process of extrapolating meaning from a piece of speech that may only be partly understood" (Field, 2008a, p.277).

In terms of the way forward, then, it needs to be recognised that the implementation of the BFE test is not a simple endeavour and will require much time and effort. It may be that initially a small sample of participating schools could take part. In this way, a baseline study could first be carried out and the consequential validity of the new test could be investigated before it was implemented on a larger scale. After all, as Wall (1996, p.334) has pointed out, the use of high-stakes tests as a means to "introduce change in the classroom are often not as effective as their designers hoped they would be". Furthermore, criticisms about centrally-driven reforms have shown that they may well suffer from a range of unintended consequences (Qi, 2007).

If the social and academic goals are to be achieved, it is up to policy makers to ensure that the new educational reforms are implemented in the correct way. Such educational

reforms increasingly rely on the introduction of new assessment procedures in order to improve the quality of education (Chalhoub-Deville, 2016). Here, broader policy issues of testing which do not concern the qualities of a test should also be considered (Bachman, 2005). McNamara (2005) has argued that all language testing is politically motivated, with tests being used to achieve certain political ends (Fulcher, 2009). In Spain, the new laws are indeed an attempt to comply with European directives and any resulting policy decisions subsequently need to be well informed.

Assessment is seen as the solution to a problem, one which, if implemented correctly, can directly support learning outcomes for students. However, a top-down approach which comes directly from policy makers with no real understanding of how new policies are to be implemented is certainly not a solution. Rather, a whole host of considerations need to be taken into account, including the design, administration and marking of any centrally-administered exams. It has been argued that "the most important element of any reform project are the individuals and their ambitions, personal agendas, openness to change and attitudes to professionalism" (Pižorn & Nagy, 2009, p.185). Such a context requires the creation of an expert group to oversee the implementation of new reforms, which should be implemented over time and include dialogue with teachers, students, parents, university admissions and any other relevant stakeholders. If correctly implemented, a new assessment system could bring positive change, leading to enhanced equity and fairness as well as providing statistics for the analysis of local and national achievement.

In Spain, results from the European Survey on Language Competences (2012) have shown that present school-based learning has failed to equip students with the real world language competencies necessary for success in the plurilingual world we live in. In response to this deficiency, the present thesis has demonstrated the necessary steps to be taken in order to develop a valid test of oral comprehension. The main purpose of such a test is both to promote learning and bring about a shift in language pedagogy—from knowledge-based to more communicative practices—and to validly interpret what has been learned.

The present thesis has, in addition, detailed the benefits of a nationally-standardised, CEFR-related language proficiency test, arguing that it is both desirable and necessary if we are to see a general improvement in the implementation of the new LOMCE national curriculum for English. It has furthermore attempted to demonstrate to the relevant bodies how such a project should be developed in accordance with a validity argument approach. Were such a test to be implemented, it would be further necessary to provide equally detailed studies of the constructs of the other language use abilities to be tested (reading, writing and speaking). These studies would ideally require the setting up of an expert body to standardise and oversee the test development cycle and orientate effective washback within the education system through the inclusion of all relevant stakeholders. Such a development process is of paramount importance if Spain is to foster the necessary foreign language skills to allow it to play a more significant role on the European stage.

RESUMEN EN ESPAÑOL

INTRODUCCIÓN

Desde su publicación, el *Marco Común Europeo de Referencia para las lenguas* (Council of Europe, 2001; en adelante, MCER) se ha convertido en la guía de referencia por excelencia para la enseñanza y aprendizaje de lenguas modernas en Europa. Su meta es fomentar la trasparencia y la coherencia en la educación proporcionando una base común para los currículos de aprendizaje de idiomas, con el fin de normalizar la comparación de estándares en los distintos países de la comunidad europea, es decir, la implantación de un marco conceptual compartido. A raíz de la política de plurilingüismo promocionada por la UE, los distintos gobiernos y departamentos dentro de la comunidad se han visto obligados a tomar en cuenta el MCER, compromiso que ha estimulado una plétora de iniciativas educativas por parte de los agentes políticos, quienes se han esforzado por incorporar a los sistemas educativos públicos un modelo de la enseñanza de las lenguas basado en las competencias lingüísticas (Lim, 2013). Como consecuencia de ello, la evaluación de estos objetivos supone actualmente un componente esencial del marco y muchos países europeos han llevado a cabo reformas de sus exámenes finales de salida en la enseñanza secundaria para así evidenciar buenas prácticas en la evaluación de lenguas. Dichos cambios podrían considerarse especialmente apremiantes en cuanto a la enseñanza y evaluación del inglés como segunda lengua, debido a su estatus como una de las lenguas instrumentales más relevantes dentro de la UE, además de su rápida y continuada extensión como la lengua más utilizada en la actualidad para la comunicación internacional (Crystal, 2012). En efecto, su estatus como *lingua franca* popular ha dado paso a la implantación de distintas políticas, tanto nacionales como internacionales, con el fin de mejorar las capacidades lingüísticas de estudiantes de inglés como segunda lengua.

En España, no obstante, la influencia del MCER es algo bastante reciente; a pesar de que actualmente la mejora de la calidad de la enseñanza de idiomas sea claramente un objetivo fundamental, los resultados de la Encuesta europea de competencias lingüísticas (Comisión Europea, 2012)—que comparó las competencias de alumnos de la Educación Secundaria Obligatoria en los distintos países de la Unión Europea—manifiestan que los usuarios españoles de inglés como segunda lengua han quedado algo a la zaga de sus conciudadanos europeos respecto a sus capacidades lingüísticas. La versión española de los resultados (INEE, 2013) sitúa a España en décimo lugar entre los catorce países europeos entrevistados, quedando destacados los malos resultados obtenidos en la destreza de comprensión oral. No obstante, la competencia lingüística resulta de especial interés para los agentes de las políticas educativas a nivel europeo y la Comisión Europea ha precisado objetivos de referencia a sus estados miembros para el año 2020, estipulando que al menos el 50 % de alumnos deberán poseer un nivel B1 o superior del MCER en una segunda lengua europea a los 15 años de edad (i. e., antes de empezar el bachillerato). En el mejor de los escenarios, una promoción de alumnos de bachillerato saldría con al menos un B2, tal como ocurre en el caso de la mayoría de los otros países europeos (véase, por ejemplo, Deygers & Zeidler, 2015; Lim, 2013).

Todo esto no quiere decir que no se estén realizando esfuerzos a nivel español por mejorar la situación en la actualidad. De hecho, se acaba de implantar varias reformas educativas importantes en el recientemente actualizado currículo nacional español, la *Ley Orgánica para la Mejora de la Calidad Educativa* (LOMCE), introducida por el Ministerio de Educación, Cultura y Deporte (*Boletín Oficial Del Estado (BOE),* 2013; en adelante, MECD). Con el fin de mejorar el aprendizaje, se ha elaborado el nuevo programa de estudios para inglés con base en un currículo enfocado en las competencias y que incluye propuestas para una nueva prueba de acceso a la universidad. Esta evaluación externa se había programado para introducirse en el curso 2017/18, con la inclusión de un nuevo componente oral de comprensión y producción. No obstante, el MECD no solo no ha conseguido poner esta reforma en marcha, sino que su falta de claridad y dirección en cuanto a la descripción de la forma y contenido de la prueba justifica en gran parte la reticencia de los gobiernos regionales a implantarla y explica el

posterior aplazamiento de los nuevos procedimientos del examen final de bachillerato (en adelante, el EFB). En verdad, parece haber poco interés por parte de los gobiernos regionales y nacionales en mejorar los distintos exámenes actualmente vigentes, a pesar de que estas anticuadas pruebas proporcionen muy poca información sobre las competencias de los alumnos (Amengual Pizarro, 2005, 2006; García Laborda, 2010, 2012; Sanz Sainz, & Fernández Álvarez, 2005) y que su sustitución sea un requisito por ley. Por tanto, el futuro de la prueba ha quedado algo incierto y evidentemente va a tener que haber más diálogo si vamos a percibir avances al respecto.

La falta de un componente de comprensión oral en una prueba tan decisiva como la de la selectividad actual es particularmente preocupante, ya que claramente constituye no solo un componente esencial en la competencia comunicativa, sino uno de los factores que más contribuyen a la exitosa adquisición de un idioma extranjero (Rubin, 1994; Vandergrift, 1999; Zhang, 2012). La inclusión de actividades de compresión oral en el aula de lenguas aumenta el contacto con materiales comprensibles para los alumnos, labor fundamental si entendemos que la audición supone casi el 50 % de todos los actos comunicativos de los usuarios de lengua adultos (Miller, 2003). Efectivamente, actividades como el acto de escuchar las explicaciones del profesor o las preguntas de otros compañeros son actividades básicas en el aula y por tanto la buena comprensión oral ha sido necesariamente vinculada con el éxito académico (Jeon, 2007). Si se pretende utilizar un título final de bachillerato como evidencia para la acceso a la universidad, entonces estas actividades resultan de clara importancia, sobre todo para aquellos cursos que se imparten con el inglés como medio de instrucción (en inglés, *English as a method of instruction*, en adelante *EMI*).

Debido a la clara importancia de la destreza de comprensión oral para las futuras actuaciones lingüísticas de alumnos, esta tesis doctoral abogará por su inclusión en la prueba final de bachillerato, además de instar a que dicha prueba mida de forma válida y fiable los niveles de dominio de los alumnos en relación con el MCER. Tanto las plazas universitarias como los fondos económicos de los departamentos son de un carácter limitado y las universidades tienen la responsabilidad de distribuir a los alumnos

equitativamente si pretendemos fomentar la igualdad de oportunidades en la educación pública. En consecuencia, es de esperar que esta tesis realice una aportación oportuna al debate mediante un riguroso estudio de validez sobre una prueba de comprensión oral para bachillerato, ideada para ser administrada en centros participantes. Concluiré que el proyecto debe realizarse mediante el desarrollo de un examen final basado en las competencias y vinculado con el MCER (CoE, 2001); por tanto, mi propuesta de investigación tratará el desarrollo de la parte de comprensión oral de una prueba vinculada con el MCER a nivel B2. El constructo para esta prueba partirá de un constructo que seguirá internacionalmente reconocidos conceptos de validez, fiabilidad, calidad e imparcialidad con el fin de crear una prueba que no solo sea relevante a su contexto de uso, sino que además tenga el objetivo de generar un efecto rebote positivo a largo plazo en la enseñanza del inglés en España.

Tras haber presentado mis principales motivaciones del presente estudio, a continuación procederé a describir las demás partes de la tesis, la cual está dividida en otros siete capítulos.

CAPÍTULO 2

Tras la introducción del Capítulo 1, el Capítulo 2 proporciona una declaración detallada del contexto actual y del problema, y ofrece razones para la necesidad de desarrollar una nueva prueba de comprensión oral vinculada con el MCER.

El Capítulo 2 comienza con un examen de la situación de las competencias lingüísticas—y en especial, las de inglés como segunda lengua—en el contexto europeo actual, en el cual intentaré situar a los estudiantes de lengua españoles. Tras exponer los resultados deficientes obtenidos por los alumnos españoles en el Estudio Europeo de Competencia Lingüística (EECL, 2013), procedo a examinar varias de las reformas que otros países europeos han realizado en evaluación para así intentar identificar algunas de las lecciones aprendidas y con la esperanza de que también sirvan para orientar sobre el contexto español.

A continuación, examino la actual Prueba de Acceso a la Universidad (en adelante, la PAU) que se emplea desde hace más de veinte años, a pesar de las continuadas críticas desfavorables recibidas desde la comunidad académica. Entre las críticas principales figura el hecho de que el examen no se atiene a estándares internacionales para la producción de exámenes de alto impacto, debido en gran parte a que no suele contar con la colaboración de expertos. El examen actual no evalúa la competencia lingüística comunicativa y por tanto no tiene validez de constructo. En efecto, González-Such, Jornet y Bakieva (2013) concluyeron que, en su forma actual, carece de definiciones de constructo y no puede considerarse ni justo ni ético. Como consecuencia, ha habido llamamientos a que se establezca un organismo de expertos a nivel nacional para que este supervise la implantación de un nuevo examen (véase p.ej., Fernández Álvarez, 2007; López Navas, 2012, 2015).

La principal crítica de la PAU es sin duda la percepción extendida de su efecto negativo en la práctica educativa en España debido a la falta de una prueba de oral. De hecho, se ha demostrado que el actual currículo impartido a menudo sufre limitaciones debido a que muchos de los profesores están, en efecto, enseñando para el examen. Fernández Álvarez (2007) observa que el 75 % de los profesores que completaron un cuestionario sobre la PAU se sentían presionados por parte de sus alumnos para que les prepararan para el examen. Por tanto, no resulta nada sorprendente que los profesores opinen que deben basar su práctica didáctica en métodos más tradicionales. García Laborda y Fernández Álvarez (2011) también notaron que, a pesar de que los profesores sí tienen interés en desarrollar componentes de comprensión y producción orales dentro del aula, las mayoría de las clases actuales suelen limitarse a prácticas de traducción y de actividades gramaticales, debido a su inclusión en la PAU. En gran medida, los profesores empleaban ejercicios no auténticos sacados de manuales de texto, con más del 75 % dedicando menos del 10 % de su tiempo en el aula a actividades de comprensión oral. Dichas afirmaciones son una clara muestra de que la correcta implantación de un efecto rebote en el aula tendría buenas posibilidades de mejorar la calidad de la enseñanza de idiomas en España, país en el que queda más que demostrado que el uso de tiempo en el aula se destina a aquellos aspectos que van a entrar en el examen.

En respuesta a dichas críticas, el gobierno ha realizado algunas tentativas de rectificación de la situación mediante la implantación de reformas educativas en los últimos años. A continuación en el capítulo, doy un breve resumen de estos decretos, resumen que termina concluyendo que a pesar de haberse aprobado una plétora de decretos nuevos, el gobierno todavía no ha logrado evidenciar ningún cambio palpable, bien en el contenido, bien en la administración del nuevo examen. Al publicarse, la LOMCE exponía claramente que la introducción de un componente de comprensión oral sería programado para el curso de 2017 y con vistas de que se hiciera obligatorio para el acceso a la universidad en el 2018. No obstante, a día de hoy todavía existe un buen grado de oposición a la implantación de esta ley y parece claro que las pretendidas reformas no se implantarán en un futuro próximo.

A continuación, facilito un breve resumen del MCER, cuya filosofía de competencia comunicativa está ahora insertada en el currículo nacional, argumentando que el único paso todavía necesario es que España implante un sistema de evaluación vinculado con el MCER similar a los ya existentes en otros países europeos dentro de la comunidad y sosteniendo que dicho paso no solo le conferirá a España un mayor papel dentro del panorama europeo, sino que también servirá para que cumpla con sus propios objetivos curriculares. Propongo el desarrollo de una prueba vinculada al MCER que no solo tenga un constructo bien formulado, sino que además cubra las cuatro macro destrezas de comprensión y producción orales y escritas a nivel B2. El nivel B2 ha sido elegido debido a que sus descriptores son los que mejor se ajustan al actual contenido curricular requerido, además del hecho de que muchos lo consideran el más apropiado tanto para los estudios académicos como para la inserción en el mercado laboral y, por tanto, es el más exigido por la mayoría de las demás universidades europeas.

CAPÍTULO 3

Tras haber perfilado la situación presente en España y haber documentado la actual inexistencia de verdaderas tentativas de desarrollar un componente de comprensión oral válido, el objetivo de esta tesis doctoral será la resolución de las carencias previamente

expuestas mediante el desarrollo de una prueba de comprensión oral para una nueva Prueba Final de Bachillerato (en adelante, PFB) a nivel B2 del MCER, el cual se debe basar en una clara definición del constructo. A tal fin, los objetivos de la revisión bibliográfica son dos. En primer lugar, trato cuestiones relacionadas con la definición de la capacidad lingüística—haciendo especial hincapié en el MCER—, para luego presentar una definición del constructo de la destreza de comprensión oral. En segundo lugar, hago resumen de las ideas actuales sobre la validación de pruebas de lengua para poder establecer así un marco coherente desde el cual se podrá evaluar la propuesta prueba de comprensión oral a nivel B2 para la nueva PFB.

Con el fin de poder entender mejor los complejos conceptos relacionados con el tema del discurso en contextos auténticos, doy a continuación una descripción exhaustiva de la literatura en cuestión. Para que quede manifiesta la validez de cualquier prueba, esta tiene que demostrar que en verdad activa los debidos procesos cognitivos que pretende evaluar. Para una prueba vinculada al MCER, cualquier noción de la capacidad lingüística de comprensión oral estará fundamentada en el modelo de la competencia comunicativa. Con el objetivo de proporcionar un modelo de capacidad de comprensión oral basado en la teoría sustantiva, se tendrán que emplear los conocimientos, destrezas y capacidades detallados en esta sección, los cuales están basados en gran parte en el modelo de procesos para la comprensión oral de Field (2008a, 2013a)

Además, y de acuerdo con perspectivas interaccionistas sobre la capacidad lingüística, cualquier definición del constructo debe tomar en cuenta el ámbito de uso de la tarea. Aquí, se debe ajustar nuestro modelo de capacidad lingüística al contexto de uso, ya que solo la introducción de tareas que tengan un auténtico contexto interactivo nos permitirá demostrar que se reproduzcan procesos cognitivos parecidos a los del ámbito de uso de la lengua de destino (en inglés, *Target language use*, o *TLU*). Por tanto, trataremos aquellos debates relacionados con los parámetros contextuales de las tareas de comprensión oral. Aquí examinaremos la literatura en lo que concierne a las variables necesariamente relacionadas con la fuente de entrada (como son su autenticidad, la elección de acento/uso de inglés cómo *lingua franca*, el canal, o la complejidad

lingüística) y las condiciones de la tarea (la cantidad de reproducciones permitidas, la vista preliminar de los ítems, o el formato de respuesta), para así orientar el desarrollo de las especificaciones para la prueba de comprensión oral PFB.

