

Sheffield Hallam University

Vocabulary and listening in English among L1-Spanish learners: a longitudinal study

AOIZ PINILLOS, Martin <<http://orcid.org/0000-0001-6016-1832>>

Available from the Sheffield Hallam University Research Archive (SHURA) at:

<http://shura.shu.ac.uk/28911/>

A Sheffield Hallam University thesis

This thesis is protected by copyright which belongs to the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Please visit <http://shura.shu.ac.uk/28911/> and <http://shura.shu.ac.uk/information.html> for further details about copyright and re-use permissions.

**VOCABULARY AND LISTENING IN ENGLISH
AMONG L1-SPANISH LEARNERS:
A LONGITUDINAL STUDY**

Martin Aoiz Pinillos
BA, MA

Thesis submitted to Sheffield Hallam University
for the degree of Doctor in Education
January 2021

I hereby declare that:

1. I have not been enrolled for another award of the University, or other academic or professional organisation, whilst undertaking my research degree.
2. None of the material contained in the thesis has been used in any other submission for an academic award.
3. I am aware of and understand the University's policy on plagiarism and certify that this thesis is my own work. The use of all published or other sources of material consulted have been properly and fully acknowledged.
4. The work undertaken towards the thesis has been conducted in accordance with the SHU Principles of Integrity in Research and the SHU Research Ethics Policy.
5. The word count of the thesis is 59,988

Name	<i>Martín Aoiz Pinillos</i>
Date	<i>January 2021</i>
Award	<i>EdD</i>
Faculty	<i>Social Sciences and Humanities</i>
Director(s) of Studies	<i>Dr Nicholas Moore Dr Jane L. Morgan</i>

To Ana

ABSTRACT

Second language listening causes situations of stress and negative perceptions among learners and teachers. Research has suggested that L2 listening and vocabulary knowledge are related. However, this relationship has been barely explored, and in most cases with inadequate instruments. This thesis is an attempt to bridge those gaps by examining the contribution of the language learners' vocabulary size to their listening ability.

A bilingual multiple-choice vocabulary test, based on the official vocabulary list in a standardized language exam, was created to assess the vocabulary size of L2-English learners. Its 81 items were delivered first orally, and then in writing. The ability to comprehend aural texts was assessed through the listening paper in the same standardized examination. 284 language learners took the vocabulary and listening tests. After an observation period of 35 weeks, the study participants were given the same tests. Both datasets were analyzed with the Rasch model to determine the participants' abilities and the item difficulties.

Evidence from data analyses supported the following findings:

- 1) A strong and positive relationship exists between L2 vocabulary knowledge and listening comprehension.
- 2) Aural and written vocabulary knowledge are two dimensions that should be assessed and investigated separately, particularly in relation to listening comprehension.
- 3) Aural vocabulary knowledge is a better predictor of listening comprehension than written vocabulary knowledge, especially among language learners with comparatively weaker listening skills.
- 4) Knowing 71.71% of the words featured in a listening comprehension test is sufficient to answer 72% of its questions correctly.
- 5) Language learners increase their aural and written vocabulary size, and improve their listening ability after attending classes for about 35 weeks.

This improvement is particularly acute among lower-level learners.

Based on these results, L2 learners, teachers and researchers should focus more on the aural form of words to improve listening comprehension.

TABLE OF CONTENTS

ABSTRACT	iv
LIST OF ACRONYMS	ix
LIST OF FIGURES	xi
LIST OF TABLES	xii
CHAPTER 1 – INTRODUCTION	1
1.1 – RESEARCH CONTEXT	2
1.2 – VOCABULARY AND LISTENING – GENERAL INTRODUCTION	4
1.3 – BRIDGING GAPS	8
1.4 – CHAPTER SUMMARY	11
CHAPTER 2 – LITERATURE REVIEW	13
2.1 – INTRODUCTION	14
2.1.1 The importance of listening	14
2.1.2 Listening in language learning	15
2.1.3 Listening in language teaching and research	19
2.2 – UNDERSTANDING LISTENING	22
2.2.1 Listening as a process: The ‘teaching approach’	22
2.2.2 Listening processes	25
2.2.3 Listening in the present study	29
2.2.3.1 <i>Bottom-up or Top-down?</i>	29
2.2.3.2 <i>Bottom-up and Top-down in weak and strong listeners</i>	32
2.3 – VOCABULARY AND LISTENING	36
2.3.1 Inadequate vocabulary size and listening performance	37
2.3.2 Positive effects of adequate vocabulary size on listening performance	40
2.3.3 Vocabulary and Listening among L1-Spanish Learners	41
2.4 – KNOWING A WORD – TEXT PROFILING	44
2.4.1 Lexical units in vocabulary studies	44
2.4.2 Mismatches in word families – Single-word items	47
2.4.2.1 <i>Polysemy</i>	48
2.4.2.2 <i>Homoforms</i>	49
2.4.2.3 <i>Proper nouns</i>	51
2.4.3 Mismatches in word families - Multiword items	52
2.4.3.1 <i>Multiword nouns</i>	52
2.4.3.2 <i>Multiword verbs</i>	54
2.4.3.3 <i>Formulaic language</i>	55

2.4.4 Reasons for the mismatches	57
2.5 – ESTIMATING VOCABULARY SIZE IN L2	60
2.5.1 Vocabulary Size, Frequency and Lexical Coverage	60
2.5.2 Vocabulary Testing and Listening Comprehension	63
2.5.2.1 <i>Unsuccessful attempts to assess the aural vocabulary size</i>	65
2.5.2.2 <i>Listening Vocabulary Size Test (LVST)</i>	68
2.6 – BRIDGING GAPS	72
2.7 – CHAPTER SUMMARY	75
CHAPTER 3 – METHODOLOGY AND METHODS	79
3.1 – METHODOLOGY	80
3.1.1 Ontological Assumptions and Epistemological Approach	80
3.1.2 Ontology and Epistemology in the present Research Study	81
3.1.2.1 <i>Vocabulary and Listening: Homogeneity, Reliability and Generalizability</i>	83
3.1.2.2 <i>Vocabulary and Listening: Ecological Validity</i>	85
3.1.2.3 <i>Vocabulary and Listening: Research Questions</i>	88
3.1.2.4 <i>Vocabulary and Listening: Research Constructs</i>	92
3.1.2.5 <i>Vocabulary and Listening: Use of the Rasch Model</i>	93
3.2 – METHODS	98
3.2.1 Vocabulary Test – Preliminary Issues	99
3.2.2 Cambridge English: Preliminary and Preliminary for Schools (PET) – Vocabulary List	100
3.2.2.1 <i>Preparing the PET Vocabulary List</i>	101
3.2.2.2 <i>The PET Vocabulary List and the BNC-COCA 1-25k</i>	101
3.2.3 Creation of a Vocabulary Test based on the PET Vocabulary List	105
3.2.4 Cambridge English: Preliminary – Adapting the Listening Paper	108
3.2.5 Preliminary Study – Refining the Vocabulary Tests	111
3.2.5.1 <i>Data collection</i>	111
3.2.5.2 <i>Data Analysis: Descriptive Statistics, Reliability, Separation</i>	115
3.2.5.3 <i>Data Analysis: Fit Statistics</i>	119
3.2.5.4 <i>Data Analysis: Descriptive Statistics and Item Difficulty</i>	125
3.2.5.5 <i>Data Analysis: Instrument Validity</i>	129
3.2.5.6 <i>Conclusions</i>	133
3.2.6 Main Study – First Data Collection – October 2019	133

3.2.6.1 <i>Descriptive statistics, reliability, and separation</i>	134
3.2.6.2 <i>Data Quality Analysis</i>	135
3.2.6.3 <i>Effect of misfit on data quality</i>	136
3.2.6.4 <i>Conclusions</i>	140
3.2.7 <i>Main Study – Second Data Collection – June 2020</i>	140
3.2.7.1 <i>Adapting to a new research environment caused by COVID-19</i>	141
3.2.7.2 <i>Data Quality Analysis</i>	141
3.2.7.3 <i>Effect of misfit on data quality</i>	144
3.2.7.4 <i>Conclusions</i>	146
3.3 – CHAPTER SUMMARY	148
CHAPTER 4 – DATA ANALYSIS AND RESULTS	151
4.1 – DATA ANALYSIS – DESCRIPTIVE STATISTICS and ITEM DIFFICULTY	152
4.1.1 First Dataset – October 2019	152
4.1.2 Second Dataset – June 2020	158
4.1.3 Conclusions	163
4.2 – RESEARCH QUESTION 1: <i>How much of the listening performance in an exam might be attributed to knowing the words in a vocabulary list?</i>	165
4.2.1 Data from October 2019	165
4.2.2 Data from June 2020	170
4.2.3 Conclusions	172
4.3 – RESEARCH QUESTION 2: <i>How much lexical coverage of a spoken text does a learner need to achieve comprehension in a listening test?</i>	174
4.3.1 Data from October 2019	174
4.3.2 Data from June 2020	178
4.3.3 Conclusions	180
4.4 – RESEARCH QUESTION 3: <i>How similar are the scores in vocabulary size tests based on recognising either the aural or the written form of words?</i>	182
4.4.1 Data analysis	182
4.4.2 Conclusions	186
4.5 – RESEARCH QUESTION 4: <i>How does the relationship between lexical knowledge and listening performance evolve over time?</i>	188
4.5.1 Data analysis	188
4.5.2 Conclusions	193
4.6 – CHAPTER SUMMARY	196
CHAPTER 5 – GENERAL DISCUSSION	199
5.1 – RESEARCH QUESTION 1: <i>How much of the listening performance in an exam might be attributed to knowing the words in a vocabulary list?</i>	200
5.2 – RESEARCH QUESTION 2: <i>How much lexical coverage of a spoken text does a learner need to achieve comprehension in a listening test?</i>	209

5.3 – RESEARCH QUESTION 3: <i>How similar are the scores in vocabulary size tests based on recognising either the aural or the written form of words?</i>	215
5.4 – RESEARCH QUESTION 4: <i>How does the relationship between lexical knowledge and listening performance evolve over time?</i>	220
5.5 – CONTEXTUALIZATION OF RESULTS	224
5.6 – CHAPTER SUMMARY	228
CHAPTER 6 – IMPLICATIONS, CONCLUSIONS and LIMITATIONS	231
6.1 – LIMITATIONS OF THE STUDY	232
6.2 – IMPLICATIONS FOR SECOND LANGUAGE RESEARCH METHODOLOGY	235
6.3 – IMPLICATIONS FOR SECOND LANGUAGE THEORY and RESEARCH	239
6.4 – IMPLICATIONS FOR SECOND LANGUAGE TEACHING	245
6.5 – IMPACT ON MY TEACHING PRACTICE	248
6.6 – CHAPTER SUMMARY	251
BIBLIOGRAPHY AND REFERENCES	253
APPENDICES	273