La validez y la validación suponen aspectos claves al inicio de cualquier proyecto de desarrollo que pretenda construir un instrumento de evaluación apropiado desde el cual se pueda proceder a usar la puntuación obtenida por el candidato para tomar decisiones pertinentes. De las muchas y variadas tentativas de proporcionar marcos de evaluación para pruebas de lengua, actualmente se favorece un enfoque desde la lógica argumentativa (véase p.ej., Kane, 1992, 2001, 2002, 2004, 2012, 2013; Kane, Crooks & Cohen, 1999). La opinión actual comprende la validez como concepto unitario, es decir, como la provisión de un argumento integrado y unificado que justifica los varios usos, interpretaciones, decisiones y consecuencias que surgen de cualquier evaluación. Es de fundamental importancia que se incorpore evidencia de validez como elemento integrado de un argumento coherente para así justificar los consiguientes usos de cualquier prueba desarrollada, y aquí el enfoque desde la lógica argumentativa para la validación de pruebas proporciona un marco tanto lúcido como equilibrado (Chapelle, 2012).

En vista de lo anteriormente mencionado, procedo a continuación a dar un bosquejo del enfoque desde la lógica argumentativa de Kane. Dicho enfoque comprende dos etapas. La primera, conocida como "la etapa de desarrollo" (Kane, 2006), requiere el diseño del *Argumento interpretativo* (AI), el cual incluye especificaciones detalladas de cualquier afirmación sobre las puntuaciones de la prueba y los objetivos deseados para poder justificar el uso de la prueba; la segunda, o "etapa de evaluación, supone la evaluación de la plausibilidad general de las interpretaciones y usos propuestos" (Kane, 2012, p.4). Este es el *Argumento de Validez* (AV) para la prueba y debe especificar el significado de la puntuación y justificar el marco teórico que subyace a la prueba. El marco incluye una serie de inferencias de validez que en su esencia tratan de conclusiones justificadas por evidencia acumulada sobre cada aspecto de la validez de la prueba. En términos del ámbito de interés, una prueba solo puede considerarse válida si proporcionamos suficiente evidencia de cada uno de los inferencias del AI. Por

consiguiente, también presento una discusión de cada inferencia según el enfoque desde la lógica argumentativa, junto con ejemplos prácticos de este tipo de enfoque para la validación de las pruebas de lengua. En este sentido, se considera que tanto el ámbito de uso de la lengua de destino como las consecuencias de la prueba juegan un papel importante. También describo el enfoque de Toulmin sobre el razonamiento lógico (1958/2003), el cual proporciona un marco conveniente y un vocabulario bien establecido desde los cuales se puede proceder a discutir distintos argumentos interpretativos (Kane, 2004). Por tanto, ha sido empleado por la mayoría de los defensores del enfoque en el argumento de la validez (p.ej., Kane, Bachman, Bachman y Palmer, Chapelle et al., Mislevy y colaboradores, entre otros) y será empleado en el presente estudio para el desarrollo de las preguntas de la investigación.

CAPÍTULO 4

El Capítulo 4 procede a desarrollar las especificaciones para la prueba planeada, que actúa como prototipo de la prueba (Alderson et al., 1995) me permitió poner en funcionamiento el constructo de comprensión oral de la PFB, y de ese modo producir una versión de la prueba que funcionara como instrumento de medición.

Una vez desarrollada la prueba, se procedió a llevar a cabo un estudio de pilotaje preliminar, el cual me permitió realizar un análisis inicial de los ítems, además de pilotar la metodología que pretendía emplear con los informes verbales en el estudio principal. Los resultados de este estudio preliminar condujeron a que se realizaran algunos cambios en la forma final de la prueba como parte de los preparativos para el estudio principal. Después de estas modificaciones, se produjo una prueba de 33 ítems, la cual, en principio, alcanza el nivel correcto respecto a su dificultad y a las especificaciones de la prueba.

Un enfoque AV fue elegido para guiar el estudio y a este fin se procedió a definir el AI y a desarrollar las preguntas de la investigación mediante el empleo de un enfoque de razonamiento lógico de Toulmin. De este modo, se podrá determinar qué tipos de

evidencia se tiene que recopilar para presentar la propuesta final del argumento de validez. Se presenta aquí el argumento inicial para la PFB, respaldado por la teoría sustantiva descrita en la revisión bibliográfica y el estudio de pilotaje preliminar y se plantean posibles refutaciones como una serie de preguntas del tipo "¿Qué sucede si…?" para cada una de las inferencias del argumento interpretativo. Posteriormente, estas refutaciones se convertirán en mis preguntas de la investigación y por tanto guiarán la investigación necesaria para definir y presentar el argumento de validez de la PFB. Concluyo el capítulo con la presentación de mis preguntas de la investigación según el orden en el cual son tratadas:

Pregunta 1 (P1): ¿Cuáles son las propiedades estadísticas de la prueba?

Pregunta 2 (P2): ¿La prueba resulta unidimensional? ¿Existe alguna varianza irrelevante en el constructo dentro de las puntuaciones de la prueba?

Pregunta 3 (P3): ¿Emplean los candidatos los conocimientos, destrezas y capacidades relevantes para resolver los ítems de la PFB?

Pregunta 4 (P4): ¿Son fiables las puntuaciones de la forma final de la prueba?

Pregunta 5 (P5): ¿Qué opinan los candidatos sobre la prueba de comprensión oral de la PFB? ¿Creen los candidatos que la prueba fomentará un efecto rebote positivo?

Pregunta 6 (P6): ¿Opinan los jueces expertos que las tareas de la prueba representan fielmente el constructo de comprensión oral del MCER?

Pregunta 7 (P7): ¿Cuál debe ser el punto de corte en la prueba para que se proporcione una evaluación exacta de candidatos a nivel B2? (¿Será posible la elaboración de formas de la prueba en paralelo?)

Pregunta 8 (P8): ¿Existe una correlación entre las puntuaciones de la prueba y medidas equivalentes del mismo constructo?

CAPÍTULO 5

Este capitulo presenta una perspectiva general del diseño de la investigación, además de una discusión de las metodologías de investigación empleadas en el estudio, antes de proceder a resumir los procedimientos de la recopilación y del análisis de los datos. Tanto en la recopilación de los datos como en su almacenamiento, se observaron todas las directrices sobre buenas prácticas y la ética en la investigación exigidos por la Universidad de Granada.

Los datos han sido recopilados de las siguientes fuentes:

1. Puntuaciones obtenidas en la prueba
2. Cuestionarios
3. Informes verbales
4. Juicios expertos

Las puntuaciones de prueba fueron analizadas mediante el uso de la teoría clásica de los tests (TCT) en primer lugar. No obstante, y debido a algunas limitaciones de este tipo de metodología para las pruebas referidas a criterios, también se procedió a aplicarles un análisis de teoría de respuesta al ítem (TRI) mediante el análisis Rasch. El modelo Rasch (George Rasch, 1960), modelo logístico de un parámetro, es a menudo el modelo de preferencia en la evaluación de las lenguas. Es un modelo de probabilidades que en el nivel más básico es dicótomo, es decir, o correcto o incorrecto. Sick (2008) destaca que la gran diferencia entre la TCT y Rasch es que la TCT es un modelo descriptivo, y por tanto dependiente de la población de la prueba en cuestión, mientras que Rasch es probabilístico y inferencial. Así, el modelo Rasch genera estimaciones tanto de las actuaciones de las personas como de la dificultad los ítems, y por tanto puede

predecir la actuación de cualquier candidato en una determinada habilidad o ante cualquier ítem de una determinada dificultad en una prueba.

Se conceptualiza a los candidatos como más o menos hábiles, ordenándolos según un continuo de habilidad de comprensión oral en el cual se requiere más o menos habilidad para solucionar los ítems de la prueba. Los ítems funcionan juntos para definir el continuo de habilidad de comprensión oral, que se expresa no como puntuación bruta sino como medida linear. Se puede proporcionar información sobre la unidimensionalidad a través de las estadísticas de *fit* (Sick, 2010), además de la investigación posterior de los residuos de Rasch. Linacre (2017a) aconseja un análisis factorial de los componentes principles de los residuos de Rasch (en inglés, *Principle components factor analysis of Rasch residuals*, o PCAR) como método principal para identificar la presencia de multidimensionalidad en los datos. Este análisis identifica diferencias estructurales entre constructos opuestos y ayuda a identificar cualquier componente secundario que podría haber sido medido por la prueba. Esta metodología se utilizó para contestar la Pregunta de investigación 2 (P2) para descubrir si la prueba en verdad resulta ser una prueba de comprensión oral unidimensional.

El cuestionario fue desarrollado para, además de recopilar datos biográficos de los candidatos, cubrir las siguientes cuestiones, consideradas las más relevantes para el presente estudio:

1. Autoevaluación de los candidatos.
2. Opiniones sobre la prueba.
3. Opiniones sobre el uso de procesos y estrategias.

Se le administró la prueba de 33 ítems a una muestra grande (N=153) de universitarios del primer curso de grado en Filología inglesa de la UGR. La prueba fue administrada en octubre del 2016. En consecuencia, solo hacía tres

meses que la mayoría de los participantes habían hecho la prueba de selectividad para el acceso a la universidad y, por tanto, se puede considerar que eran representativos de la población meta. El hecho de que los universitarios habían elegido cursar estudios de filología inglesa se consideró indicativo de que estaban interesados en la materia y que deberían poseer un nivel de capacidad aproximado al B2 del MCER en inglés. Justo al terminar la prueba, se les pidió a los candidatos que completaron el cuestionario, que incluía una pregunta de tipo abierto para que pudieran hacer cualquier comentario que les pareciera pertinente. Los datos categóricos recompilados de los cuestionarios fueron codificados y presentados en principio como datos estadísticos descriptivos en el programa SPSS. A estos, se le añadió un análisis de correlación, además de llevar a cabo una comparación de las respuestas a los ítems del cuestionario y la puntuación media obtenida en la prueba siempre que se considerara apropiado.

El proceso de desarrollo de pruebas de evaluación de lenguas puede beneficiarse tanto de métodos como los "pensar en voz alta" (en inglés, *think alouds*) o informes verbales para comprender mejor los procesos cognitivos activados en el candidato durante la realización de una prueba; en concreto, estos dos métodos gozan de un uso extendido en investigaciones sobre las estrategias de presentación de exámenes (Cohen, 2014). En investigaciones sobre la validación de pruebas, estas metodologías resultan muy útiles para contestar a preguntas como "¿Activa la prueba aquellas capacidades que pretende evaluar?" (Xi, 2008, p.186), que también es una de las preguntas de investigación del presente estudio.

No obstante, la comprensión oral es un proceso conectado e *in vivo*, lo cual dificulta la investigación (Vandergrift, 2007). Dada la imposibilidad de recopilar información de los *think alouds* durante la realización de la prueba, hay que recurrir a métodos de carácter retroactivo, como el del recuerdo estimulado (en inglés, *stimulated recall*, Gass & Mackey, 2000). Este método pretende que la administración de pies orales y apuntes—como son el cuadernillo de preguntas, la hoja de respuestas o los apuntes—, proporcionados al participante mientras

finaliza sus respuestas, facilite la recopilación de pensamientos y procesos todavía presentes en la memoria a corto plazo. Este proceso de recordatorio inmediato se graba y a continuación la grabación es sometida a un largo y costoso proceso de trascripción, segmentación y codificación. El esquema de codificación puede desarrollarse de varias formas. Por ejemplo, siguiendo el enfoque inductivo de la Teoría Fundamentada (en inglés, *grounded theory*, véase Dörnyei, 2007), las pautas a seguir salen de los propios datos. En cambio, el enfoque empleado en el presente estudio sigue el de Gu (2014) y desarrolla los códigos con base en un marco teórico.

Debido a las dificultades prácticas que suponen las grandes cantidades de tiempo exigidas por este tipo de metodología, solo se procesó una muestra muy reducida, donde cada sesión para el presente estudio duró aproximadamente 75 minutos. También, se siguió la recomendación de Dörnyei (2007) de emplear un muestreo "intencionado" y los siete participantes eran voluntarios de mis propias clases en el CLM, quienes a mi juicio poseían un nivel de capacidad de B2 en la comprensión oral. Una vez recogidos los informes, fueron trascritos e introducidos en el programa de análisis *QDA minor lite* para ser codificados. Además, se presentan aquí ejemplos relevantes del análisis cualitativo de los informes concurrentes durante el periodo de planificación preliminar antes de que hicieran la prueba para dar respaldo a las conclusiones obtenidas.

Con el fin de proporcionar evidencia de los estándares de contenidos y de vinculación con el MCER, es necesario contar con las opiniones de expertos que segunden cualquier juicio sobre los contenidos de la prueba. En este sentido, el Consejo de Europa publicó un documento guía, *The Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Council of Europe, 2009, en español, *Manual para relacionar exámenes con el MCER del Consejo de Europa*, en adelante el *Manual*), el cual establece unas directrices muy detalladas sobre el proceso para ajustar las pruebas de lengua al MCER. El presente estudio sigue las

pautas recomendadas en el *Manual* y emplea dos métodos de establecimiento de normas. Replicando el protocolo establecido en Shackleton (en preparación), se empleó una combinación de 1) *the Basket Method* (en español "el método de la cesta") y 2) *the Bookmark Method* (en español "el método del marcador"), entre otras razones porque los ocho jueces que participaban ya habían recibido formación en estos dos métodos y estaban familiarizados con su empleo. Con el fin de contestar a la pregunta **P6** el *Basket method* se empleó para especificar los contenidos de la prueba y evidenciar que los jueces expertos opinaban que sí eran representativos del nivel B2 del MCER. Estos datos se analizaron mediante un modelo de medición de muchas facetas de Rasch (*Many-Facet Rasch measurement model* en inglés, en adelante *MFRM*) en el programa FACETS (Linacre, 2017b). Al igual que en el estudio previo, los puntos de cortes se establecieron mediante el *Bookmark Method* (Mitzel, Lewis, Patz, & Green, 2001), el cual emplea parámetros de dificultad Rasch para producir un *Ordered Item Booklet* (en español "cuadernillo de ítems ordenados", en adelante *OIB*). El Capítulo 5 concluye presentando otros aspectos de la prueba que respaldan su validez interna, externa y de procedimiento.

CAPÍTULO 6

Este capítulo presenta y analiza los resultados del estudio en lo concerniente a mis preguntas de la investigación.

En primer lugar, se llevó a cabo un análisis TCT de las puntuaciones de la prueba, seguido por otro más detallado a través de la medición Rasch. Este análisis, junto con un análisis de distractores, la examinación de las curvas ICC individuales y la examinación del mapa de variables de los ítems, confirmó que se debería eliminar cuatro ítems de la prueba (los ítems 1.3F, 2.8C, 4.10 y 4.8). Aunque por lo general los demás ítems parecían funcionar bien, los análisis también señalaron los ítems 3.3 y 4.6 como posibles candidatos para ser

eliminados y después de un análisis posterior de los informes verbales, finalmente se decidió también eliminar el ítem 3.3.

Para contestar a la Pregunta 2 (P2) y proporcionar evidencia que respaldara la inferencia de evaluación, se presentan los resultados de un análisis *PCAR*, además de ejemplos relevantes procedentes del cuestionario. Los resultados *PCAR* demuestran que la prueba es unidimensional y confirma que la propiedad de independencia local es válida. El hecho de que la prueba sí se muestre unidimensional respalda la validez del constructo y sugiere que no se está evaluando ningún otro constructo que no sea el de la capacidad de comprensión oral. Por tanto, se puede considerar que dicha evidencia demuestra la ausencia de varianza irrelevante del constructo.

Otro tipo de evidencia recopilado en lo que se refiere a la varianza irrelevante del constructo se derivó de los propios candidatos mediante el análisis de los datos del cuestionario que pertenecían a las opiniones de los candidatos sobre el contenido y la administración de la prueba. Por lo general, los candidatos no consideraban que la prueba contuviera varianza irrelevante; aquellos candidatos que sí mostraban percepciones negativas eran los de menor habilidad, posiblemente como respuesta a su peor actuación en la prueba.

Los datos verbales recopilados de los *think alouds* concurrentes fueron analizados de forma cualitativa y los informes concurrentes fueron introducidos en el programa *QDA Minor Lite* y codificado de acuerdo con los temas emergentes observados en las mismas entrevistas. En este sentido se presenta una variedad de ejemplos donde se observó que los participantes demostraban que utilizaban estrategias metacognitivas al activar sus esquemas de conocimiento previo y usar el contexto de la situación para ayudarse a predecir el audio. Del mismo modo, se siguió este proceso también con los informes retrospectivos; estos resultados se presentan primero como resultados cuantitativos (como el nivel de procesamiento empleado por cada participante para poder contestar a las

preguntas) y luego también en un análisis más exhaustivo. Los resultados muestran que los candidatos sí demostraban que usaban los conocimientos y capacidades planteados por el modelo de comprensión de la PFB para solucionar los ítems. Por consiguiente, esta parte del estudio ha proporcionado buenas pruebas de la validez del constructo que respaldan tanto la inferencia de explicación como la inferencia de extrapolación del argumento de validez general.