LIST OF ACRONYMS

B1	Linguistic level included in the Common European Framework of Reference (CEFR) and equivalent to an intermediate level of proficiency.
BNC	British National Corpus. Compilation of millions of words from different sources – written and oral –, mainly from the British variety of English.
CEFR	Common European Framework of Reference. An introductory guide to this framework can be retrieved from http://www.englishprofile.org/images/pdf/GuideToCEFR.pdf (Cambridge University Press, 2013)
COCA	Corpus Of Contemporary American English. Compilation of millions of words from different sources – written and oral –, mainly from the American variety of English.
EdD	Doctorate in Education.
EFL	English as a Foreign Language.
ESL	English as a Second Language.
IELTS	International English Language Testing System. It measures the language proficiency of people who want to study or work where English is used as a language of communication.
L1	First language or mother tongue of a language user.
L2	Second language of a language user.
LCT	Listening Comprehension Test. Instrument designed to gather data on the listening comprehension of the study participants. See section 3.2.4 and Appendices 8-11.
LVST	Listening Vocabulary Size Test. Instrument created by McLean et al. (2015) to estimate the aural vocabulary size of L1-Japanese learners of English as a second language. The first items in the test are shown in Appendix 1.
LVT	Listening Vocabulary Test. Instrument designed to gather data on the aural vocabulary size of the study participants. See section 3.2.3 and Appendices 3 and 5.
MALQ	Metacognitive Awareness Listening Questionnaire. Instrument created “to assess the extent to which language learners are aware of and can regulate the process of L2 listening comprehension” (Vandergrift, Goh, Mareschal & Tafaghodtari, 2006, 432).

MNSQ	Mean Square. In the Rasch model it is the chi-square statistic divided by its degrees of freedom.
PET	Preliminary English Test. Also known as Cambridge English: Preliminary, and B1 Preliminary. Standardized examination created and administered by the University of Cambridge Local Examinations Syndicate (UCLES). The test is meant for English language learners with a B1-level according to the Common European Framework of Reference.
SLA	Second Language Acquisition.
SLL	Second Language Learning.
SLT	Second Language Teaching.
TOEFL	Test Of English as a Foreign Language. Standardized examination created and administered by Educational Testing Service. It intends to measure the academic communication skills in English for non-native speakers.
TOEIC	Test Of English for International Communication. Standardized examination created and administered by Educational Testing Service. It is meant for non-native speakers who need to measure their proficiency in English at the global workplace.
UCLES	University of Cambridge Local Examinations Syndicate. Non-teaching department of the University of Cambridge which operates under the brand name Cambridge Assessment. Within that institution, Cambridge Assessment English is responsible for the creation and administration of standardized language exams like Cambridge English: Preliminary (PET), currently known as B1 Preliminary.
VLT	Vocabulary Levels Test. Instrument designed by Nation and validated by Schmitt, Schmitt and Clapham (2001) to give an estimate of vocabulary size for second language learners of general or academic English.
VST	Vocabulary Size Test. Instrument designed by Beglar and Nation (2007) to estimate the written vocabulary size of English language users.
WRS	Word Recognition Speech: ability to map information from the speech signal onto the lexical units that information represents.
WVT	Written Vocabulary Test. Instrument designed to gather data on the written vocabulary size of the study participants. See section 3.2.3 and Appendices 4 and 6.
ZSTD	Z-Standardized. In the Rasch model, it reports the statistical significance (probability) of the chi-square (mean-square) statistics occurring by chance when the data fit the Rasch model.

LIST OF FIGURES

2.1 – Cognitive processes and knowledge sources in listening comprehension (Vandergrift & Goh, 2012, 27)	26
3.1 – Research Instruments: Design, Implementation and Use	85
3.2 – Instruments and Analyses to answer Research Questions.	91
3.3 – Wright Map – Person abilities and item difficulties for the LVT (81 items; 73 persons)	128
3.4 – Wright Map – Person abilities and item difficulties for the WVT (81 items; 73 persons)	128
4.1 – Wright Map – Person abilities and item difficulties in the LVT – (First Data Collection – October 2019)	155
4.2 – Wright Map – Person abilities and item difficulties in the WVT – (First Data Collection – October 2019)	156
4.3 – Wright Map – Person abilities and item difficulties in the LCT – (First Data Collection – October 2019)	157
4.4– Wright Map – Person abilities and item difficulties in the LVT – (Second Data Collection – June 2020)	160
4.5 – Wright Map – Person abilities and item difficulties in the WVT – (Second Data Collection – June 2020)	161
4.6 – Wright Map – Person abilities and item difficulties in the LCT – (Second Data Collection – June 2020)	163

LIST OF TABLES

3.1 – Inclusion criteria for the target population.	84
3.2 – Research Questions	89
3.3 – Items in PET Vocabulary List according to frequency bands in 1-25k BNC-COCA	103
3.4 – Compound nouns in PET Vocabulary List according to frequency bands in 1-25k BNC-COCA	104
3.5 – PET Vocabulary List including compound nouns according to frequency bands in 1-25k BNC-COCA	104
3.6 – Summary of the PET listening paper (Cambridge University Press, 2008)	109
3.7 – Results in consecutive thirds of items in vocabulary tests (results based on raw data).	114
3.8 – Reliability and Separation depending on the number of items in the test (expressed in logits).	116
3.9 – % of correct answers in the test depending on number of items considered and corresponding values for separation and reliability (expressed in logits)	118
3.10 – Preliminary Study (May 2019) – Items in the listening vocabulary test with highest misfit values	122
3.11 – Preliminary Study (May 2019) – Items in the written vocabulary test with highest misfit values	122
3.12 – Preliminary Study (May 2019) – Persons in the listening vocabulary test with highest misfit values	303
3.13 – Preliminary Study (May 2019) – Persons in the written vocabulary test with highest misfit values	303
3.14 – Overall fit values for the listening and written vocabulary test expressed in logits (81 items)	123
3.15 – Count of the options chosen by participants for the items with highest misfit (N = 73).	124
3.16 – Descriptive statistics for the listening and written vocabulary test.	126
3.17 – Person and Item separation and reliability in two random halves of items in vocabulary tests	132
3.18 – Person and Item reliability and separation in logits – Preliminary Study vs First Data Gathering	134
3.19 – Comparison of MIN, MAX, MEAN and percentage of correct answers (raw data) – Preliminary Study (May 2019) vs First Data Gathering (October 2019)	135
3.20 – Main Study (October 2019) – Items in the listening vocabulary test with highest misfit values (shaded cells represent flagged items)	304
3.21 – Main Study (October 2019) – Items in the written vocabulary test with highest misfit values (shaded cells represent flagged items)	304

3.22 – Main Study (October 2019) – Items in the listening comprehension test with highest misfit values (shaded cells represent flagged items)	305
3.23 – Main Study (October 2019) – Persons in the listening vocabulary test with highest misfit values (shaded cells represent flagged persons)	306
3.24 – Main Study (October 2019) – Persons in the written vocabulary test with highest misfit values (shaded cells represent flagged persons)	307
3.25 – Main Study (October 2019) – Persons in the listening comprehension test with highest misfit values	308
3.26 – First Data Collection (October 2019) – Percentage of misfitting items or persons vs overall counts	136
3.27 – First Data Gathering (October 2019) – Summary of fit values for items and persons expressed in logits.	136
3.28 – Summary of mean measures, standard deviation and fit values for items and persons expressed in logits. Preliminary Study vs First Data Collection.	137
3.29 – Misfitting items in preliminary study (May'19) vs First Data Gathering (October'19)	138
3.30 – Misfitting items in First Data Gathering (October 2019) in LVT, WVT and LCT with their item measures expressed in logits.	139
3.31 – Person and Item reliability and separation (logits) across datasets (May'19, October'19, and June'20)	142
3.32 – Reliability and Separation indices (logits) with and without perfect scores (June 2020)	143
3.33 – Main Study (June 2020) – Items in the listening vocabulary test with highest misfit values (shaded cells represent flagged items)	309
3.34 – Main Study (June 2020) – Items in the written vocabulary test with highest misfit values	310
3.35 – Main Study (June 2020) – Items in the listening comprehension test with highest misfit values (shaded cells represent flagged items)	311
3.36 – Main Study (June 2020) – Persons with their fit values in the LVT (shaded cells represent flagged persons)	311
3.37 – Main Study (June 2020) – Persons with their fit values in the WVT	312
3.38 – Main Study (June 2020) – Persons with their fit values in the LCT	312
3.39 – Percentage of misfitting items in first and second data collection – October'19 vs June'20	144
3.40 – Second Data Gathering (June 2020) - Overall fit values for the items and persons (expressed in logits)	144
3.41 – Mean measure, standard deviation, and fit statistics across datasets (expressed in logits)	145
3.42 – Items in Second Data Gathering (June 2020) with biggest misfit in LVT, WVT and LCT with their item measures expressed in logits.	146

4.1 – Comparison of MIN, MAX, MEAN and percentage of correct answers – Preliminary Study (May 2019) vs First Data Gathering (October 2019) – Calculations based on raw data	152
4.2 – Mean measure and standard deviation in LVT, WVT and LCT. Preliminary Study vs First Data Collection. - Results in logits.	153
4.3 – Comparison of MIN, MAX, MEAN and percentage of correct answers – May'19, October'19 and June'20 - Calculations based on raw data	158
4.4 – Mean measure and standard deviation in LVT, WVT and LCT across datasets (May'19, October'19, and June '20) – Results expressed in logits.	159
4.5 – Multiple regression analysis of the LCT (N =283) – Calculations made on the person measures – (October'19)	166
4.6 – Comparison of scores and measures in LVT and WVT according to performance in LCT (N =284)	168
4.7 – Multiple regression analysis for top LCT scores in October 2019 (N = 48)	169
4.8 – Multiple regression analysis for bottom LCT scores in October 2019 (N = 235)	169
4.9 – Pearson product-moment correlations among LVT, WVT and LCT. October 2019 vs June 2020. - Calculations made on person measures	170
4.10 – Table 4.10 – Multiple regression analysis of LCT in June 2020 (N = 283) – Calculations made on the person measures.	171
4.11 – Scores in LVT and WVT for top scorers in LCT with corresponding lexical coverage (October 2019) – Calculations based on raw scores - (N = 48)	175
4.12 – WVT and LCT results vs bands of lexical coverage according to results in LVT – October 2019 - (N = 283) – Calculations made on raw scores	177
4.13 – LVT and LCT results vs bands of lexical coverage according to results in WVT – October 2019 - (N = 283) - Calculations made on raw scores	177
4.14 – Significance and Effect Size of differences between LVT vs WVT and LVT vs LCT results. Bands based on LVT results – October 2019 - (N = 283) – Calculations based on raw scores	178
4.15 – Significance and Effect Size of differences between WVT vs LVT and WVT vs LCT results. Bands based on WVT results – October 2019 - (N = 283) – Calculations based on raw scores	178
4.16 – Descriptive statistics and coverage in LVT and WVT according to LCT results bands (Dataset, June 2020)	179
4.17 – Descriptive statistics in LVT and WVT according to LCT results bands (October'19 vs June'20)	180

4.18 – Comparing results LVT vs WVT across three datasets (May'19 – October'19 – June'20) – Calculations based on raw scores - (*) N = 282 in WVT	183
4.19 – Comparing results LVT vs WVT across three datasets (May'19 – October'19 – June'20) – Calculations based on raw scores - (*) N = 283 in WVT	184
4.20 – Comparison of MIN, MAX, MEAN person measures (expressed in logits) – May'19, October'19 and June'20	184
4.21 – Comparison of p values and effect sizes in t -tests for person measures across datasets – May'19, October'19 and June'20	185
4.22 – Significance and effect sizes for differences in mean person measures across datasets according to LCT results bands (October'19 vs June'20)	186
4.23 – Pearson product-moment correlations among LVT, WVT, and LVT based on person mean measures across three datasets – May'19 - October'19 - June'20.	189
4.24 – Comparison of multiple regression analysis of the LCT based on person measures - October'19 vs June'20	190
4.25 – Descriptive statistics in LVT and WVT based on person measures, according to LCT scores – October 2019 vs June 2020 – (*) N = 236 in WVT	190
4.26 – Descriptive statistics in LVT, WVT and LCT based on person measures in both data collections (October 2019 and June 2020) – (N = 17)	191
4.27 – Statistics for paired t -tests on differences between raw scores in LVT, WVT and LCT from October 2019 to June 2020	191
4.28 – Statistics for paired t -tests on differences between raw scores in LVT-WVT, LVT-LCT and WVT-LCT (October 2019 and June 2020)	192
4.29 – Descriptive statistics in LVT and WVT based on person measures, according to LCT scores for participants in both data collections (October 2019 and June 2020) – (N = 17)	192
4.30 – Statistics for paired t -tests on differences between raw scores in LVT, WVT and LCT from October 2019 to June 2020, according to bands of LCT results	193

CHAPTER 1 – INTRODUCTION

This dissertation is the culmination of a research journey that began four years ago. The initial question that piqued my curiosity was how I could help my Spanish-speaking students become better listeners in English. Their complaints about not being able to understand most of the aural texts in that language led me to focus my investigation on exploring factors that might facilitate their listening performance (Rubin, 1994). Among those factors, learners' vocabulary knowledge stood out as one of the best predictors for the overall proficiency in second language (Milton, Wade & Hopkins, 2010), and as a highly influential variable on language skills like listening.

1.1 – RESEARCH CONTEXT

My personal teaching context has played an important role in determining the topic of this dissertation, as well as its scope and specific focus. I am a teacher at university in Spain, and I am familiar with situations where students struggle to become competent users of English as a foreign language. Furthermore, my own experience as a language learner has influenced my stance towards the research topic in this dissertation.

The vast majority of my students in my 15 years of teaching experience have had Spanish as their first language (L1). Many of them have complained about the listening activities in their English classes because the speakers in the recordings tended to “speak too fast [or] swallow their words” (Field, 2009, 27). In some cases, they have even expressed their frustration for not being able to see any progress in their listening ability after months of hard work.

The perceived difficulty of listening in English might also be reflected in the overall test results L1-Spanish learners show with respect to other students with different mother tongues. The European Survey on Language Competences (European Commission, 2012) assessed the proficiency of thousands of foreign language learners in 15 different European Union educational systems. The European authorities had set for their citizens the objective of attaining the listening level of an independent language user, i.e., B2 according to the Common European Framework of Reference (CEFR). Only 32% of the survey participants showed to have reached that level in listening, whereas the percentage dropped to less than 13% among the Spanish learners (Costa & Albergaria-Almeida, 2015). Spain ranked in the 11th position in the survey for reading and writing, whereas its results in the listening tests were the second worst. Furthermore, about a third of the Spanish-speaking participants showed

a listening competence below the A1-level (European Commission, 2012).

Finally, this dissertation focuses on the relationship between vocabulary and listening among L1-Spanish learners with a B1-level in English (CEFR). My experience with those language learners has shown that achieving this level of language proficiency might serve to predict their future success in English. Once that level has been achieved, learners might make further progress, and arrive at an advanced level of proficiency. Those who are still struggling to consolidate their B1-level might interrupt their learning and resume it after some time. Eventually, these learners might be making no actual progress, despite the many years they might have been studying the language (Yi, 2011). Interestingly, among the recommendations for those learners aiming to achieve the B1-level of language proficiency, research has highlighted the importance of developing both L2 vocabulary and listening (Richards, 2008a).

1.2 – VOCABULARY AND LISTENING – GENERAL INTRODUCTION

The literature review carried out for this dissertation confirms that my students' complaints are similar to the ones made by other students from other language backgrounds (section 2.1.2). Second language (L2) listening might cause anxiety in many language learners (Ferris, 1998; Xu, 2011), which has a negative impact on their performance (Graham & Santos, 2015; Mills, Pajares, & Herron, 2006). Furthermore, L2 learners tend to perceive listening as something difficult to learn, where they feel the least successful, particularly when they are tested (Kim, 2002; Graham, 2006).

This perception of listening as a difficult skill seems to extend to the classrooms, as some teachers might show attitudes that are not based on actual research evidence, and that might not help their students. Many teachers tend to think that listening is impossible or really difficult to teach (Field, 2009); adopting a 'comprehension approach' (Vandergrift & Goh, 2012), where the actual teaching is equated with testing the skill (Mendelsohn, 2006; Siegel, 2013). However, research in second language teaching (SLT) has shown that there are alternative perspectives for the teaching of this skill, where the focus is set not on the product to achieve – listening comprehension – but on the abilities, processes and knowledge that a listener needs for such achievement. This stance towards listening has shown to be more effective than just testing the listener's ability (Field, 2009; Hulstijn, 2003; Richards, 2008b; Tsui & Fullilove, 1998).

The investigation of what is necessary to achieve L2 aural comprehension could lead to the exploration of which factors might hinder or facilitate that achievement (Rubin, 1994). Once their impact on listening comprehension has been determined, more efficient teaching methodologies could be offered to

either minimize their negative influence, or to increase their beneficial effect. Furthermore, these methodologies might – in turn – contribute to reduce the anxiety caused by the listening experience among some L2 listeners (Vandergrift & Baker, 2015), and enhance their sense of self-efficacy (Graham & Santos, 2015).

Among the possible factors that might help our L2 students while listening in another language, the vocabulary knowledge of the target language has shown to be clearly beneficial (Fung & Macaro, 2019; Matthews, 2018; Wang & Treffers-Daller, 2017). Furthermore, this positive impact is particularly heightened among less proficient users (Pan, Tsai, Huang & Liu, 2018). Although some researchers might draw on the model proposed by Stanovich (1980), and claim that L2 listeners have compensation strategies and mechanisms to make up for their lack of vocabulary knowledge, language teachers and learners need to understand that nothing is able to compensate for the lack of the relevant vocabulary (Milton, 2009). Furthermore, cognitive load theory provides an additional argument for the inability of such mechanism to compensate for the lack of vocabulary knowledge in certain situations: if a text has too many unknown words, our mind is likely to be overwhelmed (Paas & Sweller, 2014). Alternatively, if the person's long-term memory has a sufficient number of lexical terms stored, they will be less likely to find unknown words in a text and therefore, to tax their working memory excessively.

Unfortunately, despite the negative perception L2 practitioners have about listening, and the importance of teaching how to develop the comprehension of aural texts, listening might be considered the “Cinderella skill” (Nunan, 2002, 238) in L2 research. Compared to other language skills, listening has received little attention in the literature, probably because it might seem more difficult to

investigate (Vandergrift, 2007). Consequently, the factors that impact positively or negatively on the listeners' performance have been neglected in the literature (Graham & Santos, 2015).

Moreover, most studies have investigated the relationship between L2 vocabulary and listening comprehension by matching the scores in written vocabulary tests to the results in listening comprehension tests (Read, 2013). In other words, they have tried to determine how related the language learners' vocabulary size is to their listening ability. However, those investigations might have disregarded the possible existence of two separate dimensions in L2 vocabulary knowledge – aural and written (Milton 2009) – by focusing only on the written form of words. This decision might be particularly relevant when the vocabulary scores are subsequently matched to the listening performance.

Moreover, most of the few studies employing aural vocabulary tests to assess the vocabulary size have drawn on research instruments that might not be the most suitable for that purpose. The use of dictation tests (Bonk, 2000), or of aural versions of word-recognition tests (Milton & Hopkins, 2006) might show construct validity issues, as well as an overestimation of learners' aural vocabulary size (van Zeeland, 2014a). Fortunately, there are other vocabulary tests that target the aural form of words, while raising no concerns about their validity or reliability (McLean, Krammer & Beglar, 2015).

The importance of accurately assessing the language learners' vocabulary size is two-fold. Firstly, because the possible relationship between their lexical knowledge and their listening performance can be determined more exactly, providing thus more reliable evidence to support further claims and recommendations. Secondly, because a relevant strand in the research into L2 learning has focused on determining the minimum percentage of words a

person should be able to recognize in order to function in another language (van Zeeland & Schmitt, 2013b). Once the required percentage is determined, it is matched to the frequency of occurrence of words in that language to estimate the approximate number of lexical items a language user should know. For example, if it is necessary to know at least 95% of the words in any text to understand it, and that percentage is covered with the occurrences of the 5,000 most frequent words in a language, learners should know those 5,000 words to function adequately in the target language.