A continuación, se llevó a cabo un segundo análisis de la forma final de la prueba de 28 ítems. El análisis Rasch de los ítems por separado muestra que los ítems tienen un *infit MNSQ* entre los valores aceptables de 0.84 y 1.21, lo que confirma que todos los ítems presentes en la prueba se muestran productivos para la medición. En definitiva, el análisis Rasch muestra que la prueba de 28 ítems resulta más fiable que la original de 32 ítems. Todos los ítems se ajustan al modelo Rasch y la muestra utilizada en el estudio ha sido lo suficientemente grande como para proporcionar estimaciones precisas sobre la dificultad de los ítems. Por lo tanto, podemos concluir que se puede comparar a los candidatos según la escala de medición y que la obtención de una mayor puntuación en la prueba en verdad corresponde a un mayor nivel de habilidad de comprensión oral. En consecuencia, esta información podría utilizarse para llevar a cabo un estudio de establecimiento de normas para determinar el punto de corte apropiado que debe representar la capacidad de comprensión oral a nivel B2 del MCER.

A continuación, se presenta las opiniones de los candidatos respecto a la posibilidad de un efecto rebote positivo de la prueba, que incluyen:

1) Opiniones sobre cómo la prueba refleja las prácticas de comprensión oral en el aula.

Por lo general, los temas incluidos en la prueba eran representativos de aquellos estudiados en el instituto durante el bachillerato.

2) Opiniones sobre los procesos de comprensión oral y los usos de estrategias.

Las respuestas a las preguntas sobre los usos de estrategias demuestran que los candidatos creen que sí habían utilizado las estrategias metacognitivas incluidas en el modelo de capacidad de comprensión oral de la PFB para contestar a los ítems. El hecho de que los participantes opinaran que utilizaban estas estrategias proporciona evidencia de que las tareas activan las estrategias descritas como B2 en el MCER de manera que se pueda afirmar que, por lo general, los participantes sí empleaban los descriptores "saber hacer" a nivel B2 del marco.

3) Opiniones sobre la prueba como instrumento de medición válido de la capacidad de comprensión oral

La percepción por parte de los candidatos de que la prueba es un instrumento de medición imparcial de su capacidad es un importante aspecto ético que los desarrolladores de pruebas deben tomar en consideración. Para el presente estudio, los candidatos sí opinaban que las tareas 3 y 4 (ambas del formato tipo test) suponían unas buenas representaciones de sus niveles de capacidad de comprensión oral. No obstante, se mostraron menos satisfechos con las tareas 1 y 2 como indicadores justos de sus niveles de capacidad. Una de las posibles explicaciones en este sentido podría ser que estuvieran más familiarizados con el formato tipo test, así como que con el tiempo llegarían a practicar los hasta ahora más desconocidos formatos de apareamiento de ítems y de respuesta corta si la prueba se introdujera finalmente.

4) Opiniones sobre la instrucción actual de la comprensión oral y los posibles efectos rebote de incluir una prueba de comprensión oral en la prueba final de bachillerato

Queda evidente que los participantes opinan que no se practica lo suficiente la capacidad de comprensión oral en el instituto. Además, parece ser que los participantes creen que esta circunstancia se debe a que la habilidad de comprensión oral no se incluye actualmente en el examen de selectividad. En efecto, todas las preguntas relacionadas con la introducción de una prueba de comprensión oral en el examen final y con el posible aumento de tiempo dedicado a la mejora de esta capacidad en el aula reciben una media de 4, un total acuerdo. Por último, hubo un acuerdo abrumador sobre la propuesta de que se introdujera una prueba de comprensión oral como parte integral de la prueba final de bachillerato.

Como se ha explicado en el resumen del Capítulo 5, para tratar la Pregunta de Investigación (P6), se aplicaron los procedimientos del *Basket Method*. Los jueces analizaron las tareas siguiendo los principios del método y contestaron la pregunta "¿Qué nivel del MCER necesita tener el candidato para poder responder correctamente a este ítem?". Los resultados se analizaron con un modelo *MFRM* en el programa FACETS (Linacre, 2017b). Los resultados de este análisis muestran claramente que los jueces eran de la opinión de que el contenido se extiende por la escala de habilidad desde el nivel B1 (3 ítems) hasta el nivel C1 (2 ítems). Los jueces colocaron la mayoría de los ítems dentro de la franja de nivel B2, adjudicando que había 16 ítems que podrían ser contestados por un candidato que se encuentre en el umbral mínimo de la franja y otros 7 ítems que podrían ser contestados por un candidato que se encuentre en la parte más alta. Todo esto es evidencia de que los jueces opinaron que la prueba es representativa de un nivel comprensión oral B2 del MCER.

Los jueces recibieron una sesión de retroalimentación sobre la primera ronda antes de pasar al estudio principal (el del método *Bookmark*) que produjo un punto de corte de 17. En este punto, los participantes se mostraron satisfechos con las decisiones finales—algo también evidenciado luego por sus respuestas en el

cuestionario posterior al estudio—y afirmaron que estaban de acuerdo con que el punto de corte que habían establecido se empleara en la versión final de la prueba.

Si aplicamos una prueba de tau de Kendall a la diferencia entre la puntuación obtenida en la prueba y la propias estimaciones de los candidatos de su nivel MCER obtenidas en el cuestionario, la resultante correlación es positiva y estadísticamente significativa. El tamaño de efecto de la puntuación total obtenida en la prueba es grande y como consecuencia la hipótesis nula puede ser rechazada. Si aplicamos una prueba de independencia de Chi Cuadrado para comparar las categorías de aquellos candidatos que se autoevaluaron como nivel B1 o B2 (o más) con la categorías dicotómicas de aprobar o suspender (empleando nuestro punto de corte de 17), nos proporciona aún más evidencia de que los candidatos que se autoevaluaron como B2 o más tienen más probabilidades de aprobar la prueba que aquellos que opinaron que no tenían el nivel.

CAPÍTULO 7

Este capítulo sintetiza los resultados del estudio y presenta el argumento de validez de la PFB para la interpretación y uso de puntuaciones de la prueba final. Cada vínculo del argumento de validez se respalda por los resultados del presente estudio.

La definición del dominio se basa en la justificación de que las tareas de la prueba forman una buena representación del ámbito de uso de la lengua meta. En este sentido, las especificaciones de la prueba se basaron en gran parte en la revisión extendida de la literatura y por tanto se encuentran bien fundamentadas en la teoría sustantiva. Esta justificación recibe un mayor respaldo de las preguntas de investigación P6 y P7. Tanto aquellos procedimientos que emplearon el método *Basket* como los que usaron el *Bookmark* requirieron amplias discusiones sobre el emparejamiento de los ítems con el MCER; por

consiguiente, la validez sustantiva y del constructo recibieron un escrutinio minucioso mediante el cual el buen alineamiento del contenido fue confirmado.

Además, la pregunta P7 también sirvió para determinar el punto de corte relevante para la prueba. Según Kane (2012), la asociación de una puntuación determinada con un nivel de capacidad del MCER le confiere más sentido a los ojos de los usuarios de una prueba. Por ejemplo, si la puntuación describe un nivel de capacidad de B2 en la comprensión oral, el usuario de la puntuación recibirá información sobre los tipos de actividades que se supone que un usuario de lengua de este nivel podría desempeñar (Tannenbaum & Cho, 2014). De este modo, las puntuaciones se podrán interpretar con una facilidad que permita que los usuarios las usen para tomar decisiones relevantes. Por consiguiente, esta parte del estudio también contribuye a la inferencia de uso.

La inferencia de evaluación del AI depende de la justificación de que la prueba sea unidimensional (es decir, que sea una prueba de comprensión oral) y que no contenga aspectos de varianza irrelevante de constructo. Un análisis *PCAR* demostró que este era el caso y junto con los estadísticos de *fit* evidencia que la prueba tiene unidimensionalidad sicométrica, es decir, que no evalúa más de un solo rasgo o constructo (Sick, 2010). También se investigó la posibilidad de varianza irrelevante de constructo en una de las secciones del cuestionario. Aquí, los resultados mostraron que, por lo general, no había problemas ni con la prueba ni con su administración. No obstante, sí se destacó la cantidad de tiempo necesario para leer las preguntas como potencialmente demasiado corto, aun a pesar de que fueron los candidatos con una menor habilidad quienes declararon este hecho. Respecto a este tema, me gustaría añadir que los resultados de la pregunta P3 también respaldan la justificación. Los participantes no utilizaron estrategias de *test wiseness* (“intuición o picardía de test” en español), lo cual habría sido una seria refutación de la inferencias en el caso que se uso hubiera sido evidenciado.

La inferencia de generalización recibe respaldo del estudio preliminar de pilotaje, junto con los resultados de las preguntas P1 y P4. La forma final de la prueba contiene ítems que demuestran coherencia interna y las puntuaciones son fiables, y por tanto reproducibles, algo especialmente cierto debido al uso de un modelo Rasch para este estudio. Los ítems se ajustan bien al modelo Rasch y apuntan a que se puede tener confianza en los parámetros de dificultad producidos por este análisis. El buen diseño de las especificaciones permite que se pueda producir pruebas comparables que abarquen múltiples convocatorias, algo fácilmente alcanzable con la medición Rasch.

Las inferencias de explicación y extrapolación se basan en la justificación de que la prueba tenga buena validez del constructo. El presente estudio proporciona lo que Kane (2001) ha definido como una "forma fuerte" de validez del constructo, en el sentido de que la prueba se ha diseñado basándose en una teoría de uso para la comprensión oral, es decir, está dirigida por la teoría. Las diferencias entre las puntuaciones obtenidas en una prueba tienen que poder ser interpretadas: una puntuación más baja debería ser representativa de que el candidato tiene menos habilidad y viceversa. Para la prueba del estudio actual, los informes verbales proporcionan evidencia de que esto es así y de que fueron los candidatos de mayor capacidad quienes entendieron más los audios. Hasta cierto punto, esta justificación también recibe el respaldo de los resultados que ya evidenciaron las justificaciones anteriores, por ejemplo, los resultados de la P6 demuestran que los jueces expertos opinan que la prueba es representativa del constructo de comprensión oral a nivel B2 del MCER. Además, la comparación entre las autoevaluaciones de los propios candidatos y las afirmaciones de acreditaciones oficiales comunicadas en el cuestionario también manifestó la existencia de una correlación positiva entre ellas.

La inferencia de utilización se respalda en buena parte por la constatación del problema (Capítulo 2) y, en concreto, por el hecho de que hace muchos años que el presente examen de selectividad recibe duras críticas negativas por parte de

expertos académicos. Conforme al argumento de que los principales interesados deberían formar parte del proceso de desarrollo de la prueba (Fulcher & Davidson, 2007), este estudio también proporciona evidencia para respaldar la inferencia de utilización al recopilar la opiniones de los propios candidatos. Los resultados muestran que los candidatos consideraban que los temas tratados en la prueba sí eran representativos de los estudiados durante los dos últimos años en la escuela secundaria. No obstante, hubo opiniones diversas respecto a la fiabilidad con que la prueba representaba sus niveles de capacidad, opiniones que parecían depender en buena parte de la medida en que los participantes habían obtenido buenos resultados o no en la prueba.

Por consiguiente, es importante que reconozcamos que la inferencia de utilización es probablemente una de las partes más débiles del argumento de validez, a la vez que supone un elemento de alta importancia en el constructo, dado que el razonamiento fundamental tras la introducción de la prueba es el fomento de un efecto rebote positivo en el sistema educativo. A pesar de que los participantes sí se mostraron en principio de acuerdo con esto, resultaría fundamental la realización de otros estudios posteriores para confirmar que esto es realmente es así.

El uso destinatario es otra importante consideración. Aquí se ha respaldado la interpretación de la puntuación de la prueba, es decir, se ha demostrado que es una representación válida y fiable de la comprensión oral a nivel B2 del MCER. De este modo, la puntuación ha adquirido un valor adicional y por tanto resulta más fácil de interpretar a funcionarios de admisión universitaria. Esta consideración es de suma importancia dado que, en el caso de que la prueba se empleara para tomar decisiones sobre el acceso a la universidad, la responsabilidad de decidir y justificar el nivel de capacidad considerado necesario para cursar una carrera universitaria satisfactoriamente correspondería a los propios departamentos universitarios.

CAPÍTULO 8

Este capítulo concluye el estudio, cuyo propósito ha sido el desarrollo y la provisión de una prueba de capacidad de comprensión oral que se adecue a su propósito y que cumpla con el objetivo de producir un efecto rebote positivo dentro del sistema educativo en España. Extraigo la conclusión de que la prueba sí resulta por lo general un instrumento válido y fiable para la medición de la capacidad de comprensión oral a nivel B2 del MCER.

No obstante, también debo destacar aquí la salvedad de que futuras investigaciones deben partir de otras posibles refutaciones y que siempre habrá refutaciones adicionales que podrán (y deberán) ser expresadas por cualquiera de las partes interesadas. En este respecto, también discuto brevemente cuestiones sobre la imparcialidad y la relevancia de pruebas de alto impacto internacionales y reitero la importancia de poder contar con una acreditación independiente vinculada al MCER para el acceso a la universidad. Concluyo proporcionando varias sugerencias respecto a la implantación de una nueva PFB, haciendo especial hincapié en la importancia del papel que los profesores empeñarían en este proceso.

REFERENCES

AERA (American Educational Research Association), APA (American Psychological Association) & NCME (National Council on Measurement in Education). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Agulló, G. L., & Alastuey, M. C. B. (2017). Analysis of oral skills development in the most used English language textbooks in the second year of baccalaureate in Spain. *Porta Linguarum: Revista Internacional de Didáctica de las Lenguas Extranjeras*, *27*, 107-121.

Alderson, J. C. (1993). Judgements in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 46-57). Alexandria, VA: TESOL.

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

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*, 658–662.

Alderson, J. C. (2010). A survey of aviation English tests. *Language Testing*, *27*, 51-72.

Alderson, J. C. (2012). *Principles and practice in language testing: Compliance or conflict?* Presentation at IATEFL TEA SIG Conference: Innsbruck. Retrieved December, 2012, from http://tea.iatefl.org/inns.html

Alderson, J.C., & Banerjee, J. (2002) State of the art review: Language testing and assessment (part two). *Language Teaching*, *35*(2), 79-113.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2004). *The development of specifications for item development and classification within the Common European Framework of Reference for Languages: learning, teaching, assessment. Reading and listening. Final report of the Dutch CEF construct project*. Retrieved August 2013 from http://eprints.lancs.ac.uk/44/1/final_report.pdf

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of The Dutch CEFR Construct Project, *Language Assessment Quarterly*, *3*(1), 3-30.

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

ALTE. (1998). *Studies in second language testing 6: Multilingual glossary of language testing terms*. Cambridge, UK: Cambridge University Press.

ALTE/Council of Europe (2011) *Manual for language test development and examining. For use with the CEFR*. Retrieved December 2012, from http://www.coe.int/t/dg4/linguistic/ManualtLangageTest-Alte2011_EN.pdf

Amengual Pizarro, M. (2003). A study of different composition elements that raters respond to. *Estudios Ingleses de la Universidad Complutense*, *11*, 53-72.

Amengual Pizarro, M. (2005). Posibles sesgos en el examen de Selectividad. In H. Herrera Soler & J. García Laborda (Eds.), *Estudios y criterios para una evaluación de calidad* (pp. 121- 148). Valencia: Universidad Politécnica de Valencia.

Amengual Pizarro, M. (2006). Análisis de la prueba de inglés de Selectividad de la Universitat de les Illes Balears. *Ibérica, 11*, 29-59.

Amengual Pizarro, M. (2009). Does the English test in the Spanish university entrance examination influence the teaching of English? *English Studies, 90*(5), 582-598.

Amengual Pizarro, M. (2010). Exploring the washback effects of a high-stakes English test. *Revista Alicantina de Estudios Ingleses* (Universidad de Alicante), *23*, 149-170.

Amengual Pizarro, M., & Méndez García, M. (2012). Implementing the oral English task in the Spanish university admission examination: An international perspective of the language. *Revista de Educación, 357*, 105-127.

Anderson, J. R. (2009). *Cognitive psychology and its implications*. NY: Worth Publishers.

Andrews, S. J., Fullilove, J., & Wong, Y. (2002). Targeting washback: A case study. *System*, *30*, 207–233.

Anckar, J. (2007). *Fruits of the happy union of quality and quantity: Analysis of an MC-test of listening comprehension*. Presentation at EALTA conference. Sitges, Spain. Retrieved March 2013 from http://www.ealta.eu.org/conference/2007/docs/pres_friday/Anckar.pdf

Anckar, J. (2011). *Assessing foreign language listening comprehension by means of the multiple-choice format: Processes and products*. Retrieved March 2013 from http://www.academia.edu/2969303/Assessing_foreign_language_listening_comprehension_by_means_of_the_multiple-choice_format_processes_and_products

Aryadoust, V. (2011a). Application of the fusion model to while-listening performance tests. *SHIKEN: JALT Testing & Evaluation SIG Newsletter, 15*(2), 2-9.

Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency*. New Castle: Cambridge Scholars Publishing.

Aryadoust, V. (2015) Fitting a Mixture Rasch Model to English as a Foreign Language listening tests: The role of cognitive and background variables in explaining latent differential item functioning, *International Journal of Testing, 15*(3), 216-238.

Aryadoust, V., Goh, C., & Kim, L. O. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly*, *8*(4), 361-385.

Ashton, K, Salamoura, A., & Diaz, E. (2012). The BEDA impact project: A preliminary investigation of a bilingual programme in Spain, *Research Notes 50*, 34–42.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453-476.

Bachman, L F. (2004). *Statistical analysis for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34.

Bachman, L.F. (2006). Generalizability: A journey into the nature of empirical research in applied linguistics. In M. Chalhoub-Deville, C. Chapelle & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp.165-207). Dordrecht: John Benjamins.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, et al. (Eds.), *Language testing reconsidered* (pp. 41–71). Ottawa: University of Ottawa Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L.F., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*, 189-208.

Badger, R., & Yan, X. (2012). The use of tactics and strategies by Chinese students in the Listening components of IELTS. In L. Taylor & C. J. Weir (Eds.), *IELTS Collected papers 2: Research in reading and listening assessment. Studies in language testing*, *34*. Cambridge: UCLES/CUP

Banerjee, J. (2004). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the CEF: Section D: Qualitative analysis methods*. Retrieved March, 2013 from http://www.coe.int/t/dg4/linguistic/CEF-ref-supp-SectionD.pdf

Barta, E. (2010). Test takers' listening comprehension sub-skills and strategies. *WoPaLP, 4*. Retrieved March, 2013 from http://langped.elte.hu/WoPaLParticles/W4Barta.pdf

Batty, A. O. (2015). A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing*, *32*(1), 3–20.

Bejar I. (2008). Standard Setting: What is it? Why is it important? *Educational Testing Service R&D Connections, 7,* October 2008, pp. 1-5. Retrieved January 2017 from: https://www.ets.org/Media/Research/pdf/RD_Connections7.pdf

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper*. TOEFL Monograph Series, MS 19. New Jersey: Educational Testing Service.

Berne, J. (1995). How does varying pre-listening activities affect second language listening comprehension? *Hispania*, *78*, 316-29.

Biber, D., Conrad, S. , Reppen, R. , Byrd, P., & Helt, M. (2002), Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, *36*, 9-48.

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2011). *What makes listening difficult? Factors affecting second language listening comprehension*. College Park, MD: University of Maryland.

Boletín Oficial del Estado (BOE). (2006). Ley Orgánica 2/2006, de 3 de mayo, de Educación. Retrieved June 2015 from http://www.boe.es/boe/dias/2006/05/04/pdfs/A17158-17207.pdf

Boletín Oficial del Estado (BOE). (2007). Real Decreto 1467/2007, de 2 de noviembre. Retrieved June 2015 from https://www.boe.es/boe/dias/2007/11/06/pdfs/A45381-45477.pdf

Boletín Oficial del Estado (BOE). (2008). Real Decreto 1892/2008, de 14 de noviembre. Retrieved June 2015 from https://www.boe.es/boe/dias/2008/11/24/pdfs/A46932-46946.pdf

Boletín Oficial del Estado (BOE). (2012). Real Decreto 961/2012, de 3 de julio. Retrieved June 2015 from https://boe.es/boe/dias/2012/07/03/pdfs/BOE-A-2012-8849.pdf

Boletín Oficial del Estado (BOE). (2013). Ley Organica 295 8/2013, de 9 de diciembre. Retrieved June 2015 from https://www.boe.es/buscar/pdf/2013/BOE-A-2013-12886-consolidado.pdf

Boletín Oficial del Estado (BOE). (2014). Real Decreto/412/2014, de 6 de junio. Retrieved June 2015 from https://www.boe.es/boe/dias/2014/06/07/pdfs/BOE-A-2014-6008.pdf

Boletín Oficial del Estado (BOE). (2015). Real Decreto/1105/2014, de 26 de diciembre. Retrieved June 2015 from https://www.boe.es/boe/dias/2015/01/03/pdfs/BOE-A-2015-37.pdf

Boletín Oficial del Estado (BOE). (2016). Orden ECD/1941/2016, de 22 de diciembre. Retrieved Jan 2018 from https://www.boe.es/buscar/pdf/2016/BOE-A-2016-12219-consolidado.pdf

Boletín Oficial del Estado (BOE). (2018). Orden ECD/42/2018, de 25 de enero. Retrieved March 2018 from https://www.boe.es/boe/dias/2018/01/26/pdfs/BOE-A-2018-984.pdf

Boletín Oficial de la Junta de Andalucía (BOJA). (2008). Orden 169/2008, de 5 de agosto. Retrieved June 2015 from Boletín Oficial del Estado (BOE). (2006). Ley orgánica 2/2006, de 3 de mayo, de Educación. Retrieved June 2015 from http://www.juntadeandalucia.es/boja/2008/169/d2.pdf

Bond, T. G. (2003). Validity and assessment: A Rasch measurement perspective. *Metodología de las Ciencias del Comportamiento 5*(2), 179–194. Retrieved June, 2013 from http://eprints.jcu.edu.au/1799/1/bondValidity.pdf

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

Bonk, W. J. (2000). Second language lexical knowledge and listening comprehension. *International Journal of Listening Comprehension*, *14*, 14–31.

Boroughs, R. (2003). The change process at paper level. Paper 4: Listening. In C. J. Weir & M. Milanovich (Eds.), *Continuity and change: Revising the Cambridge Proficiency in English Examination: 1913–2002* (pp. 315–366). Cambridge: Cambridge University Press,

Breeze, R., & Roothooft, H. (2014). Teacher perspectives on implementing Cambridge English: Young Learners exams in Spanish schools. *Cambridge English: Research Notes*, *57*, 3-13.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, *19*(4), 369–394.

Brown, A. (1993). The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing*, *10*, 277–303.

Brown, G. (2008). Selective listening. *System*, *36*, 10–21.

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge: Cambridge University Press.

Brown, J. D. (2012). Classical test theory. *The Routledge handbook of language testing*. London: Routledge. Retrieved January 2017 from https://www.routledgehandbooks.com/doi/10.4324/9780203181287.ch22

Brunfaut, T., & Harding, L. (2014). *Linking the GEPT listening test to the Common European Framework of Reference*. Taiwan: Language Training and Testing Centre.

Brunfaut, T., & Révész, A. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, *35*, 31–65.

Brunfaut, T., & Révész, A. (2015), The role of task and listener characteristics in second language listening. *TESOL Quarterly*, *49*, 141–168.

Buck, G. (1991). The testing of listening comprehension: an introspective study. *Language Testing*, *8*(1), 67-91.

Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing*, *11*(2), 145-170.

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

Buck C., & Tatsuoka, K. (1998) Application of the rule-space procedure to language testing: examining attributes of a free response listening test. *Language Testing,15*, 119–157.

Bueno Alastuey, M., & Luque, Agulló, G. (2012). Competencias en lengua extranjera exigibles en la Prueba de Acceso a la Universidad: una propuesta para la evaluación de los aspectos orales. *Revista de Educación*, *357*, 81-104.

Canagarajah, A. S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly*, *3*(3), 229–242.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richard & R. Schmidt (Eds.), *Language and communication* (pp. 2-28). Harlow: Longman.

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

Carlsen, C, H,. (2018). The adequacy of the B2 level as university entrance requirement, *Language Assessment Quarterly*, 15:1, 75-89.

Carrell, P. L., Dunkel, P. A., & Mallaun, P. (2002). *The effects of notetaking, lecture length, and topic on the listening component of the TOEFL 2000 test* (TOEFL Monograph No. 23). Princeton, NJ: Educational Testing Service.

Carrell, P. L., Dunkel, P. A., & Mollaun, P. (2004). The effects of notetaking, lecture length, and topic on a computer-based test of ESL listening comprehension. *Applied Language Learning, 14,* 83–105.

Celce-Murcia, M., Dérnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics*, *6*, 5-35.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, *20*(4), 369–383.

Chalhoub-Deville, M. (2009). The intersection of test impact, validation, and educational reform policy. *Annual Review of Applied Linguistics*, *29*, 118-131.

Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing*, *33*(4), 453-472.

Chamot, A, U. (2005). Language learning strategy instruction: Current issues and research. *Annual Review of Applied Linguistics*, *25*, 112-130.

Chang, A. C., & Read, J. (2006). The effects of listening support on the listening performance of EFL learners. *TESOL Quarterly*, *40*(2), 375–397.

Chang, A. C., & Read, J. (2008). Reducing listening test anxiety through various forms of listening support. *TESL-EJ*, *12*, 1–25.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. Bachman & A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). Cambridge: Cambridge University Press.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple... *Language Testing*, *29*(1), 19–27.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). (Eds.) *Building a validity argument for the test of English as a foreign language*. New York: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3-13.

Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing*, *20*(4), 409-439.

Chapelle, C.A., & Voss, E. (2016). Utilizing technology in language assessment. In Shohamy E., Or I., & May S. (Eds.) *Language testing and assessment*. *Encyclopedia of language and education* (pp.149-161). Dordrecht: Springer Netherland.

Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing, 35*(1), 3-25.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study. Studies in language testing, 21*. Cambridge, MA: Cambridge University Press.

Cheng, L. (2008). Washback, impact and consequences. In E. Shohamy & N. H. Hornberger (Eds.), *Language testing and assessment. Encyclopedia of language and education* (pp. 349–364). New York, NY: Springer.

Cheng, L. (2013). Consequences, impact, and washback. In A. J. Kunnan (Ed.), *The companion to language assessment* (Chapter 68). Hoboken, NJ, USA: John Wiley & Sons, Inc. Retrieved January, 2016 from http://doi.wiley.com/10.1002/9781118411360.wbcla071

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

Choi, I. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing*, *25*(1), 39-62.

Cizek, G. J. (2006). Standard setting. In S. M. Downing & T. M. Haladyna (Eds.) *Handbook of test development.* Mahwah: Lawrence Erlbaum Associations.

Cizek, G. J. (2012). The forms and functions of evaluations in the standard setting process. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 165 - 178). New York: Routledge.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 212-225.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks: SAGE Publications.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests *Educational and Psychological Measurement, 68*, 397-412.

Cohen, A. D. (1998). *Strategies in learning and using a second language*. Harlow: Longman.

Cohen, A.D. (2000). Exploring strategies in test-taking: Fine-tuning verbal reports from respondents. In G. Ekbatani & H. Pierson (Eds.), *Learner-directed assessment in ESL* (pp.127-150). Mahwah, NJ: Lawrence Erlbaum.

Cohen, A. D. (2005). Coming to terms with language learner strategies: What do strategy experts think about the terminology and where would they direct their research? *Working paper No. 12*. Retrieved September, 2013 from http://www.crie.org.nz/research-papers/Andrew%20Cohen%20WP12.pdf

Cohen, A. D. (2007a). Coming to terms with language learner strategies: Surveying the experts. In A. D. Cohen & E. Macaro (Eds.), *Language learner strategies: 30 years of research and practice* (pp. 29-45). Oxford: Oxford University Press.

Cohen, A. D. (2007b). The coming of age for research on test-taking strategies. In J. Fox, M. Wesche, D. Bayliss, et al. (Eds.), *Language testing reconsidered* (pp. 89-111). Ottawa: University of Ottawa Press.

Cohen, A. D. (2011). L2 learner strategies.(Ch. 41). In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning, Vol. II - Part V. Methods and instruction in second language teaching* (pp. 681-698). Abingdon, England: Routledge.

Cohen, A. D. (2012). Test-taking strategies and task design. In G. Fulcher & F. Davidson (Eds.) *The Routledge handbook of language testing* (262-277), New York and London: Routledge.

Cohen, A. D. (2013). Using Test-Wiseness Strategy Research in Task Development. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 893–905). Hoboken, NJ, USA: John Wiley & Sons, Inc. Retrieved January, 2016 from http://doi.wiley.com/10.1002/9781118411360.wbcla006

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education*, *12*(4), 343-366.

Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study. *System*, *29*(1), 1-14.

Coste, D. (2007). Contextualising uses of the Common European Framework of Reference for Languages. In F. Goullier (Ed.) *Report on The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities* (pp. 38-47). Strasbourg: Council of Europe.

Couet, R., & Arnaiz, P. (Coord.).(2009). *El marco común europeo de referencia para las lenguas: Adecuación del documento a la enseñanza universitaria.* Las Palmas de Gran Canaria: Romenable.

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

Council of Europe. (2003). *Manual for relating examinations to the Common European Framework of Reference for Languages* [Preliminary pilot version]. Strasbourg: Council of Europe.

Council of Europe. (2008). *Recommendation CM/Rec (2008)/7 of the Committee of Ministers to member states on the use of the Council of Europe's Common European Framework of Reference for Languages (CEFR) and the promotion of plurilingualism.* Retrieved June 2015, from http://www.coe.int/t/dg4/linguistic/Source/SourceForum07/Rec%20CM%202008-7_EN.doc.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages (CEF). A manual*. Strasbourg: Council of Europe. Retrieved September 2013 from http://www.coe.int/T/DG4/Linguistic/Manuel1_EN.asp

Council of Europe. (2017). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors*. Strasbourg: Council of Europe.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Boston, MA: Pearson Education, Inc.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

CRUE. (2016). *Documento marco de política lingüística para la internacionalización del sistema universitario español*. Retrieved January 2017 from https://www.upf.edu/documents/6602910/7420475/2016_+Documento+Marco+Pol%C3%ADtica+Linguistica+Internacionalización+SUE.pdf/87a63fff-c2e1-f4fa-47d9-b8db1be5b407

Crystal, D. (2012). *English as a global language*. Cambridge: Cambridge University Press.

Davidson, F. (2012). Test specifications. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. New Jersey: John Wiley & Sons.

Davidson, F., & Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language Teaching*, *40*, 231-241.

Davies, A. (1997). Demands of being professional in language testing. *Language Testing*, *14*(3), 328–339.

Davies, A. (2008). Ethics, professionalism, rights and codes. In E. Shohamy (Ed.), *Language testing and assessment. Encyclopedia of language and education* (pp. 429-443). Dordrecht: Springer Netherland.

Davies, A. (2010). Test fairness: A response. *Language Testing, 27*(2) 171–176.

Davies, A. (2012). Kane, validity and soundness. *Language Testing, 29*(1) 37 –42.

Davies, A., & Elder, C. (2005) Validity and validation in language testing. In E. Hinkle (Ed.), *Handbook of research in second language teaching and learning* (volume 1) (pp. 795–813). Mahwah, NJ: Lawrence Erlbaum.

Dearden, J. (2015). *English as a medium of instruction - a growing global phenomenon*. London: British Council.

De Jong, J. (2014). *Extending and complementing the Common European Framework*. Presentation at EALTA conference. Warwick, UK. Retrieved January 2016 from http://www.ealta.eu.org/conference/2014/presentations/John%20deJong%20EALTA%202014.pdf

De Jong, J., & Zheng, Y. (2011). *Research note: Applying EALTA guidelines: A practical case study on Pearson Test of English Academic*. Retrieved January 2018 from https://pearsonpte.com/wp-content/uploads/2014/07/RN_ApplyingEALTAGuidelines_2010.pdf

De Jong, J., & Zheng, Y. (2016). Linking to the CEFR: validation using a priori and a posteriori evidence. In, Banerjee, J., & Tsagari, D. (Eds.) *Contemporary Second Language Assessment* (pp. 83-100). London: ,Bloomsbury Academic.

Derwing, T., & Munro, M. (2001). What speaking rates do non-native listeners prefer? *Applied Linguistics, 22*(3), 324–337.

Deygers, B., & Zeidler, B. (2015). *The CEFR & university entrance tests. A state of affairs in Europe*. ALTE Bergen, 8 May 2015.

Deygers, B., Zeidler, B., Vilcu, D., & Hamnes Carlsen, C. (2017). One framework to unite them all? Use of the CEFR in European university entrance policies. *Language Assessment Quarterly*, *15*(1), 3-15.

Díez Bedmar, M. B. (2012). The use of the Common European Framework of Reference for Languages to evaluate compositions in the English exam section of the university admission examination. *Revista de Educación*, *357*, 55-80.

Díez-Bedmar, M. B. (2017). Fine-tuning descriptors for CEFR B1 level: Insights from learner corpora. *ELT Journal*, ccx052–ccx052.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative and mixed methodologies*. Oxford: Oxford University Press.

Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development,* (pp. 3-26). Mahwah, NJ: Erlbaum.

Duff, P.A. (2006). Beyond generalizability: Contextualization, complexity, and credibility in applied linguistics research. In M. Chalhoub-Deville, C. Chapelle & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives,* (pp. 65–95). Philadelphia: John Benjamins.

EALTA. (2006). *The EALTA guidelines for good practice in language testing and assessment.* Retrieved December 2012 from http://www.ealta.eu.org/documents/archive/guidelines/English.pdf

Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt: Peter Lang.

Eckes, T. (2012). Examinee-centered standard setting for large-scale assessments: The prototype group method. *Psychological Test and Assessment Modeling*, *54*, 257–283.

Eckes, T., Ellis, M., Kalnberzina, V., Pižorn, K., Springer, C., Szollás, K., & Tsagari, C. (2005). Progress and problems in reforming public language examinations in Europe: Cameos from the Baltic States, Greece, Hungary, Poland, Slovenia, France, and Germany. *Language Testing*, *22*, 355–377.

Ehrich, J, F., & Henderson, D, B., (2018). Rasch analysis of the metacognitive awareness listening questionnaire (MALQ), *International Journal of Listening*, (p1-13).

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.

Elkhafaifi, E. (2005). The effect of pre-listening activities on listening comprehension in Arabic learners. *Foreign Language Annals*, *38*, 505–513.

Elliot, M., & Wilson. J. (2013). Context validity. In Geranpayeh, A., & L.Taylor (Eds.) *Examining listening: Research and practice in assessing second language listening (*pp.152-241). Cambridge: Cambridge University Press.

Engelhard, G. (2009). Evaluating the judgments of standard-setting panelists using Rasch measurement theory. In E. V. Smith Jr. & G. E. Stone (Eds.), *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models* (pp. 312-346). Maple Grove, MN: JAM Press.