Those percentages are based on the previous assessment of learners' vocabulary size. If the vocabulary tests are not sensitive enough they might overestimate or underestimate the actual vocabulary size, leading to inaccuracies in the number of words necessary to function in a language. As teaching and learning plans might be based on those figures, the impact on the classrooms is clear. An apparently minimal variation in the percentage of necessary words might imply learning thousands of new lexical forms (section 2.5.1).

1.3 – BRIDGING GAPS

In the previous sections we have seen the pertinence of carrying out research into L2 listening comprehension, especially among L1-Spanish speakers who want to become proficient listeners in English, and how their vocabulary size might contribute positively to their listening performance. The following paragraphs present the aims of this dissertation to contribute to the body of research into L2 vocabulary and listening comprehension.

First of all, as this dissertation investigates the facilitating effect of vocabulary, it focuses on the processes and elements leading to listening comprehension, rather than on the final product to be achieved. This investigation intends thus to add to the few studies about the relationship between L2 vocabulary size and listening comprehension. Furthermore, it presents an additional perspective on this topic, as it is a partial replication of previous research studies (McLean et al., 2015; Stæhr, 2009), although on a more linguistically homogenous population of L1-Spanish speakers. In particular, and based on the gaps detected in previous studies (section 2.6), this investigation intends to:

- 1) estimate the relationship between L2 vocabulary size and listening comprehension over time,
- 2) determine the differences between L2 aural and written vocabulary,
- 3) estimate the minimal L2 vocabulary size necessary to achieve listening comprehension.

Apart from adding to the existing research into L2 vocabulary and listening comprehension, this study contributes to that body of knowledge with its novel and unique design. This investigation intends to increase the accuracy of both estimations by focusing on the validity and reliability of the instruments

employed for the data collection, and in their subsequent analysis. In particular, this design aims at:

- 1) increasing the overall validity of the study by drawing on the same framework to create its research instruments,
- 2) enhancing the validity of those instruments by using two versions – aural and written – of the same vocabulary test, in a bilingual format,
- 3) refining the reliability of the research instruments by carrying out a preliminary study to determine the best performing items,
- 4) increasing the accuracy of the assessment by employing multiple measures (Webb, 2002), administered one after the other to the same participants at two moments in time,
- 5) improving the overall reliability of the study by drawing on the Rasch model, a more thorough and conservative approach to data analysis.

One last set of objectives are related to the fact that the present research study is part of a dissertation for a professional doctorate in Education (EdD). I want to help my own students, as well as other language learners, and find evidence in support of more beneficial methodologies and approaches than the ones currently in use (Vandergrift, 2007). Those novel perspectives might eventually contribute to change their perceptions about listening. Moreover, statistics indicate that L1-Spanish speakers might need more help than other students of English in Europe, particularly with listening comprehension (section 1.2). Therefore, both the preliminary and the main study are carried out in language centres in Spain, within a population of L2-English learners attending classes at a B1-level.

The following chapters will elaborate on the contribution of this dissertation to the field of L2 learning. Chapter 2 will present what researchers have already

said about L2 listening comprehension and its relationship with lexical knowledge. In particular, it intends to show the possible gaps in previous research that the present investigation might contribute to mitigate (section 2.6). Chapter 3 will deal with the methodology and methods employed in this dissertation. A thorough account of all the methodological decisions taken with respect to the study design will be provided, as well as a detailed description of the unique contribution of the Rasch model, and a discussion of the quality of the instruments employed in the data collection. Chapter 4 will show the different data analyses performed to find evidence in support of the answers to the research questions. Chapter 5 will present those answers and contextualise them by drawing on some studies already discussed in the literature review. Finally, Chapter 6 will discuss the impact of those answers on future research into L2 vocabulary and listening, on theoretical and methodological approaches to this research topic, and on the specific aspects of L2 classroom practice affected by the findings and claims of the present study.

1.4 – CHAPTER SUMMARY

This chapter has introduced the broad topic in this dissertation: second language vocabulary and listening comprehension. After a brief presentation of the negative results and perceptions L2 listening might generate, we have seen the need for more research in the field, particularly investigations based on more adequate methodological approaches. In the final section of this introductory chapter, we have presented how this study might contribute to bridge some of the gaps mentioned in the previous sections by mentioning its main aims.

CHAPTER 2 – LITERATURE REVIEW

Helping students be better listeners in a second language has a direct positive impact on their overall linguistic performance, because in some cases, most of the language they acquire is through the linguistic information they hear (Richards, 2008b; Rost, 2006). However, reality in the language classrooms tells us that most students are simply tested in their listening skills instead of being taught how to be more proficient in that respect (Field, 2009; Vandergrift & Goh, 2012). Some research studies (e.g., Ferris, 1998; Graham, 2002; Graham & Santos, 2015; Kim, 2002; Mills et al, 2006; Xu, 2011) have investigated the problems L2 learners face when dealing with this language skill. Other studies have addressed the possible factors that might bear an influence on the listening performance (e.g., Boyle, 1984; Mendelson, 2001; Rubin, 1994; Tsui & Fullilove, 1998; Vandergrift & Baker, 2015). Among those studies that have focused on the listening skill, the vocabulary size of those L2 learners has been pointed out as one of the possible predictors of their listening performance (Field, 1998, 2009; Goh, 2005, Rost; 2005, 2011; van Zeeland, 2018; Vandergrift & Baker, 2015; Wang & Treffers-Daller, 2017).

This study intends to explore the contribution of language learners' vocabulary size on their ability to understand aural texts. Consequently, the first sections in this literature review will present the reasons why listening is the language skill under study, as well as a brief description of the listening model that underpins the investigation (section 2.2). Based on that model, section 2.3 will address the possible influence of L2 vocabulary knowledge on listening performance. The final sections of this chapter will focus on more practical issues as they deal with questions such as *knowing* a word or *quantifying* the vocabulary size, and with concepts like corpora and frequency vocabulary lists.

2.1 – INTRODUCTION

Experts in second language acquisition (SLA) claim that ‘listening’ and ‘listening comprehension’ are synonymous (Richards, 2008b). Although many authors have attempted to define listening comprehension (Buck, 2001; Rost, 2011), they all tend to see listening comprehension as a process of making sense of the linguistic input delivered orally to a person (Yi’an, 1998). Two important aspects stand out from this definition. First, that the purpose of listening is comprehension (making sense). Second, that the listener plays an active role – “fully as active as when speaking” (Mendelsohn, 2001, 34) – because they combine the use of multiple resources to achieve comprehension.

2.1.1 The importance of listening

Listening is probably the most important skill to obtain comprehensible input (LeLoup & Ponterio, 2007), and we can consider it the “foundation of language acquisition and communication ability” (Rost & Wilson, 2013, xiii). In fact, adults spend about half of the time they need to communicate just listening to what other people are saying (Siegel, 2015). In the case of L2 learners, listening might be the primary source to acquire the target language (Rost, 2006). This might be the case with classrooms and methodologies where oral interaction is a priority, whereas skills like reading or writing might be the main focus of interest in other contexts of language learning, like a doctoral thesis, for example.

Research has shown that listening is pivotal in many forms of communication, and that understanding oral input is essential in several daily situations. Therefore, any enhancement in the way our L2 students use their listening skills

will have a noticeable impact on their overall linguistic performance, especially in informal education and in settings where communication is encouraged (Vandergrift & Baker, 2015).

The importance of carrying out investigations on L2 listening can also be seen in the tests results European citizens have shown in that particular skill. The European Survey on Language Competences assessed the proficiency of 54,000 Europeans in learning a foreign language across 15 educational systems in 14 different countries. Costa and Albergaria-Almeida (2015) concluded that only 32% of the participants showed a B2-level in listening as stated in the Common European Framework of Reference (CEFR). According to the educational authorities in most European countries, students should have that linguistic level for the first foreign language they are learning by the end of their secondary education (Bařdak, Balcon & Motiejunaite, 2017). The results in this survey are comparatively worse for the target population of the present study, i.e., L1-Spanish speakers who are learning English as a foreign language. Among the Spanish participants less than 15% of them showed a proficiency of a B2-user in listening, and almost a third of them showed a linguistic competence below the A1-level in that skill (European Commission, 2012).

2.1.2 Listening in language learning

The previous section has shown how being able to understand aural texts is important in our everyday lives, particularly among language learners. However, they perceive listening as something difficult to learn, where they feel the least successful, particularly when they are tested (Kim, 2002; Graham, 2006). Several reasons might account for this perception of difficulty, and the

subsequent feelings of low self-efficacy. Students might feel less 'prepared' for the listening part of a language test, when compared to other sections of the same examination. This perception from students might come from the belief held by many teachers that effective listening is synonymous with task completion (Graham, Santos & Francis-Brophy, 2014). If the students' results in a test are poor, the logical consequence is to think that success has not been achieved.

Secondly, L2 learners might perceive listening as difficult because there exists little ecological validity (section 3.1.2.2) in the way listening skills are taught and tested, which in turn might lead them to feel 'less in control'. In real-life situations, listeners have different aids at their disposal that might help them overcome possible communication problems (Alderson, 2005). They can interrupt the speaker's discourse, ask for clarification of some parts of the content, or draw on paralinguistic features like body language. Unlike what happens in most real-life situations (Lynch, 1997), in the vast majority of listening tests, the input is unidirectional – preventing any interaction with the speakers –, conveyed by people who are perfect strangers to them, and about topics or situations that are usually foreign to the listener, and that might attract little interest on their part. In this respect, Field's unexpected finding about L2 listeners' beliefs is revealing (Field, 2012). Contrary to his initial hypothesis, he concluded that the ecological validity of the listening task is a more decisive factor for the students to perceive its relative difficulty than the cognitive demands it might pose on them.

Another aspect of those negative perceptions on L2 listening is the anxiety it causes among learners. We can understand anxiety here as a "state of anticipatory apprehension over possible deleterious happenings" (Bandura,

1997, 137). In this definition, perceptions of self-efficacy are key to successfully manage that apprehension over future negative events (Mills et al., 2006). In this respect, research has documented how language learners lack confidence in their oral abilities, as well as the stress and anxiety they feel when listening is at stake (Ferris, 1998; Xu, 2011). Anxiety influences the listener's performance (Graham & Santos, 2015; Mills et al., 2006) because if "learners are worrying about not understanding, they are not giving their full attention to the task at hand" (Arnold, 2000, 784). Alternatively, the more confident and less anxious the listener feels, the better their performance in listening comprehension (Brunfaut & Revesz, 2014; Vandergrift & Baker, 2015). Furthermore, language instructors may contribute to their L2 students' comprehension by enhancing their confidence in their own abilities in a foreign language (Mills et al., 2006).