Engelhard, G. (2011). Evaluating the bookmark judgments of standard-setting panelists. *Educational and Psychological Measurement*, *71*, 909–924.

European Commission, CILT. (2006). *ELAN: Effects on the European Union economy of shortages of foreign language skills in enterprise*. Retrieved December 2015 from http://ec.europa.eu/languages/policy/strategic-framework/documents/elan_en.pdf

European Commission Survey Lang. (2012). *First European survey on language competences: Final report*. Retrieved June 2015 from http://ec.europa.eu/languages/policy/strategic-framework/documents/language-survey-final-report_en.pdf

Ericsson, K.A., & Fox, M.C. (2011). Thinking aloud is not a form of introspection but a qualitatively different methodology: Reply to Schooler (2011*). Psychological Bulletin*, *137*(2), 351-354.

Fernández Álvarez, M. (2007). *Propuesta metodológica para la creación de un nuevo examen de inglés en las pruebas de acceso a la universidad.* Unpublished PhD dissertation. University of Granada, Spain. Retrieved August 2015 from http://digibug.ugr.es/handle/10481/1449#.Vp4ZvUsxIds

Field, A. (2009). *Discovering statistics using SPSS*. CA: Sage Publications.

Field, J. (2008a). *Listening in the language classroom*. Cambridge: Cambridge University Press.

Field, J. (2008b). Bricks or mortar: Which parts of the input does a second language listener rely on? *TESOL Quarterly*, *42*(3), 411-432.

Field, J. (2008c). Guest editor's introduction. Emergent and divergent: A view of second language listening research. *System, 36*, 2–9.

Field, J. (2008d). Revising segmentation hypotheses in first and second language listening. *System*, *36*, 35-51.

Field, J. (2009). *Two bites of the cherry: The effects of replay on the listener*. Paper presented at the BAAL Annual Meeting 2009, Newcastle.

Field, J. (2011). The elusive skill: How can we test L2 listening validly? In Powell-Davies, P. (Ed.). *New directions: Assessment and evaluation. A collection of papers* (pp.139-145), British Council, Malaysia. Retrieved September, 2013 from http://www.britishcouncil.jp/sites/britishcouncil.jp/files/eng-new-directions-en.pdf

Field, J. (2012a). *Cognitive validity in language testing: Theory and practice*. Presentation given at CRELLA Summer Research Seminar. Retrieved September, 2013 from http://www.beds.ac.uk/__data/assets/pdf_file/0007/215845/Cognitive-valdity-summerseminar-Read-Only-Compatibility-Mode.pdf

Field, J. (2012b) The cognitive validity of the lecture-based paper in the IELTS listening test. In L. Taylor & C. J. Weir (Eds.) IELTS Collected Papers 2: Research in reading and listening assessment. *Studies in Language Testing*, *34*. Cambridge: UCLES/CUP.

Field, J. (2013a). Cognitive validity. In Geranpayeh, A., & L.Taylor (Eds.). *Examining listening: Research and practice in assessing second language Listening* (pp. 77–151). Cambridge: Cambridge University Press.

Field, J. (2013b). *Good at listening or good at listening tests?* Conference presentation, ANUPI 2013. Huatulco. Retrieved January 2016 from https://www.google.es/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwiYx_HP_bDLAhWD8RQKHc9eAqkQFggcMAA&url=http%3A%2F%2Fwww.beds.ac.uk%2F__data%2Fassets%2Fpowerpoint_doc%2F0010%2F297037%2FJF-Mexico-ANUPI-2013-Good-at.pptx&usg=AFQjCNGZJTwzwoXwIg4TfZswkMRg5kR7ag&bvm=bv.116274245,d.d24

Field, J. (2015). *The effects of single and double play upon listening test outcomes and cognitive processing*. ARAGs Research Reports Online; Vol. AR/2015/003. London: The British Council.

Field, J. (2016). *A profile of the academic listener. Language testing forum conference*. UKALTA, November 25th, University of Reading. Retrieved February 2017 from http://ukalta.org/wp-content/uploads/2016/10/Field_LTF2016.pdf

Field, J. (2017). *Mind the gap: Listening tests versus real world listening*. IATEFL TEASIG conference, October 2017, CRELLA, University of Bedfordshire. Retrieved February 2018 from https://tea.iatefl.org/wp-content/uploads/2015/10/John-Field_Mind-the-gap-TEA-SIG-Oct-17-delivered.pdf

Figueras, N., & Noijons, J. (Eds.). (2009). *Linking to the CEFR levels: Research perspectives*. Arnhem, The Netherlands: CITO.

Flowerdew, J., & Miller, L. (2005). *Second language listening: Theory and practice.* Cambridge: Cambridge University Press.

Fortanet-Gómez, I. (2013). *CLIL in tertiary education: Towards a multilingual language policy.* Canada: Short Run Press.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, *16*, 2–32.

Froetscher, D. (2016). A new national exam: A case of washback. In J. Banerjee & D. Tsagari (Eds.), *Contemporary second language assessment* (pp. 61–81). London: Continuum.

Fulcher, G. (1999). Assessment in English for academic purposes: Putting content validity in its place. *Applied Linguistics*, *20*(2), 221-236.

Fulcher, G. (2003). *Testing second language speaking*. Harlow, Essex: Pearson Education Limited.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly*, *1*(4), 253-266.

Fulcher, G. (2008). Testing Times Ahead? *Liaison Magazine*, *1*, 20-24. Published by the UK Subject Centre for Languages, Linguistics and Area Studies, University of Southampton.

Fulcher, G. (2009). Test use and political philosophy. *Annual Review of Applied Linguistics*, *29*, 3-20.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London and New York: Routledge.

Fulcher, G., & Davidon, F. (2009) Test architecture, test retrofit. *Language Testing*, *26*(1) 123–144.

Fulcher, G., & Davidson, F. (2012). Introduction. In G. Fulcher & F. Davidson (Eds.) *The Routledge handbook of language testing* (pp.1-17). New York and London: Routledge.

Fulcher, G., & Owen, N. (2016). Dealing with the demands of language testing and assessment. *The Routledge handbook of English language teaching*, 109-120.

García Laborda, J. (2006). Designing an Internet based tool for oral evaluation in the University Access Examination in Spain. In J. Colpaert, W. Decco, S. Van Bueren & Godfroid, A. (Eds.), *CALL and monitoring the learner: Proceedings of the international CALL conference*, *Antwerp* (pp. 92-94), Antwerp: Universiteit Antwerpen.

García Laborda, J. (2007). On the net: Introducing standardized EFL/ ESL exams. *Language Learning & Technology*, *11*(2), 3-9.

García Laborda, J. (2010). ¿Necesitan las universidades españolas una prueba de acceso informatizada? El caso de la definición del constructo y la previsión del efecto en la enseñanza para idiomas extranjeros. *Revista de Orientación y Psicopadagogía*, *21*(1), 71-80.

García Laborda, J. (2012). Preliminary findings of the PAULEX Project: A proposal for the internet-based Valencian university entrance examination. *Journal of Language Teaching & Research, 3*(2), 250-255.

García Laborda, J. G., Bakieva, M., Gonzalez-Such, J., & Pavon, A. S. (2010). Item transformation for computer assisted language testing: The adaptation of the Spanish university entrance examination. *Procedia: Social and Behavioral Sciences, 2*, 3586-3590.

García Laborda, J., Bejarano, L.G., & Simons, M. (2012). ¿Cuánto aprendí en lasecundaria? Las actitudes de los estudiantes universitarios de primer año respecto a la relación enseñanza-aprendizaje de su segunda lengua en la escuela secundaria en tres contextos internacionales. *Educación XX1*, *15*(2), 159-184.

García Laborda, J., & Fernández Álvarez, M. (2011). Teachers' opinions towards the integration of oral tasks in the Spanish university examination. *International Journal of Language Studies*, *5*(3), 1-12.

García Laborda, J., & Fernández Álvarez, M. (2012). Actitudes de los profesores de Bachillerato adscritos a la Universidad de Alcalá y a la Universidad Pública de Navarra ante la preparación y efecto de la Prueba de Acceso a la Universidad. *Revista de Educación*, *357*, 29-54.

García Laborda, J., & Gimeno Sanz, A. (2007). Adaptación del examen de inglés de las pruebas de acceso a la universidad a un entorno informático: estudio sobre las tipologías de pregunta. In R. Monroy & A. Sánchez (Eds.), *25 años de lingüística aplicada en España: Hitos y Retos* (pp. 723-730). Murcia: EDINUM. Retrieved June 2015, from http://www.um.es/lacell/aesla/contenido/pdf/6/garcia4.pdf

Garcia Laborda, J., Gimeno Sanz, A., & Martínez Sáez, A. (2008). *Anticipating washback in a computer based university entrance examination: Key issues*. Paper session presented at the CALL conference 2008, Antwerp. Retrieved June 2015 from http://www.eric.ed.gov/PDFS/ED504035.pdf

García Laborda, J., Magal Royo, T., & Bárcena Madera, E. (2015). An Overview of the needs of technology in language testing in Spain. *Procedia: Social and Behavioral Sciences, 186*, 87 – 90.

Garcia Laborda, J., Magal Royo, T., & Bakieva, M. (2010). A first approach to the analysis of student motivation in the trial version of the computer based university entrance examination. *Proceedings of the 14th Edition. International CALL Research Conference*, 84-87.

García Laborda, J., & Martín-Monje, E. (2013). Item and test construct definition for the new Spanish baccalaureate final evaluation: A proposal. *International Journal of English Studies*,*13*(2), 69-88.

Gass, S., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Geranpayeh, A., & Taylor, L. (2008). Examining listening: Developments and issues in assessing second language listening. *Cambridge ESOL: Research Notes, 32*, 2–5, Retrieved March, 2013 from http://www.cambridgeesol.org/rs_notes/rs_nts32.pdf.

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching*, *40*, 97–118.

Gilmore, A. (2011). "I prefer not text": Developing Japanese learners' communicative competence with authentic materials. *Language Learning*, *61*, 786–819.

Gila González, B. (1996). Encuesta sobre la selectividad. *GRETA, 4*(2), 90-92.

Gimeno Sanz, A., & De-Sequeira, J. (2009). Designing feedback to support language acquisition using the "ingenio" authoring tool. *Procedia: Social and Behavioral Sciences, 1*, 1239-1243.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing*, *19*(2), 133-167.

Goh, C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*. *28*(1), 55-75.

Goh, C. M. (2002). Exploring listening comprehension tactics and their interaction patterns. *System*. *30*(2),185-206.

Goh, C. M., & Aryadoust, V. (2015). Examining the notion of listening subskill divisibility and its implications for second language listening. *International Journal of Listening*, *29*(3), 109-133.

Goh, C. M., & Taib, Y. (2006). Metacognitive instruction in listening for young learners. *ELT Journal*, *60*(3), 222-232.

Goh, C. M., & Hu, G. (2014). Exploring the relationship between metacognitive awareness and listening performance with questionnaire data. *Language Awareness*, *23*, 255–274.

González-Such, J., Jornet, J. M., & Bakieva, M. (2013). Consideraciones metodológicas sobre la evaluación de la competencia oral en L2. *Revista Electrónica de Investigación Educativa*, *15*(3), 1-20. Retrieved August 2015 from http://redie.uabc.mx/vol15no3/contenido-glez-jornet.html

Goullier, F. (2007). *Report on The Common European Framework of Reference for Languages (CEFR) and the development of language policies: Challenges and responsibilities*. Strasbourg: Language Policy Division, Council of Europe. Retrieved June 2015 from http://www.coe.int/t/dg4/linguistic/Publications_en.asp

Graham, S., & Macaro, E. (2008). Strategy instruction in listening for lower- intermediate learners of French. *Language Learning*, *58*, 747–783.

Graham, S. J., Santos, D., & Vanderplank, R. (2008). Listening comprehension and strategy use: A longitudinal exploration. *System*, *36*, 52–68.

Graham, S. J., Santos, D., & Vanderplank, R. (2010). Strategy clusters and sources of knowledge in French L2 listening comprehension. *Innovation in Language Learning and Teaching*, *4*(1), 1-20.

Green, A. (1998). *Verbal Protocol Analysis in language testing research: A handbook.* Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

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

Green, A. (2008). English profile: Functional progression in materials for ELT. *Research Notes, 33*, 19-25.

Green. A. (2015, December 2nd). Learning oriented test preparation. (Webinar). IATEFL TEASIG.

Green, A. (2018). Linking tests of English for Academic Purposes to the CEFR: The score user's perspective, *Language Assessment Quarterly*, *15*(1), 59-74.

Green, A., & Inoue, C. (2016). *Linking speaking exams to the CEFR: Issues and Challenges*. CRELLA Winter Research Seminar 30 November, 2016. Retrieved January 2017 from: https://www.beds.ac.uk/__data/assets/pdf_file/0017/525311/Green-Inoue-Winter-Seminar-2016.pdf

Green, R. (2013). *Statistical analyses for language test developers*. Basingstoke: Palgrave Macmillan.

Green, R. (2017). *Designing listening tests: A practical approach*. Basingstoke: Palgrave Macmillan.

Green, R. & C. Spoettl. (2011). *Building up a pool of standard setting judges: problems solutions and insights*. EALTA, Siena, Italy 5th-8th of May, 2011. Retrieved January 2016 from: http://www.ealta.eu.org/conference/2011/friday/EALTA2011_Green_Spoettl.pdf

Griffiths, M. (2017). The impact of international speaking and listening assessments on primary school bilingual learning: Insights from survey research. *Educación bilingüe: Tendencias educativas y conceptos claves: Bilingual educational: Trends and key concepts*, 145-158.

Gruba, P. (1993). A comparison study of audio and video in language testing. *JALT Journal*, *16*(1), 85-88.

Gu, Y. (2014). To code or not to code: Dilemmas in analysing think-aloud protocols in learning strategies research. *System*, *43*, 74-81.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R., & Jirka, S. (2006) Anchor-based methods for judgmentally estimating item statistics. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 399-420). Mahwah, NJ: Erlbaum.

Hambleton, R., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (pp. 433-470). Westport, CT: American Council on Education/ Praeger Publishers.

Hamp-Lyons, L. (1997). The intersection of test impact, validation, and educational reform policy. *Language Testing*, *14*, 295-303.

Hamp-Lyons, L. (2000). Social, professional, and individual responsibility in language testing. *System 28*(4), 579–591.

Hamp-Lyons, L., & N. Tavares. (2011). Interactive assessment - a dialogic and collaborative approach to assessing learners' oral language. In D. Tsagari and I. Csepes (Eds.), *Classroom-based language assessment* (pp.29-46). Frankfurt: Peter Lang.

Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening: Research perspective* (pp. 241- 268). Cambridge: Cambridge University Press.

Hanushek, E. A., Ruhose, J., & Woessmann, L. (2015). *Economic gains for U.S. states from educational reform*. Cambridge, Mass: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w21770

Hanushek, E. A., & Woessmann, L. (2010). *The economics of international differences in educational achievement* (No. w15949). National Bureau of Economic Research. Retrieved July 2015 from http://www.nber.org/papers/w15949.pdf

Harding, L. (2008). Accent and academic listening assessment: A study of test-taker perceptions. *Melbourne Papers in Language Testing*, *13*(1), 1-33.

Harding, L. (2011). *Accent and listening assessment: A validation study of the use of speakers with L2 accents on an academic English listening test*. Frankfurt: Peter Lang.

Harding, L. (2012). *Language testing, world Englishes and English as a lingua franca: The case for evidence-based change*. Invited keynote address, CIP symposium 2012, University of Copenhagen, Denmark. Retrieved January 2016 from http://cip.ku.dk/arrangementer/tidligere/symposium_2012/Luke_Harding.pdf

Harding, L. (2014a). Communicative language testing: Current issues and future research. *Language Assessment Quarterly*, *11*(2), 186-197.

Harding, L. (2014b). *Adaptability and ELF communication: Next steps for communicative language testing?* (Invited plenary presentation). Centro de Lenguas Modernas, University of Granada: IATEFL Testing Evaluation and Assessment SIG conference. Granada.

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, *20*(3), 245-270.

Harris, T. (2002). Chasing windmills: why spoken English is not taught in Spanish secondary schools. *English Speaking Board*, *35*(1), 26-30.

Harsch, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly*, *11*(2), 152-169.

Harsch, C., & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly*, *12*(4), 333-362.

Hawkey, R. A. H. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000 (Studies in Language Testing 24)*. Cambridge, UK: Cambridge University Press and Cambridge ESOL.

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, *9*(1), 1-11.

Herrera Soler, H. (1999). Is the English test in the Spanish university entrance examination as discriminating as it should be? *Estudios Ingleses de la Universidad Complutense*, *7*, 87-103.

Herrera Soler, H. (2000-2001). The effect of gender and working place of raters on university entrance examination scores. *Revista Española de Lingüística Aplicada*, *14*, 161-180.

Hirai, A., & R. Koizumi. (2009). Development of a practical speaking test with a positive impact on learning using a story retelling technique. *Language Assessment Quarterly*, *6*, 151–167.

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*, 403–430.

Huff, K. (2003). *An item modeling approach to providing descriptive score reports*. Unpublished doctoral dissertation, University of Massachusetts Amherst.

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

Hulstijn, J. H. (2003). Connectionist models of language processing and the training of listening skills with the aid of multimedia software. *Computer Assisted Language Learning*, *16*, 413–425.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*, 663-667.

Hulstijn, J. H. (2011). Language Proficiency in Native and Nonnative Speakers: An Agenda for Research and Suggestions for Second-Language Assessment, *Language Assessment Quarterly*, *8*(3), 229-249.