Previous research studies have pointed out several reasons for those perceptions of listening as difficult skill to learn, and a major source of anxiety among L2 learners. Firstly, listening is transitory. The different nature of written and spoken texts is a crucial factor here. When a person reads a text, they have permanent access to the input, and they can get back to its relevant parts if necessary, because the text is present in time and space (Ridgway, 2000). Investigations on how the L2 reader's eyes behave while reading confirm that they do not process the written input in a linear and straightforward fashion, and that they draw on the permanent nature of the written text. Compared to native readers, language learners take longer to read a text, fixate more on some parts, divert less their attention to previous parts of the text, and skip fewer words (Cop, Drieghe & Duyck, 2015). On the other hand, aural texts are only present in time, which implies processing online the information conveyed, without the possibility of getting back to a previous passage unless the delivery

is interrupted, and a repetition or clarification provided. Furthermore, this resource is usually available to the listener only in real-life situations, whereas in L2 classrooms – and especially when learners' listening performance is assessed – this possibility is seldom made available to them.

A second aspect of L2 listening that might cause those perceptions of difficulty and feelings of anxiety might be that the input L2 learners receive is a *block without spaces*, as the words are pronounced in connected speech. Again, the intrinsically different nature of written and aural texts plays a major role in the way language users approach them, and on how their working memory is taxed (Ridgway, 2000). When a person reads a text in the target language, they can benefit from the presence of blank spaces to delimit the individual words (Field, 2009), determining where one word ends and the next begins. Furthermore, when new ideas are introduced in a written discourse, they are easily perceived because a new paragraph or section begins. However, in oral texts – particularly in more spontaneous and unscripted texts – those transitions and marks might be harder to find.

A third factor that renders listening a comparatively more difficult skill is that listeners might need to pay attention to additional and complementary ways to deliver the message (Zhang & Graham, 2020). In listening comprehension, the context might be more important than in reading, as the speaker has other 'channels' to convey their message. Listeners might need to pay attention not only to the lexical and syntactic output the speaker is conveying – as they would with a written message – but also to the particular stress or intonation employed by the speaker. Moreover, although in some L2 classrooms the listening events are usually one-way (Lynch, 1997), and limited to double playing short audio passages (Field, 2009), in real-life situations listeners might need to make use

of visual elements accompanying the aural text such as the speaker's body language, or the use of multimedia elements like diagrams or video.

These three factors – the transitory nature of aural input, the blurred contours of words in connected speech, and the importance of context and multimodality in some listening events – might tax language learners' working memory in listening more than in reading (Ridgeway, 2000; Vandergrift & Baker, 2015). If their working memory – responsible for the temporary storage and manipulation of information for language processing – is overwhelmed, a breakdown in comprehension might occur, which might cause subsequent feelings of anxiety and low perceptions of self-efficacy (section 2.2.3.2).

2.1.3 Listening in language teaching and research

The relative difficulty of listening might also become apparent among L2 teachers. Many language classrooms have reduced the instruction of listening to the mere checking of correct answers on the part of the learners (Vandergrift & Goh, 2012). Some teachers might believe that listening is difficult to teach, and a skill where tangible outcomes are hard to achieve (Field, 2009). There might exist cases where “many teachers are themselves unsure of how to teach listening in a principled manner” (Vandergrift & Goh, 2012, 4). Consequently, language teachers only play recordings and check answers, without asking why or how their students have arrived at those answers (Mendelsohn, 2006), and leave thus no room for the actual teaching of the skill.

This lack of systematicity among language teachers with respect to teaching listening might be the consequence of the comparative scarcity of research into the skill (van Zeeland, 2014b; Wang & Treffers-Daller, 2017). If teachers do not

have access to soundly designed research into L2 listening, with readily applicable findings to be implemented in the language classrooms, they might continue to believe that checking their students' answers is the only possible manner to teach how to be better L2 listeners. However, another reason might be the lack of consistency between what teachers think is necessary with respect to teaching this skill, and what they really do in their classrooms. Graham et al. (2014) found that L2 teachers saw listening as a teachable skill with clear ideas from research about how to do it. Yet, few of them introduced activities meant to teach their students how to listen. For example, they thought that identifying word boundaries in connected speech was important to listening comprehension. However, only a minority used this finding from previous research studies (Goh, 2000; Rost, 2005) and asked their students to identify those word boundaries while listening.

Moreover, listening might be perceived as difficult by publishers and textbooks authors, as it has also received the least systematic attention from them (Vandergrift & Goh, 2012). Although some research findings have been available for decades, there seems to exist a gulf between what research says and what publishers offer L2 language learners and teachers (Graham & Santos, 2015).

With respect to how research has approached listening, researchers might have considered it as something difficult to investigate. Its transitory nature might account for the limited number of L2 research studies that have focused on listening (Alderson, 2005). It might be the least researched of the four skills because the ephemeral nature of the aural input makes it clearly more difficult to be analysed and studied (Vandergrift, 2007). Furthermore, language research might tend to employ more written than spoken material simply

because “it is easier to do experimentally” (Anderson, 2020, 417).

Moreover, when researchers study the processes that lead to reading comprehension – the other traditionally considered receptive language skill – they can make use of devices to record the reader’s eye movements, and inferentially link the results to different mental processes (Lynch, 1998; Conklin & Pellicer-Sánchez, 2016). Comparatively, the difficulty of accessing the listeners’ minds, and the complexity listening as a construct might have led many researchers to draw on reading-based findings in their investigations about this skill (van Zeeland, 2014b; sections 2.5.2.1 and 5.2).

The first part of this section has highlighted two key aspects of how listening is understood in the present study. Firstly, that the purpose of listening is to make sense (i.e., comprehension). Secondly, that listeners play an active role in making that sense while drawing on a variety of resources. Then L2 learners’ perceptions on listening have been analysed, and their importance within their learning experience contextualised. The last part of this section has focused on the idea that listening is not only difficult for learners to learn, but also for researchers to study, and for teachers and publishers to deal with. Although different studies have been discussed, more research on L2 listening is necessary to prevent it from becoming “a source of frustration to learners and an area in which it seems difficult to make progress” (Graham, 2011, 139).

2.2 – UNDERSTANDING LISTENING

Section 2.1 has highlighted the importance of listening in second language learning and addressed how it is perceived among language learners and teachers, as well as the possible reasons for those perceptions and their consequences. Now the discussion focuses on the listening model that underpins the present study.

The ‘comprehension approach’ (Field, 2009) – equating listening to a product to be gained – might be considered ‘commonplace’ in second language classrooms (Vandergrift & Goh, 2012). However, focusing on the listening ability, on its processes and knowledge sources to achieve comprehension of aural texts is certainly a more efficient way to help our students develop as proficient listeners in their L2s (Hulstijn, 2003; Tsui & Fullilove, 1998). In other words, to help language learners become better listeners it is necessary to understand what listening comprehension entails. In fact, this study focuses on investigating the possible influence of one element (vocabulary) on the listening performance of L2 learners.

2.2.1 Listening as a process: The ‘teaching approach’

Several experts in second language teaching have argued that a successful listening pedagogy has to derive from studying listening as a process, not as a product (Richards, 2008b; Field, 2009). Consequently, teachers need to understand how learners engage in it, what difficulties they have, and how they deal with those difficulties (Graham & Santos 2015). This pedagogy focuses on the ability itself, and approaches the phenomenon of listening from a more holistic perspective that might encompass all the processes and demands –

internal and external – that affect comprehension (Vandergrift & Goh, 2012).

This shift of focus from the product to the process, lends itself to the analysis of the possible difficulties inherent in understanding aural messages, so that part of the variability in learners' listening performance might be accounted for. By ascertaining the reasons why some listeners are more successful than others, language teachers might then be able to devise methodologies to help their students more efficiently. In this new model, the comprehension approach is considered a means to an end: "instead of simply checking answers, the instructor operates diagnostically, establishing precisely why certain answers (correct or incorrect) have been given" (Field 2009, 95). By analysing the answers given and the reasons why the students have chosen them, the teacher might begin to have a clearer picture of what is actually happening in their students' minds when they listen to a text in a foreign language. Once teachers have accessed this information, they could devise methodologies to help their students overcome the difficulties they are experiencing (Field, 1998). In this respect, several authors have analysed listening texts in an attempt to compile the possible factors that might affect their relative difficulty (e.g., Buck, 2001; Rubin, 1994), so that teachers are able to anticipate those 'obstacles' and design activities that might help their students surmount them.

A second strand towards a process approach to teaching listening is advocated by authors who have focused on providing L2 learners with strategy instruction to overcome possible difficulties and become better listeners. Instead of diagnosing what causes listening breakdowns, they based their offer on an initial diagnostic test, and a subsequent needs analysis. One illustrative example in this strand is the use of the Metacognitive Awareness Listening Questionnaire (MALQ) "to assess the extent to which language learners are

aware of and can regulate the process of L2 listening comprehension” (Vandergrift et al., 2006, 432). Another example might be the awareness-raising activities proposed by Graham and Santos (2015) along with some strategies to help learners become better listeners. In both cases, language learners are informed about the existence of the metacognitive resources at their disposal – like planning and evaluation, or directed attention – and then instructed in how to make the best use of them.

Nevertheless, despite all research in support of a teaching approach to listening – and unlike the case of reading or writing – few teachers and even fewer published methods have adopted stance towards the skill (Field, 2010; Siegel, 2015; Tomlinson, 2013). The absence of a real pedagogy of L2 listening in the language classrooms might be attributable to a lack of understanding of what listening really is and what the listening processes entail (Mendelsohn, 2001). Furthermore, language teachers might be unaware of a range of activities that might help effectively in the development of subskills and strategies necessary to listen competently (Siegel, 2013).

Another reason for the persistence of product-based approaches to listening in the L2 classrooms might be the washback effect generated by some language testing institutions and high-stakes public examinations, with tests that have not evolved over the years (Goh, 2008), and that identify the skill of listening with choosing or writing the *right* answer to a question. Leading institutions in the market of language testing like Cambridge Assessment English or Education Testing Service conceive the listening paper in their language examinations as a series of unidirectional listening tasks. Every year, millions of language students take a standardized language exam to certify their level of proficiency because they might need a visa, apply for a job, or begin their university studies

abroad. They need to be prepared for that exam, so because of the washback effect, teachers offer them listening exercises and tasks replicating the ones featured in the test. This methodology might lead students to a *de-facto* identification between unidirectional listening and listening in real life, as well as identifying listening success with a high score in the listening section of a standardized test. This kind of test might be extremely reliable from a psychometric point of view, but they certainly lack certain ecological validity when compared to real-life listening situations (section 3.1.2.2).