Ilc, G., & Stopar, A. (2015). Validating the Slovenian national alignment to CEFR: The case of the B2 reading comprehension examination in English. *Language Testing*. *32*(4) 443–462.

ILTA. (2000). *Code of ethics*. Retrieved December, 2012 from http://www.iltaonline.com/images/pdfs/ILTA_Code.pdf

ILTA. (2007). *Guidelines for practice*. Retrieved December, 2012 from http://www.iltaonline.com/images/pdfs/ILTA_Guidelines.pdf

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*, 219–244.

Instituto Nacional de Evaluación Educativa (INEE). (2013). *Estudio europeo de competencia lingüística*. Location: Ministerio de Educación, Cultura y Deporte.

Iwashita, N., & Elder, C. (1997). Expert feedback? Assessing the role of test-taker reactions to a proficiency test for teachers of Japanese. *Melbourne Papers in Language Testing, 6*, 53–67.

Jia, Y. (2013). *Justifying the use of a second language oral test as an exit test in Hong Kong: An application of assessment use argument framework*. Los Angeles, CA: University of California.

Jenkins, J. (2007). *English as a lingua franca: Attitude and identity*. Oxford: Oxford University Press.

Jenkins, J., & Leung, C. (2013). English as a Lingua Franca. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1605–1616). Hoboken, NJ, USA: John Wiley & Sons, Inc. Retrieved, December 2015 from http://doi.wiley.com/10.1002/9781118411360.wbcla047

Jensen, C., & Hansen, C. (1995). The effect of prior knowledge on EAP listening-test performance. *Language Testing, 12*, 99–119.

Jensen, C., Hansen, C., Green, S. B., & Akey, T. (1997) An investigation of item difficulty incorporating the structure of listening tests: a hierarchical linear modeling analysis. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96* (pp. 151-164). Tampere: University of Jyvaskyla.

Jeon J. (2007). *A study of listening comprehension of academic lectures within the construction-integration model*. Unpublished Doctoral dissertation. Ohio State University, Columbus.

Jones, N., & Saville, N. (2009). European language policy: Assessment, learning, and the CEFR. *Annual Review of Applied Linguistics*, *29*, 51-63.

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching*, *44*, 137-166.

Junta de Andalucía. Consejería de Educación. (2005). *Plan de fomento del plurilingüismo. Una política lingüıstica para la sociedad andaluza*. Seville: Junta de Andalucía.

Kaftandjieva, F. (2004). *Standard-setting. Section B of the reference supplement to the preliminary version of the manual for relating language examinations to the Common European Framework of Reference for Languages:Learning, teaching and assessment*. Strasbourg: Council of Europe.

Kaftandjieva, F. (2009). Basket Procedure: The Breadbasket or the Basket Case of Standard Setting Methods? In N. Figueras, & J. Noijons (Eds.), *Linking to the CEFR levels: Research perspectives* (pp. 21-34). Arnhem: Cito, EALTA, Council of Europe.

Kaftandijeva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests. A comparative analysis of six recent methods with an application to tests of reading in EFL*. Cito, Arnhem: EALTA. Retrieved June 2017 from: www.ealta.eu.org/documentsresources/FK_second_doctorate.pdf

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-461.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4)*,* 319-342.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(2), 31–41.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2*(1), 135-170.

Kane, M. (2006). Content –related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 131-153) London: Lawrence Erlbaum Associates, Publishers.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1) 3 – 17.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kane, M. (2017). Using empirical results to validate performance standards BT. In S. Blömeke & J-E. Gustafsson (Eds.), *Standard setting in education: The Nordic countries in an international perspective* (pp. 11–29). Cham: Springer International Publishing.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5- 17.

Kanistra, V., & Harsch, C. (2017). *Using the Item Descriptor Matching method to enhance validity when aligning test to the CEFR.* 4th Meeting of the EALTA CEFR Special Interest Group. Retrieved January 2018 from: http://www.ealta.eu.org/events/SIG_CEFR_london2017/presentations/Presentations/1400-1530/CEFR%20SIG%20Voula%20and%20Claudia.pdf

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer-Verlag.

Kostin, I. (2004). *Exploring item characteristics that are related to the difficulty of TOEFL dialogue items*. TOEFL Research Report No. RR-79. Princeton, NJ: Educational Testing Service. Retrieved January 2016 from https://www.ets.org/Media/Research/pdf/RR-04-11.pdf

Koyama, D., Sun, A., & Ockey, G. J. (2016). The effects of item preview on video-based multiple-choice listening assessments. *Language Learning & Technology*, *20*(1), 148–165. Retrieved January 2016 from http://llt.msu.edu/issues/february2016/koyamasunockey.pdf

Kunnan, A. (2008). Towards a model of test evaluation: Using the test fairness and test context frameworks. In L. Taylor & C. Weir (Eds.), *Multilingualism and assessment* (pp.229–251).Cambridge: UCLES/Cambridge University Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Lee, Y.W., & Sawaki, Y. (2009). Cognitive diagnostic approaches to language assessment: An overview. *Language Assessment Quarterly*, *6*(3), 172–189.

Lee, H., & Winke, P. (2013). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing*, *30*(1), 99–123.

Li, Z. (2013a). The issues of construct definition and assessment authenticity in video-based listening comprehension tests: Using an argument-based validation approach. *International Journal of Language Studies*, *7*(2), 61–82.

Li, Z. (2013b). An empirical study of test-taking strategies in a video listening placement test. *Language and Communication Quarterly*, *1*(2), 2.

Liao, Y. (2009). *A construct validation study of the GEPT reading and listening sections: Re-examining the models of L2 reading and listening abilities and their relations to lexico-grammatical knowledge*. Unpublished doctoral dissertation, Teachers College, Columbia University.

Lim, G. S. (2013). Assessing English in Europe. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1700–1708). Hoboken, NJ, USA: John Wiley & Sons, Inc. Retrieved June 2015 from http://doi.wiley.com/

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85-106.

Linacre, J. M. (2005). Rasch Dichotomous Model vs. One-parameter Logistic Model (1PL 1-PL). *Rasch Measurement Transactions*, *19*(3), 1032. Retrieved January 2017 from http://www.rasch.org/rmt/rmt193h.htm

Linacre, J. M. (2008). *Rasch online forum*. See https://www.rasch.org/forum2008.htm

Linacre J.M. (2009). Unidimensional models in a multidimensional world. *Rasch Measurement Transactions*, *23*(2), 1209.

Linacre, J. M. (2013). *Rasch online forum*. See http://www.rasch.org/rmt/rmt221j.htm

Linacre, J. M. (2017a). *Winsteps® Rasch measurement computer program user's guide*. Beaverton, Oregon: Winsteps.com

Linacre, J. M. (2017b). *Facets computer program for many-facet Rasch measurement, version 3.80.0*. Beaverton, Oregon: Winsteps.com

Linn, R. L. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives*, *11*, 31.

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, *9*, 645-653.

Little, D. (2011). The Common European Framework of Reference for Languages: A research agenda. *Language Teaching*, *44*(3), 381-393.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

López Navas, M, D. (2012). *Comparative study of two foreign language examinations for university entry In England and Spain.* Unpublished doctoral thesis. Universitat Politechnica de Valencia. Retrieved July 2015 from https://riunet.upv.es/handle/10251/18055?show=full

López Navas, M, D. (2015). University entry English exam policy in Spain: A proposal for change. EALTA conference presentation. Copenhagen.

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing*, *24*(4), 489-515.

Luxia, Q. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22*(2), 142-173.

Lynch, T. (2010). *Teaching second language listening*. Oxford. Oxford University Press.

Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research. *Journal of English for Academic Purposes*, *10*(2), 79-88.

Macaro, E. (2006). Strategies for language learning and for language use: Revising the theoretical framework. *The Modern Language Journal, 90*, 320–337.

Macaro, E., Graham, S., & Vanderplank, R. (2007). A review of listening strategies: Focus on sources of knowledge and on success. In A. D. Cohen & E. Macaro (Eds.), *Language learner strategies: 30 years of research and practice* (pp. 165–185). Oxford: Oxford University Press.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum.

Magal-Royo, T., & Laborda, J. G. (2017). Multimodal interactivity in foreign language testing. In D. A. Dahl (Ed.), *Multimodal Interaction with W3C Standards* (pp. 351-365). Cham: Springer International Publishing.

Major, R, C. Fitzmaurice, Bunta, F., & Balasubramanian, C. (2002). The effects of nonna tive accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, *36*(2) 173-190.

Martín Monje, E. (2012): La nueva prueba oral en el examen de inglés de la Prueba de Acceso a la Universidad. Una propuesta metodológica. *Revista de Educación*, *357,* 143-161.

Martinez Saez, A., Sevilla Pavón, A., & Gimeno Sanz, A. (2009). Resultados encuesta profesores 2o bachillerato nueva prueba de lengua extranjera PAU LOGSE. Retrieved June 2015 from http://www.upv.es/ingles/documentos/informe.pdf

Martyniuk, W. (Ed.). (2010). *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's draft manual.* Cambridge, UK: Cambridge University Press.

Matthews, J., & Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System*, 52, 1–13.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. (2000). *Language testing.* Oxford: Oxford University Press.

McNamara, T. (2005). 21st century shibboleth: Language tests, identity and intergroup conflict. *Language Policy, 4*(4), 351–370.

McNamara, T. (2006) Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3(1),* 31–51.

McNamara, T. (2007). Language assessment in foreign language education: The struggle over constructs. *The Modern Language Journal*, *91*(2), 280-282.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell Publishing.

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly. 8*(2), 161-178.

Mecartty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehen-sion by foreign language learners of Spanish. *Applied Language Learning*, *11*(2), 323–348.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*, 241-256.

Miller, L. (2003). Developing listening skills with authentic materials. *ESL Magazine 6*(1), 16-19.

Ministry of Education, Culture & Sports. (2014). Estrategia para la Internacionalización de las Universidades Españolas 2015-2020. Retrieved January 2017 from https://sede.educacion.gob.es/publiventa/estrategia-para-la-internacionalizacion-de-las-universidades-espanolas-2015-2020/universidad/21475

Mislevy, R. J. (2007). Validity by design. *Educational Researcher*, *36*(8), 463–469.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003), A brief introduction to evidence-centered design. *ETS Research Report Series*.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6-20.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. In S. M. Downing., & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah: Lawrence Erlbaum Associates.

Mislevy, R., L. Steinberg., & R. Almond (2002). Design and analysis in task-based language assessment. *Language Testing 19*(4), 477–496.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–62.

Mitzel, H. C., Lewis, D. M., Patz, R.J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 249- 281). Mahwah, NJ: Lawrence Erlbaum Assoc.

Moe, E. (2009). Jack of more trades? Could standard-setting serve several functions? In N. Figueras & J. Noijons, (Eds.), *Linking to the CEFR levels: Research perspectives* (pp.131–138). Arnhem: Cito_EALTA.

Munby, J. (1978). *Communicative syllabus design*. Cambridge: Cambridge University Press.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227.

Nagao, J., Tadaki, T., Takeda, M., & Wicking, P. (2012). The attitudes of teachers and students towards a PET-based curriculum at a Japanese university. *Cambridge ESOL Research Notes, 47*, 27-36. UCLES. Retrieved January 2016 from http://www.cambridgeenglish.org/images/22669-rv-research-notes-47.pdf

Nguyen, T. N. H. (2008). *An investigation into the validity of two EFL (English as a foreign language) listening tests: IELTS and TOEFL iBT*. PhD thesis, Department of Linguistics and Applied Linguistics, The University of Melbourne. Retrieved March, 2013 from http://dtl.unimelb.edu.au.

Nissan, S., DeVincenzi, F., & Tang, K. L. (1996). An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension. *TOEFL Research Reports,51*.

Norris, J. M. (Ed.) (2002). Special issue: Task-based language assessment. *Language Testing*, *19*(4).

North, B. (2004). Europe's framework promotes language discussion, not directives, *Guardian Weekly*. Retrieved March 206 from Education Guardian. co.uk/tefl/story/0,,1191130,00.html

North, B. (2007). The CEFR Common Reference levels: Validated reference points and local strategies. In F. Goullier (Ed.), *Report on The Common European Framework of Reference for Languages (CEFR) and the development of language policies: Challenges and responsibilities* (pp.19-28). Strasbourg: Language Policy Division, Council of Europe.

North, B. (2014). Putting the Common European Framework of Reference to good use. *Language Teaching*, *47*, 228-249.

North, B., & Jones, N. (2009). *Relating language examinations to the Common European Framework of Reference for Languages. Further material on maintaining standards across languages, contexts and administrations by exploiting teacher judgment and IRT scaling*. Strasbourg: Council of Europe.

North, B., Martyniuk, W., & Panthier J. (2010). Introduction: The manual for relating language examinations to the Common European Framework of Reference for Languages in the context of the Council of Europe's work on language education. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 1–17). Cambridge, UK: Cambridge University Press.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Journal of Pragmatics*, *38*, 1928–1942.

Ockey, G. J., & French, R. (2014). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*. Retrieved January 2016 from http://applij.oxfordjournals.org/content/early/2014/11/03/applin.amu060.abstract

O'Dowd, R. (2015). *The training and accreditation of teachers for English medium instruction: A survey of European universities*. Retrieved January 2017 from http:// sgroup.be/sites/default/files/EMI%20Survey_Report_ODowd.pdf

OECD (2010). *PISA 2009 results: Executive summary*. Paris, OECD Publishing.

OECD (2015a). *Skills outlook 2015, Youth, skills and employability.* Paris, OECD Publishing.

OECD (2015b). *Education policy outlook 2015: Making reforms happen*. Paris, OECD Publishing.

Osada, N. (2001). What strategy do less proficient learners employ in listening comprehension? A reappraisal of bottom-up and top-down processing. *Journal of Pan-Pacific Association of Applied Linguistics*, *5*(1), 73–90.

O'Sullivan, B. (2008). *City & guilds communicator level IESOL examination (B2) CEFR linking project case study report*. City & Guilds Research Report. Retrieved September, 2013 from http://www.cityandguilds.com/documents/ind_general_learning_esol/CG_Communicator_Report_BOS.pdf

O' Sullivan, B. (2011). Introduction. In B. O'Sullivan (Ed.), *Language testing: Theories and practices* (pp.1-12). Oxford: Palgrave Macmillan.

O'Sullivan, B. (2015). *Technical report linking the APTIS reporting scales to the CEFR*. British Council. Retrieved January 2016 from: https://www.britishcouncil.org/sites/default/files/tech_003_barry_osullivan_linking_aptis_v4_single_pages_0.pdf

O'Sullivan, B., Weir, C.J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19*(1), 33-56.

Oxford, R. L. (2011). Strategies for learning a second or foreign language. *Language Teaching*, *44*(2), 167-180.

Oxford, R. L. (2017). *Teaching and researching language learning strategies: Self-regulation in context*. New York, NY: Routledge.

Pallant, J. (2007). *SPSS survival manual* (3rd ed.). PA: Open University Press.

Papageorgiou, S. (2009). *Setting performance standards in Europe: The judges' contribution to relating language examinations to the Common European Framework of Reference*. Berlin: Peter Lang.

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, *27*(2), 261–282.

Papageorgiou, S., Stevens, R., & Goodwin, S. (2012) The relative difficulty of dialogic and monologic input in a second-language listening comprehension test, *Language Assessment Quarterly*, *9*(4), 375-397.

Papageorgiou, S., & Tannenbaum, R. J. (2016). Situating standard setting within argument-based validity. *Language Assessment Quarterly*, *13*(2), 109–123.

Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish listening exam: Test usefulness evaluation. *Language Assessment Quarterly*, *7*(2), 137-159.

Pearson. (2010). *Aligning PTE academic test scores to the Common European Framework of Reference for Languages*. Retrieved January 2018 from https://pearsonpte.com/organizations/researchers/research-notes/

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing*, *20*(1), 26-56.

Phakiti, A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing*, *25*(2) 237-272.

Phakiti, A. (2013). Questionnaire development and analysis. In A. J. Kunnan (Ed.), *The companion to language assessment* (Chapter 74). Hoboken, NJ, USA: John Wiley & Sons, Inc. Retrieved January 2017 from http://doi.wiley.com/10.1002/9781118411360.wbcla068

Pizorn, K., & Moe, E. (2012) A validation study of the national assessment instruments for young English language learners in Norway and Slovenia. *CEPS Journal*, *2*(3), 75-96.

Pizorn, K., & Nagy, E. (2009). The politics of examination reform in central Europe. In J. C. Alderson (Ed.), *The politics of language education: Individuals and institutions* (pp. 185– 202). Bristol, England: Multilingual Matters.

Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development* (1. Ed). Berlin: Springer Science + Business Media, B.V.

Qi, L. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. *Assessment in Education, 14*(1), 51–74.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rea‐Dickins, P., & Poehner, M.E. (2011): Addressing issues of access and fairness in education through Dynamic Assessment. *Assessment in Education: Principles, Policy & Practice*, *18*(2), 95-97.

Read, J. (2002). The use of interactive input in EAP assessment. *Journal of English for Academic Purposes*, *1*, 105–109.

Reckase, M. D. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice*, *25*, 4–18.

Reckase, M. D. (2009). *Multidimensional item response theory (Vol. 150)*. New York, NY: Springer.

Reckase, M.D. (2010). NCME 2009 presidential address: What I think I know. *Educational Measurement: Issues and Practice, 29*(3), 3–7.

Revesz, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, *35*(1), 31-65.

Richards, J. C. (1983). Listening comprehension: Approach, design, and procedure. *TESOL Quarterly*, *17*, 219–240.

Richards, J. C. (2008). *Teaching listening and speaking. From theory to practice.* Cambridge: Cambridge University Press.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13.