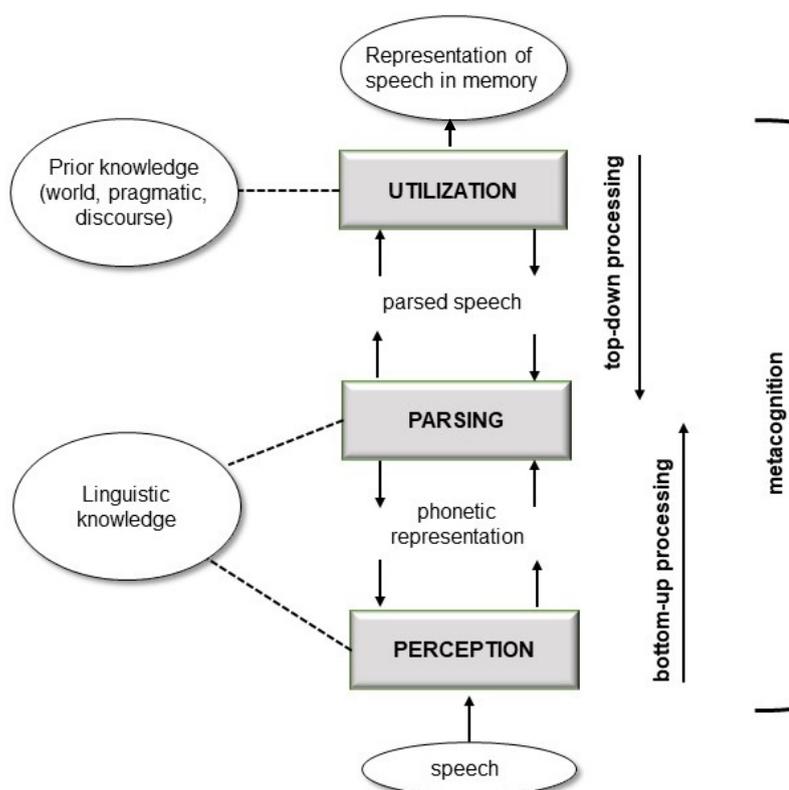
This section has presented an alternative view to the comprehension approach. This pedagogy considers listening as an ability underpinned by a series of processes that lead to the achievement of the final goal of comprehension. Furthermore, it suggests analysing the different processes involved in listening, and the intrinsic difficulties they might pose to language students. Then, either remedial activities, or instruction on strategies are offered to students to help them overcome the obstacles to comprehension and become better listeners. The following section introduces different perspectives found in the literature about those listening processes, and presents the listening model that underpins the present study.

2.2.2 Listening processes

Listening is certainly a complex skill that involves a series of psycholinguistic abilities, processes, subskills, and knowledge sources (Field, 2009; Rost, 2011). Vandergrift & Goh (2012) presented a thorough account of what L2 listening comprehension entails, and identified four sets of cognitive processes: 1) controlled and automatic processing, 2) perception, parsing and utilisation, 3)

metacognition, and 4) top-down and bottom-up processing. They also highlighted the importance of both linguistic knowledge (e.g., phonological or vocabulary knowledge) and prior knowledge (e.g., background and pragmatic knowledge) to be a successful listener. Figure 2.1 shows the interrelationships between the sets of processes and the different knowledge sources.

Figure 2.1 – Cognitive processes and knowledge sources in listening comprehension (Vandergrift & Goh, 2012, 27)



Automatic versus controlled processing refers to how rapidly and accurately language learners are able to access the knowledge sources necessary to process aural texts. The ephemeral nature of the auditory signal is one of the reasons why listening is perceived as a difficult skill (section 2.1.2) because it forces the listener to process that input almost online. Research has emphasized the importance of having a high degree of automaticity in processing the acoustic input so that attentional resources are free to focus on higher-level information (Field, 2009; Hulstijn, 2003). Generally speaking, good

L2 listeners are those who have *automatized* some of the listening processes, and are able to focus their attention on aspects of wider meaning (Field, 2009).

The framework of perception, parsing and utilization is based on Anderson's (2020) model of listening comprehension, one of the most widely cited in L2 research (Zhang, 2018). In the first phase – perception – listeners use bottom-up processing to recognise sounds and get a phonetic representation. Then, this representation is parsed to activate potential word candidates by using both word-based cues like the onset or salience, and meaning cues like the context or the topic (van Zeeland, 2014a). In the final stage of utilization, information from the perception and parsing stages is related to information stored in long-term memory. This representation is not sequential, but the three phases have a two-way relationship with each other.

Metacognition refers to the language learners' awareness of the cognitive processes that take place while listening, as well as their ability to monitor, regulate and make an orchestrated use of them. Again, successful listeners use metacognition more to regulate the listening processes and achieve comprehension (Graham, Santos & Vanderplank, 2008; Vandergrift & Goh, 2012).

In the literature the distinction between bottom-up and top-down is probably the most widely used approach to L2 listening (van Zeeland, 2014a). Bottom-up processing is identified with linguistic processing. The focus is on sounds, phonemes and parts of the words that we hear (Graham & Santos, 2015), so that we are led by the input we receive in real time (Rost, 2011). On the other hand, research considers top-down processing as equivalent to semantic and pragmatic processing. In this case, higher-level mental processes help us build ongoing and tentative representations of what the message might be like.

These mental processes make use of our previous experiences, and of what we expect from that particular listening situation (Rost, 2011).

Research has claimed that bottom-up and top-down processes do not refer to particular levels of processing aural input, but to the direction towards which these processes are heading. In a bottom-up process, small or lower-level units are progressively reshaped into larger ones; whereas in a top-down process, larger units exercise an influence over the way in which smaller ones are perceived (Field, 2009; Rost, 2006). Furthermore, these processes are not considered to be alternatives, but “mutually dependent and highly interconnected” (Field, 2008b, 3). In other words, listeners employ both directions of processing when trying to understand aural input. They might try to recognise and decode individual words in bottom-up processes to form larger structures of discourse, while using contextual cues and world knowledge for top-down processing to check that those larger structures have been correctly formed.

This section has shown a listening model where the skill consists of a series of complex processes involving not only linguistic knowledge, but also other sources of information that the listener may have like their familiarity with the listening situation and the auditory input. Furthermore, the processes to understand aural messages may begin in the auditory signal and finish in the listeners’ minds or the other way round, in an overlapping and iterative sequence that might take milliseconds. The following section intends to highlight the importance of bottom-up processing in this listening model, and connect it to the methodological choices of the present investigation.

2.2.3 Listening in the present study

Both language teachers and learners need to be offered more effective approaches to listening than the currently used in most classrooms. A more effective listening pedagogy is possible when it is based on the analysis of the different processes involved in listening comprehension (section 2.2.1). This section draws on an analysis of what L2 comprehension consists of (section 2.2.2), and discusses the impact of either bottom-up or top-down processing on the listening ability.

2.2.3.1 Bottom-up and Top-down.

Many researchers have advocated for the introduction of both bottom-up and top-down activities to help L2 learners become better listeners (Andringa, Olsthoorn, van Beuningen, Schoonen & Hulstijn, 2012; Field, 1998; Graham & Santos, 2015; Mendelsohn, 2006; Rost, 2011; Vandergrift & Tafaghodtari, 2010). Bottom-up practice deals with recognizing and identifying sounds, phonemes, and words, focusing on the available linguistic knowledge. Top-down activities focus on the use the listener makes of other sources of knowledge, and of strategies to arrive at comprehending the aural input. It could have a compensatory function, as the listener can use that information strategically and compensate for inadequate linguistic knowledge like not being able to notice or recognise words in connected speech (Field, 2004; Vandergrift, 2007).

Although both types of processing are connected to each other and mutually dependent, the question is which one is more effective and should be favoured in the language classrooms. The answer is that both types of processing should

be employed, but in due course (Field, 2009). On the one hand, some research studies have catalogued good listeners as those who make better use of their inferencing skills to help them understand the message, and check understanding and monitor the whole process of arriving at meaning (Goh, 2002; Vandergrift, 2003). These studies claim that the listener may transfer knowledge from their first language (L1) to understand the speaker's mood through their intonation patterns, or from past experiences to anticipate the next step in highly standardized procedures like checking-in at a hotel, or going through security at the airport. These information sources – prior knowledge in the model presented in Figure 2.1 – might be used to confirm hypotheses, discard competing options from tentatively decoded units of meaning, or to monitor overall understanding.

Moreover, Field (1998) claims that there seems to exist an interactive-compensatory mechanism for some kind of automatic trade-off between the amount of information available to the listener from the aural input and the clues they have from the context. He relates his claim to the interactive-compensatory model that states that “a deficit in any particular process will result in a greater reliance on other knowledge sources regardless of their level in the processing hierarchy” (Stanovich, 1980, 32). Similarly, other authors suggest that in situations where comprehension of aural texts is limited, L2 listeners may use compensatory strategies and additional sources of information available to them (Vandergrift, 2004; Yi'an, 1998). This use of compensatory mechanisms might be even more necessary among lower-level language learners because those listeners “are limited by working memory constraints” (Vandergrift, 2004, 6).

However, these resources are not available to all listeners at all times. Figure 2.1 has clearly shown how the overall process to achieve comprehension

begins with the aural input, which is perceived and forms a phonetic representation. L2 listeners might already experience difficulties the moment they are exposed to the acoustic input, at the beginning of the bottom-up processing, as they struggle to decode the information they are receiving (Goh, 2000; Graham, Santos & Vanderplank, 2010). They might even fail to notice some of the words within the aural input they are trying to process, which clearly diminishes their opportunities to make lexical inferences (van Zeeland, 2014b). The reasoning follows Schmidt's 'noticing hypothesis' (Schmidt, 1990): if the listener is unaware of the presence of a word embedded in a piece of connected speech, they will probably fail to activate any top-down processes to infer the meaning from the context, or from their own prior knowledge.

The process of building meaning from aural input might be seen as an ongoing task primarily based on understanding – and noticing – what the speaker has said before. This assumption might imply that top-down strategies are not a good alternative to poor decoding skills, because “co-text depends entirely for its reliability upon whether the listener’s decoding skills are adequate or not!” (Field, 2009, 136). In other words, although contextual and co-textual evidence from the utterance might help the listener understand better the aural message, they might fail to do so if the listener is unable to decode sufficient building blocks or basic units of meaning in the message in an accurate manner. For example, a listener that decodes the input ‘I can’t’ as ‘I can’ might certainly be misguided in their assumptions about the overall message to be conveyed, in the subsequent inferences activated to help further understand that message, and in the process of checking-up whether larger structures of meaning have been formed accurately. This failure to decode a simple unit of meaning correctly – or even notice it – might affect the decoding of subsequent units and

the building of larger units of meaning.

Noticing and efficiently processing the aural input with the help of linguistic knowledge are 'at the basis of listening' (van Zeeland, 2014a). The next section will address how the use of the different knowledge sources in listening comprehension create two different patterns among weak and strong listeners.

2.2.3.2 Bottom-up and Top-down in weak and strong listeners

Some lower-level L2 listeners hardly ever make an "orchestrated use of bottom-up and top-down sources of information" (Graham & Santos, 2015, 13). It is an involuntary decision forced by the circumstances (Fung & Macaro, 2019). They are so overwhelmed by the input and the necessary efforts to decode it, that they are unable to allocate attentional resources to both bottom-up and top-down processes (Field, 1999, 2008b, 2009; Hulstijn, 2003; Rost, 2006; Vandergrift, 2003).

It seems reasonable to assume that one cannot achieve the overall comprehension of a message if they have difficulties in distinguishing the minor components of that message (van Zeeland, 2014b). In fact, research argues that many students are simply so overwhelmed with the online processing of new information that they cannot retain and interpret it (Conrad, 1985), so that no form of association or fixation occurs in long-term memory for subsequent processes of listening comprehension (Goh, 2000). This inability to cope with the online processing of oral input might occur even if the listener recognises all the words as they are spoken (Goh, 2002; Ohata, 2006).

According to the cognitive load theory, human beings might have two types of memory: working and long-term. Working memory is "the mental workspace used for the short-term storage and manipulation of information required for

diverse cognitive tasks” (Wiley, Sanchez & Jaeger, 2014, 599), while long-term memory is where our prior knowledge is stored. Our working memory is severely limited in terms of capacity and duration when it has to deal with novel information, whereas there is no limitation in the information held in our long-term memory (Paas & Sweller, 2014). Within this framework, learning occurs when new information passes from the working memory to the long-term memory to be stored there. In this process, the role of long-term memory is vital to facilitate the work of the working memory: prior knowledge might be used to reduce the uncertainty of dealing with too many novel elements at the same time (Paas & Sweller, 2014).

This description of the roles and capacities of both types of memory – working and long-term – might help explain how weaker L2 listeners fall into a *vicious circle* when they have to understand a spoken text: they fail to understand some lower units of meaning (words) because they do not know them, or because they are unable to notice them in connected speech. They could activate top-down processing to bridge the gap, but they are so busy at lower level units that they cannot pay attention to the correct use of those compensatory sources. Alternatively, if they pay more attention to the use of top-down strategies, they might miss part of the acoustic input, which, in turn might expand the gap in comprehension.

The opposite might be true for stronger listeners – those who are considered more successful or proficient in L2 listening – as they could benefit from a *virtuous circle*. They might experience little trouble at the level of decoding the input and parsing it, so they could pay full attention to the use of top-down resources to check understanding, monitor hypotheses, and build larger units of meaning. These resources might, in turn, help in the further decoding of lower

units as the discourse unfolds, because they might confirm the few tentative competing guesses those expert listeners may have made. As some commentators have explained (Hulstijn, 2003; Field, 2009), these language users are able to free up attentional resources because they have automatized the lower-level decoding processes (section 2.2.2).

Although top-down processing and information sources are relevant to achieve listening comprehension, it is necessary to find out what aspects of bottom-up processing could be developed through instruction, so that language learners are not only taught how to compensate for problems at that level (Rost, 2011). In the discussion of bottom-up and top-down processing, the importance of noticing (Schmidt, 1990) has been highlighted. One of the factors that may affect noticing is the salience of the word in the auditory input (van Zeeland, 2014a). Based on how our long-term memory interacts with our working memory (Pass & Sweller, 2014), we could conclude that the familiarity the listener has with the words in the auditory input – prior knowledge stored in their long-term memory – might play a clearly positive role in both noticing and decoding those words. Furthermore, the listeners' prior knowledge of the words featured in the input might also play a positive part in processing the aural text and arriving at its comprehension. In other words, it seems reasonable to assume that if the listener already knows the word, it might be easier for them to notice it in the input, and decode and process it in an almost automatic fashion (section 2.2.2), so that their working memory is not taxed and attentional resources are available to process the unknown parts of the auditory input.

The following section addresses the influence of the language learners' vocabulary size on their ability to understand aural texts. This exploration aims to present research evidence in support of more effective strategies and

methodologies for the language classroom (section 2.1). Those proposals might facilitate the noticing and bottom-up processing of the auditory input, and the subsequent use of top-down resources and processes (Mendelsohn, 2001; Yi'an, 1998). By doing this, many language learners might be helped to break that vicious circle they experience when listening.

2.3 – VOCABULARY AND LISTENING

The previous sections have shown why listening is the focus of the present study. First, we have seen how more efficient listening pedagogies are necessary, and how they should be based on the analysis of what listening comprehension really entails. Therefore, the overall listening comprehension model that underlies the present study has been introduced, with a special focus on bottom-up and top-down processing, and how its use might facilitate listening success.

When L2 listeners experience difficulties at the bottom-up level, a ‘compensation’ strategy might be activated and top-down processes are used to bridge the gap (section 2.2.3). Alternatively, when the linguistic input presents no difficulties to be understood, a ‘facilitating’ mode is activated in the listener, and top-down processes are used to help them decode the linguistic input more efficiently (Yi’an, 1998). However, we should bear in mind that there do exist situations where the linguistic knowledge a listener has “is so low that no amount of strategic behaviour can compensate and overcome the comprehension problem” (Fung & Macaro, 2019, 4).

In this respect, the importance of vocabulary in understanding aural input, particularly in L2 classrooms, is clear because no compensation strategy is an adequate substitute for the vocabulary knowledge (Milton, 2009). First, I will address the negative effects of insufficient vocabulary knowledge on the listening performance of L2 learners, and attempt to account for the reasons for that insufficiency. Then, I will discuss some research evidence that supports the positive relationship between vocabulary knowledge and listening performance in L2 environments. The section will then highlight the importance of studying that relationship among L1 Spanish learners of English, the target population of

my research study.

2.3.1 Inadequate vocabulary size and listening performance

Despite the problems that listening causes among L2 practitioners, and the importance of teaching our L2 students how to develop their listening competence in their target language (section 2.1), the factors that affect this skill have traditionally received little attention in the literature (Graham & Santos, 2015; Mendelsohn, 2001; Vandergrift & Goh, 2012). Boyle (1984) was one of the first researchers to investigate which factors affect listening comprehension in L2 environments. He asked students to list the issues with the biggest impact on their listening comprehension, and they place knowing the vocabulary in a much higher position than their teachers did. Since then, research has abundantly highlighted the importance of vocabulary in listening comprehension (for example, Brown, 2006; Chang & Millet, 2014; Cheng & Matthews, 2018; Field, 2008a; Fung & Macaro, 2019; Hulstijn, 2003; Kelly, 1991; Matthews, 2018; Milton, 2009, 2013; van Zeeland, 2014b; Wang & Treffers-Daller, 2017).

The biggest problems L2 learners might have when they listen are text problems, the difficulties that derive from lacking the necessary vocabulary, or from their inability to recognize an already known word within rapid connected speech (Cross, 2009). Furthermore, not knowing the words might be the most important obstacle to auditory comprehension (Field, 2008a; Kelly, 1991). As we have seen in section 2.2.3, if the listener does not know a word, it might be more difficult for them – or even impossible – to notice that word, or to determine where the word begins and ends, or to parse it onto a lexical unit and retrieve its meaning. The cognitive load

theory claims that understanding a text when there are too many unknown elements in it – particularly when those elements are highly interactive with each other – will imply a heavier intrinsic cognitive load (Paas & Sweller, 2014). If we accept that our working memory is limited in the number of elements it can process simultaneously, and in the duration of that processing, we might assume the existence of situations where the load is excessive to process. Alternatively, it seems plausible to accept that the more elements are stored in long-term memory, the lower the chance of finding novel information items in a text and, therefore, the lower the chance for our working memory of suffering a cognitive overload (section 2.2.3.2). In this respect, the cognitive load theory might provide an additional source of rationale to justify the exploration of correlations between inadequate vocabulary levels and poor listening performance.

L2 learners sometimes feel anxious when they listen to native speakers and think that they “speak too fast [or] swallow their words” (Field, 2009, 27). They might even complain about being unable to understand most of the input in a listening task, although they can later recognize and understand the same words in the corresponding transcript of the recording (Cai & Lee, 2010; Goh, 2005; van Zeeland, 2014b). One possible explanation for this phenomenon might be that students tend to identify knowing a word with just knowing what it means and recognizing its written form, neglecting how the word is pronounced or acoustically perceived (Nation, 2001). This phenomenon might lead some learners to be completely unable to comprehend connected speech in L2 even if they do know all the words in their written form (Bonk, 2000). Therefore, researchers, teachers and learners should assume that knowing a word might also imply being able to recognize it within a spoken text (van Zeeland, 2018).

There might be a further explanation for this inability of some L2 learners to recognize the words in connected speech. They might tend to learn a set of citation forms or perfect phonemes, which are likely to present different acoustic qualities when they are embedded into connected speech (Field 2009). In connected speech, the acoustic signal might be inconsistent when compared to citation forms, causing a great deal of variation depending on the context it occurs, which might lead L2 listeners to make an extra effort to recognize it. A listener might need to be flexible enough to accept that some forms may sound differently depending on their neighbours in the acoustic stream (Bonk, 2000; Field, 2008a).

The fact that some language learners are unable to notice or decode words when they are perceived acoustically indicates the existence of two different vocabulary knowledge dimensions: written and aural. Research has claimed that being able to recognize a word in its written and aural form is different (McLean et al., 2015; Milton & Hopkins, 2006), and should be assessed separately (Cheng & Matthews, 2018; van Zeeland, 2017; Zhao & Ji, 2018). However, apart from the present study, only one investigation has attempted to study those differences on the same population in an empirical study (Masrai, 2020).

This section has discussed how insufficient vocabulary knowledge might impact negatively on the performance of L2 listeners. This negative impact is particularly acute among lower proficiencies in the target language (Bonk, 2000; Fung & Macaro, 2019; Goh, 2000; Matthews, 2018; Tsui & Fullilove, 1998). The following section focuses on research claiming that having an sufficient vocabulary knowledge in the target language is indicative of adequate levels of listening comprehension.