Romero Garcia, J.A. (2003). *Impacto de la prueba de selectividad en el desarrollo y evaluación de las destrezas orales. Estudio de un caso y pilotaje de pruebas orales.* Unpublished PhD Thesis. University of Granada.

Rost, M. (2011). *Teaching and researching listening*. Harlow: Pearson Education Limited.

Rost, M. (2014). Listening in a multilingual world: The challenges of second language (L2) listening. *International Journal of Listening*, *28*(3), 131-148

Rubin, J. (1994). A review of second language listening comprehension research. *The Modern Language Journal*, *78*, 199-221.

Rubio, F. D., & Rodríguez, L. T. (2012). Estudio sobre prácticas docentes en evaluación de la lengua inglesa en la ESO. *Revista de Curriculum y Formacion del Professorado*, *16*(1), 295-316.

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, *1*, 185–216.

Sakai, H. (2009). Effect of repetition of exposure and proficiency level in L2 listening tests. *TESOL Quarterly*, *43*, 360–71.

Salisbury, K. (2005). *The edge of expertise: Towards an understanding of listening test item writing as professional practice*. Unpublished doctoral thesis, King's College London. Retrieved September, 2013 from https://kclpure.kcl.ac.uk/portal/files/2936144/419477.pdf

Sanz Sainz, I., & Fernández Álvarez, M. (2005). *The university entrance exam in Spain: Present situation and possible solutions*. Poster session presented at the EALTA conference, Voss (Norway).

Sarig, G. (1989). Testing meaning construction: can we do it fairly? *Language Testing*, *6*(1), 77- 94.

Saville, N. (2012). Applying a model for investigating the impact of language assessment within educational contexts: The Cambridge ESOL approach. *Research Notes, 50*, 4-8.

Saville-Troike, M. (2005). *Introducing second language acquisition*. Cambridge. Cambridge University Press.

Sawaki, Y., Kim, H. J., & Gentile, C. (2009). Q-Matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly*, *6*(3), 190–209.

Schedl, M. (2010). Background and goals of the TOEIC listening and reading test redesign project. *TOEIC Compendium*, *2*, 1-18.

Schumacker, R. E., & Smith Jr, E. V. (2007). Reliability. A Rasch perspective. *Educational and Psychological Measurement*, *67*(3), 394-409.

Seidlhofer, B. (2005). Key concepts in ELT - English as a lingua franca. *ELT Journal, 59*(4), 339-341.

Seigel, J. (2011). Readers respond: Thoughts on L2 listening pedagogy. *ELT Journal*, *65*(3), 318–321.

Sevilla-Pavón, A, Gimeno-Sanz, A., & García-Laborda, J. (2017). Actitudes docentes hacia los ejercicios de la Prueba de Acceso a la Universidad informatizada. *Educação e Pesquisa*, *43*(4), 1179-1200.

Shackleton, C. (2014). Measuring CEFR B1 listening proficiency: Cognitive validity in a listening test. In *Selected proceedings from the IATEFL TEA SIG 2014 conference (Granada*).

Shackleton, C. (forthcoming). Linking the University of Granada *CertAcles* listening test to the CEFR. *Revista de Educación.*

Shang, H. (2008). Listening strategy use and linguistic patterns in listening comprehension by EFL learners. *The International Journal of Listening*, *22*(1), 29–45.

Sherman, J. (1997). The effect of question preview in listening comprehension tests. *Language Testing*, *14*(2), 185–213.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing*, *18*(4), 373-391.

Shohamy, E. (2007). Tests as power tools: Looking back, looking forward. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, and C. E. Turner (eds.), *Language testing reconsidered* (pp. 141-152). Ottawa, ON: University of Ottawa Press.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. *Language Testing*, *8*(1), 23–40.

Sick, J. (2008). Rasch measurement in language education part 2: measurement scales and invariance. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *12*(2), 26-31.

Sick, J. (2010). Rasch measurement in language education Part 5: Assumptions and requirements of Rasch measurement. *Shiken: JALT Testing & Evaluation SIG Newsletter, 14*(2), 23 - 29.

Siegel, J. (2015). *Exploring listening strategy instruction through action research*. Basingstoke, UK: Palgrave Macmillan.

Simons, M., & Colpaert, J. (2014). *Time for a new CEFR? Recommendations from the field*. EALTA conference presentation 30th May 2014. Warwick.

Song, M. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, *25*(4), 435-464.

Staehr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, *31*, 577–607.

Stemler, S, E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4). Retrieved January 2017 from: http://PAREonline.net/getvn.asp?v=9&n=4

Suárez-Álvarez, J. ,González-Prieto, C. ,Fernández-Alonso, R, Gil, G., & Muñiz, J. (2014). Evaluación psicométrica de la expresión oral en inglés de las Pruebas de Acceso a la Universidad. *Revista de Educación*, *364*, 93-118.

Suvorov, R. (2009). Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format. In C. A. Chapelle, H. G. Jun & I. Katz (Eds.), *Developing and evaluating language learning materials* (pp. 53–68). Ames, IA: Iowa State University.

Suvorov, R. (2015). Interacting with visuals in L2 listening tests: an eye-tracking study. *ARAGs Research Reports Online; Vol. AR-A/2015/001*. London: The British Council.

Szabó, G., & Märcz, R. (2012). *'Interface' validity. Investigating the potential role of face validity in content validation*. Presentation EALTA conference. Innsbruck.

Tannenbaum, R, J., & Cho, Y. (2014). Critical Factors to Consider in Evaluating Standard-Setting Studies to Map Language Test Scores to Frameworks of Language Proficiency. *Language Assessment Quarterly, 11*(3), 233-249.

Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *ELT Journal*, *60*(1) (51-60).

Taylor, L. (2013). Introduction. In Geranpayeh, A & L.Taylor (Eds). *Examining Listening: Research and Practice in Assessing Second Language Listening* (pp.1-35). Cambridge: Cambridge University Press.

Taylor, L., & Garenpeyah, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes*, *10,* 89–101.

Taylor, L., & Garenpeyah, A. (2013). Conclusions and recommendations. In Geranpayeh, A & L.Taylor (Eds). *Examining Listening: Research and Practice in Assessing Second Language Listening* (pp. 332-334). Cambridge: Cambridge University Press.

Toulmin, S. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Toulmin, S. E. (2003). *The uses of argument* (updated edition). Cambridge, UK: Cambridge University Press.

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 M. Milanovic & C. J. Weir (Eds.), *Studies in language testing: Volume 11: Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 138-149). Cambridge, UK: Cambridge University Press.

Tragant, E., Miralpeix, I., Serrano, R., Pahissa, I., Navés, T., Gilabert, R., & Serra, N. (2014). Cómo se enseña inglés en un grupo de institutos donde se obtienen resultados destacables en a prueba de lengua inglesa en las PAU. *Revista de Educación*, *363*, 60-82.

Tsui, A. B., & Fullilove, J. (1998). Bottom-up or top-down processing as a discriminator of L2 listening performance. *Applied Linguistics*, *19*, 432–451.

Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. London and New York: Longman.

Vandergrift, L. (1997). The comprehension strategies of second language (French) listeners: A descriptive study. *Foreign Language Annals*, *30*(3), 387–409.

Vandergrift, L. (1999). Facilitating second language listening comprehension: Acquiring successful strategies. *ELT Journal, 53*(3): 168- 176.

Vandergrift, L. (2003). Orchestrating strategy use: Toward a model of the skilled second language listener. *Language Learning*, *53*(3), 463–496.

Vandergrift, L. (2004). Listening to learn or learning to listen. *Annual Review of Applied Linguistics*, *24*, 3-25.

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching, 40*, 191-210.

Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, *65*(2), 390–416.

Vandergrift, L., Goh, C., Mareschal, C., & Tafaghodtari, M. H. (2006). The Metacognitive Awareness Listening Questionnaire (MALQ): Development and validation. *Language Learning*, *56*(3), 431–462.

Vandergrift, L., & Goh, C. (2009). Teaching and testing listening comprehension . In M. Long & C. Doughty (Eds.), *The handbook of language teaching* (pp. 395 -411). Oxford: Blackwell.

Vandergrift, L., & Goh, C. (2012). *Teaching and learning second language listening: Metacognition in action*. New York: Routledge.

Vandergrift, L., & Tafaghodtari, M. H. (2010). Teaching students how to listen does make a difference: An empirical study. *Language Learning*, *60*, 470-497.

VanPatten, B. (Ed.) (2004). *Processing instruction: Theory, research, and commentary*. Mahwah, New Jersey: Lawrence Erlbaum.

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*, 457–479.

Wagner, E. (2002). Video listening tests: A pilot study. Teachers College, Columbia University. *Working Papers in TESOL & Applied Linguistics, 2*(1). Retrieved September, 2013 from http://journals.tc-library.org/index.php/tesol/article/view/7/8

Wagner, E. (2004). A construct validation study of the extended listening sections of the ECRE and MELAB. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *2*, 1-23.

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning and Technology*, *11*(1), 67- 86.

Wagner, E. (2010). The effect of the use of video texts on ESL listening test- taker performance. *Language Testing*, *27*(4), 493-513.

Wagner, E. (2013a). Assessing listening. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 47–63). Hoboken, NJ, USA: John Wiley & Sons, Inc. Retrieved from http://doi.wiley.com/10.1002/9781118411360.wbcla094

Wagner, E. (2013b). An investigation of how the channel of input and access to test questions affect L2 listening test performance, *Language Assessment Quarterly*, *10*(2), 178-195.

Wagner, E., & Toth, P. (2014). Teaching and testing L2 Spanish listening using scripted versus unscripted texts. *Foreign Language Annals*, 47, 404-422.

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing, 13*, 334–354.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System, 28*(4), 499-509.

Wall, D. (2013). Washback. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 93-106). Abingdon, UK: Routledge.

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, MS-34. Princeton, NJ: Educational Testing Service.

Walsh, S. (2011). *Exploring classroom discourse language in action*. Oxon: Routledge.

Wang, H., Choi, I., Schmidgall, J., & Bachman, L. F. (2012). Review of Pearson Test of English Academic: Building an assessment use argument *Language Testing 29*, 603-619.

Wang, J. (2010). *A study of the role of the 'teacher factor' in washback*. Unpublished doctoral thesis. McGill University. Retrieved December 2017 from http://digitool.library.mcgill.ca/webclient/StreamGate?folder_id=0&dvs=1520188661667~871

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing*, *13*(3), 318–33.

Watts, F., & Garcia Carbonell, A. (1999). Control de calidad en la calificación de la prueba de inglés de selectividad. *Aula Abierta*, *73*, 173-190.

Watts, F., & Garcia Carbonell, A. (2005). Control de calidad en la calificación de la prueba de lengua inglesa de selectividad. In H. Herrera Soler & J. García Laborda, (Eds.), *Estudios y criterios para una selectividad de calidad en el examen de inglés* (pp. 99-115). Valencia: UPV.

Weir, C. J. (1993). *Understanding and developing language tests*. New York: Prentice Hall.

Weir, C. J. (2005a). Limitations of the council of Europe's framework of reference (CEFR) in developing comparable examinations and tests. *Language Testing*, *22*(3), 281–300.

Weir, C. J. (2005b). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Weir, C. J. (2010). *1980-2010 Testing communicative language use: A brief overview Language testing forum. 30 years of ltf: Issues in language testing revisited.* Lancaster University. Retrieved September, 2013 from http://www.beds.ac.uk/__data/assets/pdf_file/0011/83855/Weir_LTF2010.pdf

Weir, C. J. (2013). Conclusions and recommendations. In C. J. Weir, I. Vidakovic & E. Galaczi. *Measured constructs: A history of the constructs underlying Cambridge English language (ESOL) examinations 1913-2012* (pp.420-444). Cambridge: Cambridge University Press.

Wolfe, E. W., & Smith, J. E. (2007a). Instrument development tools and activities for measure validation using Rasch models: part II--validation activities. *Journal of Applied Measurement*, *8*(2), 204-234.

Wolfe, E. W., & Smith, J. E. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part I-instrument development tools. *Journal of Applied Measurement*, *8*(1), 97-123.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement: Transactions of the Rasch Measurement SIG, 8*(3), 370.

Wright, B. D., & Mok, M. M. (2004). An overview of the family of Rasch measurement models. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp.1-24). Maple Grove MN: JAM Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement Solutions.

Wu, Y. (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, *15*, 21–44.

Wu, Y. (2014). *Ensuring quality and fairness in the Asian EFL context: Challenges and opportunities*. ALTE conference. Taipei, Taiwan. Retrieved December 2017 from http://events.cambridgeenglish.org/alte-2014/docs/presentations/alte2014-jessica-wu.pdf

Xi, X. (2008). Methods of test validation. In E. Shohamy (Ed.), *Language Testing and Assessment, Volume 7* of *Encyclopedia of Language and Education* (pp. 177-196). Dordrecht: Springer Netherland.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing 27*(2), 147–170.

Yanagawa, K. (2012). *A partial validation of the contextual validity of the centre listening test in Japan*. Unpublished PhD thesis, University of Bedfordshire. Retrieved January 2016 from http://uobrep.openrepository.com/uobrep/bitstream/10547/267493/1/Yanagawa.pdf

Yanagawa, K., & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System*, *36*, 107-122.

Yang, P. (2000) *Effects of test-wiseness upon performance on the test of English as a foreign language*, unpublished PhD dissertation, University of Alberta, Edmonton, CN. Retrieved January, 2016 from http://www.nlc-bnc.ca/obj/s4/f2/dsk1/tape3/PQDD_0009/NQ59700.pdf

Yi'an, W. (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple choice task. *Language Testing*, *15*(1), 21–44.

Zhang, L., Goh, C., & Kunnan, A. (2014) Analysis of Test Takers' Metacognitive and Cognitive Strategy Use and EFL Reading Test Performance: A Multi-Sample SEM Approach, *Language Assessment Quarterly*, *11*(1), 76-102.

Zhang, Yan. (2012). The impact of listening strategy on listening comprehension. *Theory and Practice in Language Studies, 2*(3), 625- 629.

Zhao, Y. (1997). The effects of listeners' control of speech rate on second language comprehension. *Applied Linguistics, 18*(1), 49– 68.

Zheng, Y., & De Jong, J. (2011). *Establishing construct and concurrent validity of Pearson Test of English academic*. Retrieved January 2018 from https://pearsonpte.com/organizations/researchers/research-notes/

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, *20*(2), 79–87.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp.65-82). Charlotte, NC, US: IAP Information Age Publishing.

APPENDIX 1

As part of my Doctoral studies on the doctoral program *Lenguas, Textos y Contextos*, at the University of Granada. I am carrying a validity study of a CEFR B2 listening test. This will involve test results, a questionnaire and an analysis of how students complete a listening test as well as interviews with students and consultation with experts.

I have approached you because I am interested in discovering the way non-native speakers do listening tasks. I would be very grateful if you would agree to take part.

I will ask you to complete a listening test and then we will look at your test paper and I will ask you to talk about the things you did in order to complete the questions. I will record you whilst you explain the listening process when doing the test. I am going to transcribe portions of the recordings, and will use the information as part of my study.

You are free to withdraw from the study at any time. At every stage, your name will remain confidential. The data will be kept securely and will be used for academic purposes only.

If you have any queries about the study, please feel free to contact me at csarah@ugr.es.

Signed

Caroline Shackleton

csarah@urg.es

Consent Form

Project title: An initial validity argument for a new B2 CEFR related baccalaureate listening test.

1. I have read and had explained to me by **Caroline Shackleton** the Information Sheet relating to this project.

2. I have had explained to me the purposes of the project and what will be required of me, and any questions have been answered to my satisfaction. I agree to the arrangements described in the Information Sheet in so far as they relate to my participation.

3. I understand that my participation is entirely voluntary and that I have the right to withdraw from the project any time.

4. I have received a copy of this Consent Form and of the accompanying Information Sheet.

Name:

Signed:

Date:

APPENDIX 2: CEFR B2 listening test and questionnaire

B2 Listening Test

I would be grateful if you could complete this test and questionnaire for research analysis for my doctorate studies at the Department of Linguistics here at Granada University. This information is for academic research purposes only and all reported data will be anonymous; names will only be collected in order to give feedback to students about their performance on the test and will not be kept on record. This study is completely voluntary and can be withdrawn from at any time. For further information about this study, please contact me, Caroline Shackleton at csarah@ugr.es, or my supervisor Tony Harris at tharris@ugr.es

Please tick box to agree to taking part in this study ☐

Name: _____

Many thanks for your help.

Background Information:

Please mark the following with a cross.

SEX Male ☐ Female ☐

AGE Under 18 ☐ 18-21 ☐ 21-25 ☐ 26-35 ☐ Over 35 ☐

What level do you think you have in **listening** comprehension? A2 ☐ B1 ☐ B2 ☐ C1 or higher ☐

Do you have an official accreditation of your language proficiency?
Which? e.g. Cambridge FCE, PET, ACLES B1 etc.

_____

Secondary School:

Complete the table with a number 1 - 4:

1 = I completely disagree.
2 = I disagree
3 = I quite agree
4 = I completely agree

At school we did sufficient listening practice in class	
At school we didn't do much listening practise because listening is not on the *Selectividad* exam	
Listening is important for learning a language	
I think there should be a listening section on the *Selectividad* exam	
I practiced listening outside school e.g. I watched series in English, I had English friends, I studied in a private academy	

In my *English Selectividad* exam I scored _____

TASK 1

Opinions about sports

Listen to Simon giving his opinions about sport. Choose the correct answer (**1-7**) for each question (**A-I**). There is **one** extra question that you do not need to use. There is an example (0) at the beginning.

Should……..?
A sports people who take drugs be banned for life?
B TV replays be used in football games?
C women's boxing competitions be an Olympic sport?
D foreigners play in sports teams in your country?
E yoga be an Olympic sport?
F TV channels compete to buy the most popular sports?
G more competitive sports be shown on TV?
H sports stars be allowed to do publicity?
I foreigners be able to buy sports teams in your country?

Answer	
0	***E***
Q1	
Q2	
Q3	
Q4	
Q5	
Q6	
Q7	

TASK 2

Moving to the USA

Listen to Jean talking about moving from Mexico to Charlottesville in the USA. Choose the correct answer (**A, B, C** or **D**) for questions **1-8**. There is an example (0) at the beginning.

First you have 45 seconds to study the questions. Then you will hear the recording twice.

Glossary
UVA = University of Virginia

0. Jean's main aim in the USA is to _____

A. do his doctorate studies
B. find a good job
C. learn to speak English
D. finish his degree

Q1. Finding somewhere to live was _____

A. almost impossible
B. relatively easy
C. time consuming
D. a learning process

Q2. After arriving at the airport he _____

A. drove to his new flat
B. was collected by a teacher
C. went to the university
D. contacted a flat agency

Q3. Arriving in the USA was hard as he didn't know _____

A. where to buy things
B. what food to buy
C. any English
D. many people

Q4. He had to learn about _____

A. the education system
B. shop opening times
C. how to get around
D. how to write in English

Q5. Compared to Mexico City, he found Charlottesville to be _____

A. interesting
B. relaxing
C. expensive
D. conservative

Q6. He feels accepted in Charlottesville because _____

A. of his job expertise
B. of his university position
C. Americans respect people
D. Americans are polite

Q7. He finds that it's difficult for people to _____

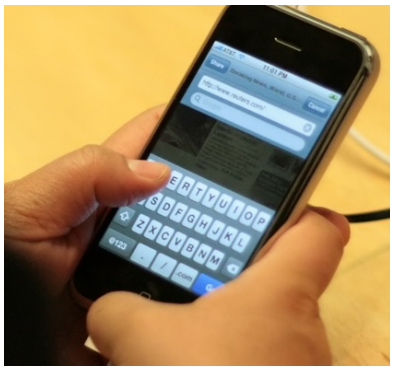
A. get to know him
B. understand his accent
C. work out his nationality
D. say his name properly

Q8. He wants Americans to know _____

A. his country’s history
B. his country’s problems
C. he is from Mexico
D. he is from a big city

TASK 3

Text messaging



Listen to a radio interview with Dr. Caroline Tagg about text messaging. Choose the correct answer (**A, B, C** or **D**) for questions **1 - 8**. There is an example (0) at the beginning.

First you have 45 seconds to study the questions. Then you will hear the recording twice.

0. Caroline's job involves _____

A. solving linguistic problems
B. researching language use
C. the newest technology
D. teaching English

Q1. When Caroline began her investigation, Facebook wasn't _____

A. available to many people
B. used outside the USA
C. an option for research
D. free for its users

Q2. In the beginning she had _____

A. an easy time collecting her data
B. difficulties understanding text messages
C. to get data from people close to her
D. to obtain written copyright permission

Q3. Most of her data was collected _____

A. from existing corpus
B. during the third year
C. by her research group
D. over a long time

Q4. She thinks that the age of the people who participated _____

A. had a lot of influence on her results
B. was appropriate for the data she needed
C. represented how young people speak
D. affected the methods of data collection

Q5. Her results showed that people _____

A. sent long messages
B. used abbreviations
C. expressed dislike for text messages
D. included lots of different *emoticons*

Q6. Investigation shows that people like to _____

A. share messages with many people
B. state their feelings correctly
C. write their messages quickly
D. use abbreviations and idioms

Q7. People normally use emoticons _____

A. to show negative feelings
B. in short messages
C. to express happiness
D. from a small selection

Q8. Caroline says that interestingly, people use emoticons to _____

A. change the look of their messages
B. change the tone of their messages
C. make their messages fashionable
D. make their messages more secret

TASK 4

Geography trip

Listen to an informative talk about a Geography trip. First you have 45 seconds to study the notes below. Then you will hear the recording twice. Listen and complete the notes (1-10) in a maximum of ***THREE*** words. There is one example (0) at the beginning.

GEOGRAPY TRIP

This talk about ___**(0)** *camping procedure*__

Financial details:
Deposit already paid
Students must pay Total Cost of the trip by **(Q1)** _____
Check website for bank transfer info

Additional safety info:
Insurance for all activities: (see website)
Students must tell the company about their **(Q2)** _____

Camping details:
Tents
Limited storage, only overnight bags taken
Check *information pack* for details of **(Q3)** _____

Informal meeting on arrival, giving info about facilities etc.
Everybody working at the camp wearing **(Q4)** _____

Entry/Exits: need to write name at **(Q5)** _____

Code of Conduct gives more information about **(Q6)** _____

Study Schedule:
Excursion details in program
Day 2: Go **(Q7)** _____

Data for project to be collected from **(Q8)** _____

Need to compare collected info with results from **(Q9)** _____

Use different methods: Final reports might include **(Q10)** _____
Results should be explained in analysis section!

Please answer the following questions about the listening comprehension test (tasks 1- 4)

1. **How authentic did you find the audios used in the tasks?**

		Not authenic	Not very authenic	Quite authenic	Very authenic
1.1	Questions about sport				
1.2	Moving to the USA				
1.3	Text messaging				
1.4	Geography trip				

2. **How difficult did you find the listening audios?**

		Not difficult	Not very difficult	Quite difficult	Very difficult
2.1	Questions about sport				
2.2	Moving to the USA				
2.3	Text messaging				
2.4	Geography trip				

3. **How difficult did you find the questions?**

		Not difficult	Not very difficult	Quite difficult	Very difficult
3.1	Questions about sport				
3.2	Moving to the USA				
3.3	Text messaging				
3.4	Geography trip				

4. **How familiar did you find the topics used in the tasks?**

		Not familiar	Not very familiar	Quite familiar	Very familiar
4.1	Questions about sport				
4.2	Moving to the USA				
4.3	Text messaging				
4.4	Geography trip				

5. **How suitable did you find the amount of time to:**

		Too Little	Not quite enough	Enough	Too much
5.1	Read the questions				
5.2	Answer the questions				

6. **How do you feel about the test as a fair measure of your English listening ability?**

		Not Satisfied	Not very satisfied	Quite satisfied	Very satisfied
6.1	Questions about sport				
6.2	Moving to the USA				
6.3	Text messaging				
6.4	Geography trip				

7. Complete the table with a number 1 - 4:

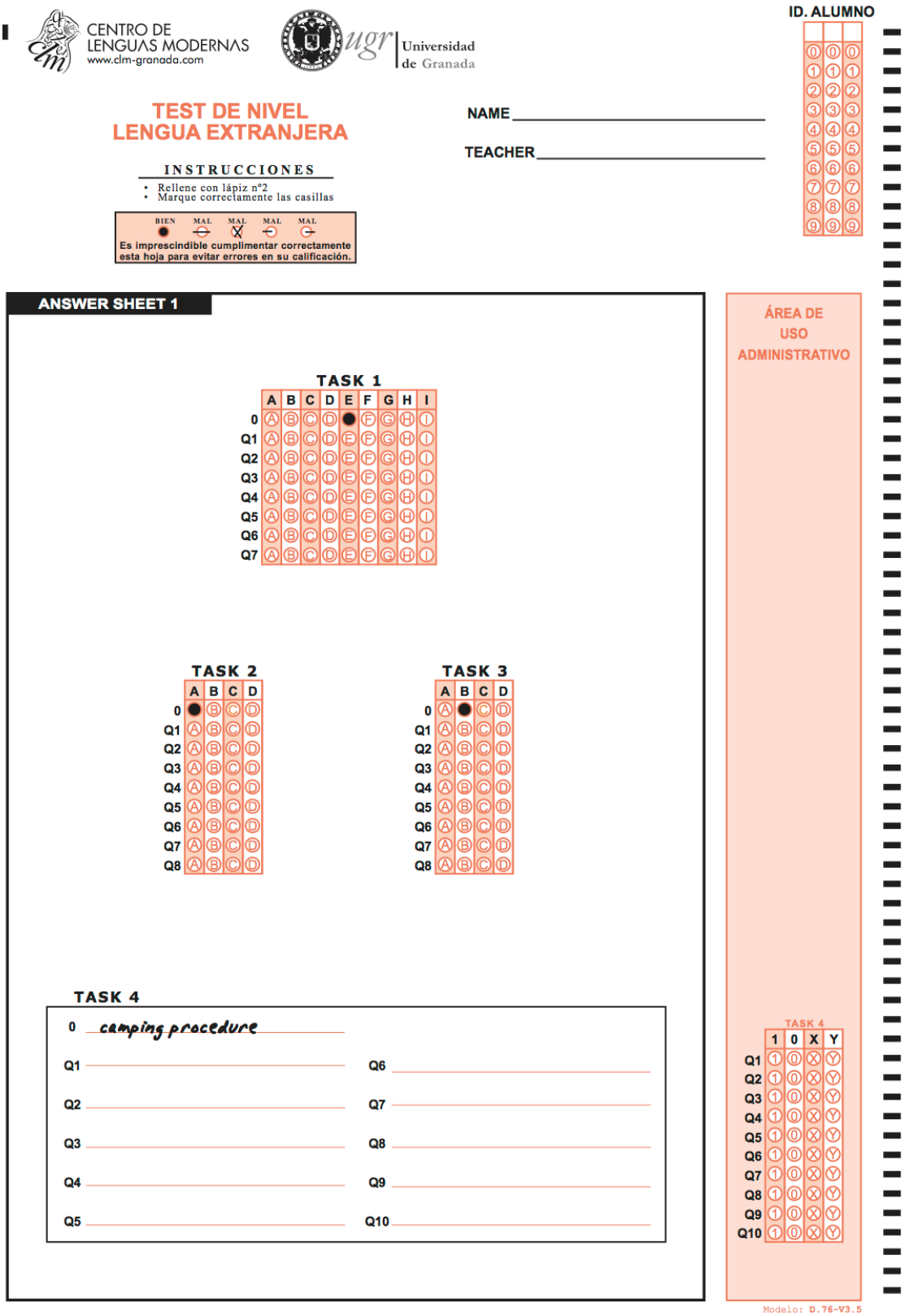
1 = I completely disagree
2 = I disagree
3 = I quite agree
4 = I completely agree

	Questions about sport	Moving to the USA	Test messaging	Geography trip
The instructions for the task were clear				
The quality of the recording was good				
I recognised the accent of the speaker(s)				
The speaker spoke at normal speed				
The topic was typical of those I studied at school				
The audio was similar to those I studied at school				
I could understand the main ideas of the audio				
In general I could follow the audio				
I guessed the answer				

I used the words I understood to guess the meaning of words I didn't understand	
As I listen, I compared what I understand with what I know about the topic	
Before I started to listen I had a plan in my head for how I was going to listen	
I wish we had used authentic audios at school	
It would be useful to learn strategies to understand authentic audios at school	
I think my listening ability would be better if we had practised more at school	

Any other comments you would like to add?

APPENDIX 3 — Answer sheet



CENTRO DE LENGUAS MODERNAS
www.clm-granada.com

ugr Universidad de Granada

ID. ALUMNO

TEST DE NIVEL
LENGUA EXTRANJERA

NAME _____

TEACHER _____

INSTRUCCIONES

- Rellene con lápiz nº2
- Marque correctamente las casillas

BIEN MAL MAL MAL MAL

Es imprescindible cumplimentar correctamente esta hoja para evitar errores en su calificación.

ANSWER SHEET 1

ÁREA DE USO ADMINISTRATIVO

TASK 1

	A	B	C	D	E	F	G	H	I
0					●				
Q1									
Q2									
Q3									
Q4									
Q5									
Q6									
Q7									

TASK 2

	A	B	C	D
0	●			
Q1				
Q2				
Q3				
Q4				
Q5				
Q6				
Q7				
Q8				

TASK 3

	A	B	C	D
0		●		
Q1				
Q2				
Q3				
Q4				
Q5				
Q6				
Q7				
Q8				

TASK 4

0 camping procedure

Q1 _____ Q6 _____

Q2 _____ Q7 _____

Q3 _____ Q8 _____

Q4 _____ Q9 _____

Q5 _____ Q10 _____

TASK 4

	1	0	X	Y
Q1				
Q2				
Q3				
Q4				
Q5				
Q6				
Q7				
Q8				
Q9				
Q10				

Modelo: D.76-V3.5

APPENDIX 4: Final CEFR B2 BFE listening test

TASK 1

Opinions about sports

Listen to Simon answering questions about sport. Choose the correct answer (**1-6**) for each question (**A-I**). There is **one** extra question that you do not need to use. There is an example (0) at the beginning.

First you have 45 seconds to study the questions. Then you will hear the recording twice.

Should……..?
A sports people who take drugs be banned for life?
B TV replays be used in football games?
C women's boxing competitions be an Olympic sport?
D foreigners play in sports teams in your country?
E yoga be an Olympic sport?
F more competitive sports be shown on TV?
G sports stars be allowed to do publicity?
H foreigners be able to buy sports teams in your country?

Answer	
0	***E***
Q1	
Q2	
Q3	
Q4	
Q5	
Q6	

TASK 2

Moving to the USA

Listen to Jean talking about moving from Mexico to Charlottesville in the USA. Choose the correct answer (**A, B, C** or **D**) for questions **7- 13**. There is an example (0) at the beginning.

First you have 45 seconds to study the questions. Then you will hear the recording twice.

Glossary
UVA = University of Virginia

0. Jean's main aim in the USA is to _____

A. do his doctorate studies
B. find a good job
C. learn to speak English
D. finish his degree

Q7. Finding somewhere to live was _____

A. almost impossible
B. relatively easy
C. time consuming
D. a learning process

Q8. After arriving at the airport he _____

A. drove to his new flat
B. was collected by a teacher
C. went to the university
D. contacted a flat agency

Q9. Arriving in the USA was hard as he didn't know _____

A. where to buy things
B. what food to buy
C. any English
D. many people

Q10. He had to learn about _____

A. the education system
B. shop opening times
C. how to get around
D. how to write in English

Q11. Compared to Mexico City, he found Charlottesville to be _____

A. interesting
B. relaxing
C. expensive
D. conservative

Q12. He feels accepted in Charlottesville because _____

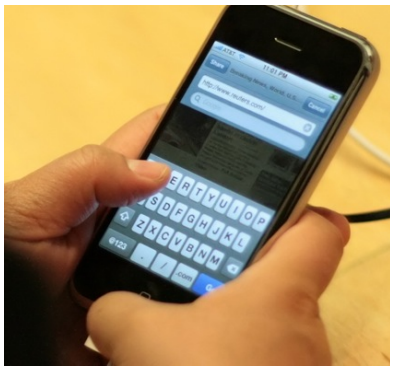
A. of his job expertise
B. of his university position
C. Americans respect people
D. Americans are polite

Q13. He finds that it's difficult for people to _____

A. get to know him
B. understand his accent
C. work out his nationality
D. say his name properly

TASK 3

Text messaging



Listen to a radio interview with Dr. Caroline Tagg about text messaging. Choose the correct answer (**A, B, C** or **D**) for questions **14 - 20**. There is an example (0) at the beginning.

First you have 45 seconds to study the questions. Then you will hear the recording twice.

0. Caroline's job involves _____

A. solving linguistic problems
B. researching language use
C. the newest technology
D. teaching English

Q14. When Caroline began her investigation, Facebook wasn't _____

A. available to many people
B. used outside the USA
C. an option for research
D. free for its users

Q15. In the beginning she had _____

A. an easy time collecting her data
B. difficulties understanding text messages
C. to get data from people close to her
D. to obtain written copyright permission

Q16. She thinks that the age of the people who participated _____

A. had a lot of influence on her results
B. was appropriate for the data she needed
C. represented how young people speak
D. affected the methods of data collection

Q17. Her results showed that people _____

A. sent long messages
B. used abbreviations
C. expressed dislike for text messages
D. included lots of different *emoticons*

Q18. Investigation shows that people like to _____

A. share messages with many people
B. state their feelings correctly
C. write their messages quickly
D. use abbreviations and idioms

Q19. People normally use emoticons _____

A. to show negative feelings
B. in short messages
C. to express happiness
D. from a small selection

Q20. Caroline says that interestingly, people use emoticons to _____

A. change the look of their messages
B. change the tone of their messages
C. make their messages fashionable
D. make their messages more secret

TASK 4

Geography trip

Listen to an informative talk about a geography trip. Listen and answer the questions **22-28** in a maximum of *THREE* words. There is an example (0) at the beginning.

First you have 45 seconds to study the questions. Then you will hear the recording twice.

GEOGRAPY TRIP

This talk about ___**(0)** *camping procedure*___

Financial details:
Deposit already paid
Students must pay Total Cost of the trip by **(Q21)** _____
Check website for bank transfer info

Additional safety info:
Insurance for all activities: (see website)
Students must tell the company about their **(Q22)** _____

Camping details:
Tents
Limited storage, only overnight bags taken
Check *information pack* for details of **(Q23)** _____

Informal meeting on arrival, giving info about facilities etc.
Everybody working at the camp wearing **(Q24)** _____

Entry/Exits: need to write name at **(Q25)** gu_____

Code of Conduct gives more information about **(Q26)** _____

Study Schedule:
Excursion details in program
Day 2: Go **(Q27)** _____

Need to compare collected info with results from **(Q28)** _____