2.3.2 Positive effects of adequate vocabulary size on listening performance

Research studies have shown a strong positive correlation between being a proficient listener and efficiently accessing a large vocabulary (Andringa et al., 2012; Matthews & Cheng, 2015; Milton et al., 2010; Stæhr, 2009). These studies have supported the claim that sufficient listening comprehension levels are clearly related to a higher familiarity with the words in the spoken text (section 2.2.3.2); whereas limitations in vocabulary knowledge seldom co-occur with those comprehension levels (Bonk, 2000; Goh, 2000).

Alderson (2005) generalized this positive correlation and claimed that L2 learners' vocabulary size is largely responsible for their overall language ability. He studied the correlation between scores in a vocabulary test and other language skills, and set that correlation at .61 in the case of vocabulary and listening (Alderson, 2005). In a similar line of research, other studies have shown that L2 learners' vocabulary size might be able to explain the variance in their listening comprehension scores in percentages that range from 23% (Bonk, 2000) to 65% (Masrai, 2020). Furthermore, in a recently published study, Masrai (2020) has investigated the joint contribution of both aural and written vocabulary to explain the language learners' listening ability. Both vocabulary measures were significant predictor, although the impact of written vocabulary knowledge was "marginal" when compared to the predictive power of listening success that the aural vocabulary size showed (Masrai, 2020, 22). The differences that the present study has found between aural and written vocabulary will be presented and discussed in sections 4.2, 4.4, 5.1 and 5.3 of this dissertation.

Moreover, this positive influence of vocabulary on listening comprehension seems to be particularly relevant among students with lower proficiency in the

target language (Pan et al., 2018), and might explain a large percentage of the variation in their ability to infer the meaning of unknown vocabulary in a text (van Zeeland, 2014b). These high figures might have led other researchers to consider language learners' vocabulary size a good indicator of their listening success (Cheng & Matthews, 2018; van Zeeland, 2018). The actual impact of learners' vocabulary size on their ability to understand and produce texts in a second language will be the focus of Section 2.5.1.

2.3.3 Vocabulary and Listening among L1-Spanish Learners

Another argument in favour of investigating the beneficial relationship between L2 vocabulary knowledge and listening refers to the target population in the present study: L1-Spanish students enrolled in intermediate-level classes of English. My experience with Spanish-speakers learning English as a foreign language has shown me that the B1 level might be considered a 'cut-off point'. Achieving this level of language proficiency might serve to distinguish between those learners who make further progress and reach an advanced level of proficiency in English, and those who will keep on floating within the same level for several years. Many of the learners who have reached this *plateau* of learning are unable to communicate in English, despite the many years they might have been studying the language (Yi, 2011). Consequently, teachers and researchers could focus on those aspects that previous investigations have highlighted as troublesome for intermediate-level learners of English. Among the language issues to focus on, Richards (2008a) recommended helping those students develop their L2 vocabulary and listening proficiency. A similar call for enhancing our students' vocabulary learning can be found in other studies (Fung & Macaro, 2019; Pan et al., 2018).

The need for vocabulary learning and listening instruction might be particularly acute among L1-Spanish learners, the target population of my research study. More than 2,000 Europeans with different L1s participated in a study of their vocabulary size in English, and one of the findings was that the scores in the vocabulary tests were comparatively worse among the learners whose mother tongue was Spanish (Alderson, 2005). Moreover, another survey carried out on school students in Spain showed that less than 15% of them had reached the level of a B2-user in listening, and that almost a third were placed below the A1-level in that particular skill (European Commission, 2012).

A possible reason for those results might be that the amount of input, output, and interaction in the language classroom in the European Union is insufficient (Suzuki, Nakata & Dekeyser, 2019), and that the time allocated to foreign language instruction is relatively small (Baïdak et al., 2017). However, the data published by the European Commission (2012) reveals that those secondary students in Spain might have already had about 1,500 hours of contact classes and practice in English by the end of their primary education. Furthermore, guidelines on the Common European Framework of Reference (CEFR) claim that an average learner might need about 100 hours of work – in the classroom and outside it – to achieve the initial level of A1. After about 1,300 hours of classes and practice in the target language, this same average learner should be able to achieve a B2-level (Cambridge University Press, 2013).

Although the present study is confirmatory in its nature (section 3.1.2) and it is beyond its scope to investigate the reasons for the current situation of L2 listening in Spain, one obvious conclusion of matching these data to the results at secondary schools in Spain is that something might be wrong with the way students are taught, when almost a third of them are below the A1 level in

listening. The reason for those poor results might not be that too little time is allocated to teaching L2s at school (Baïdak et al., 2017), or that an insufficient amount of input, output and interaction is offered in the language classrooms in Spain (Suzuki et al., 2019), as all those learners have attended more than 1,000 class hours. In any case, L2 learners in Spain need more research that supports the design and implementation of effective methodologies to help them be more proficient listeners (section 2.2.2). In the particular case of the possible impact of learners' vocabulary size on their ability to comprehend aural texts, this is the first study to investigate this relationship among adult L1-Spanish speakers.

This section has addressed the relationship between the language learners' vocabulary size and their ability to understand spoken texts. There is a significantly positive correlation between vocabulary and listening performance in L2 (Andringa et al., 2012; Matthews & Cheng, 2015; Milton et al., 2010; Stæhr, 2009). Both the negative effects on listening when learners have inadequate lexical levels, and the beneficial influence of a sufficient vocabulary size have been addressed. Interestingly, research has claimed that the impact of lexical knowledge is particularly relevant among those learners with a lower level of proficiency (Pan et al., 2018). The section has finished by referring to the particular relevance of investigating the relationship between vocabulary knowledge and listening performance among L1-Spanish learners of English as a foreign language, the target population in the present research study.

2.4 – KNOWING A WORD – TEXT PROFILING

Some studies mentioned in this chapter have based part of their claims on estimating the vocabulary size of groups of language learners. Before discussing the instruments employed in the past to assess that knowledge (section 2.5), it is necessary to address the literature on what knowing a word might entail, and what unit research might use to ‘quantify’ the vocabulary size of a person.

Establishing the unit of quantification is essential in any study that intends to compare results both within elements of that investigation and with similar studies. However, some experts in L2 vocabulary research have claimed that the counting units used in previous vocabulary studies have been imperfectly defined in the literature (Schmitt et al., 2017). Others have highlighted the idea that researchers into L2 vocabulary need to agree on the vocabulary unit under investigation (Bauer & Nation, 1993); or insisted on the importance of clearly reporting the unit of counting (Gyllstad, 2013).

2.4.1 Lexical units in vocabulary studies

For the sake of comprehension, it might be necessary to explain now the standard definitions found in the literature on vocabulary and analyses of texts for different lexical units (Milton, 2009; Nation, 2001, 2006, 2016; Stæhr, 2008). Token refers to all the words that are to be found in a given text, regardless of their form or the number of times they are repeated in a text. Token, in this sense, is synonymous with ‘running words’.

Types are all the different words that can be found in a text, so that repeated instances are counted in this case as one. Therefore, in the sentence ‘*The*

student studied the new words with the rest of the class, there are 12 tokens and 8 types, because the term '*the*' repeats itself four times.

Lemmas are the result of grouping the types from the same headword and their basic inflections like plural, third person for present simple, or past tense and past participle (Nation, 2001). Lemmas usually group items with a common part of speech, so the sentence '*Workers usually complain about having to eat at work*' has 9 tokens, 9 types and 8 lemmas ('workers' and 'work' are nouns here). Lemmas might be the preferred unit of counting when productive vocabulary is to be assessed (Webb, Sasao & Ballance, 2017), as test-takers can be told to use the correct part of speech for a given context, as in 'work', which can be a noun or a verb.

A further level in establishing a morpho-lexical unit is the word family. In this case, types and lemmas that are similar in their morphology are grouped together. For example, in the sentence: '*The singer sang a lullaby her mother used to sing to her*', there are twelve tokens, ten types, nine lemmas and eight word families because 'singer', 'sang' and 'sing' are considered members of the same family. The main difference with a lemma is that a word family might include different parts of speech as in 'strong' and 'strongly', whereas a lemma only includes elements with the same part of speech, as in 'heavy', 'heavier', 'heaviest' (but not 'heavily').

For receptive vocabulary, word families are the recommended unit of counting (Milton, 2009; Nation, 2016; Nation & Coxhead, 2014). Researchers assume thus that knowing one or two members of a word family facilitate the receptive knowledge of other members with little learning burden, i.e., little effort on the part of the learner (Nation & Webb, 2011).

Nevertheless, L2 vocabulary researchers need to decide how encompassing the term word family is going to be in their studies. Bauer and Nation (1993, 253) defined it as “a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately”. They distinguished up to seven possible levels of grouping within a word family, depending on the affixes used to modify the base word: the deeper the level, the more affixes are included, and the more encompassing the word family is. Altogether, Bauer and Nation came up with 91 possible ways to modify a base word. In level 1, each different form of a word is considered a word family, whereas in level 7, derivation includes classical roots and affixes like in ‘Francophone’ or ‘embolism’ (Bauer & Nation 1993).

Once the different levels of grouping for a word family have been determined, investigators could analyse texts, and assess how often each word appeared. The result of analysing a compilation of thousands of texts (i.e., a corpus) is the creation of wordlists based on frequency of occurrence. Examples of those lists are Nation’s 1-14k based on the British National Corpus (Nation, 2006), or his more recent compilation of the 25,000 most frequent words in English based on the BNC and the Corpus of Contemporary American English (Nation, 2012, 2019). In both cases, Nation grouped words up to the sixth level in Bauer and Nation’s typology (1993), considering thus that classical roots and affixes turn the base word into a different word family. When investigations into L2 vocabulary employ websites and software packages like *Compleat* (Cobb, 2019) for text profiling based on those frequency lists, we need to assume that those studies consider all words up to level 6 (Bauer & Nation, 1993) members of the same family.

Researchers have often claimed that language learners find it relatively easy to

know the form of other members of a word family and understand their meaning (Bauer & Nation, 1993), particularly when their derivations and inflections are very common (Milton, 2009). Learners who already know one member of a word family are expected to recognize most of the other members (van Zeeland, 2018), even if their linguistic proficiency is minimal (Beglar & Nation, 2007). However, there might be a lack of consensus about which derivations should be included in a word family (Stæhr, 2008). Furthermore, the facilitating effect of knowing one word to learn the other members of its family might only be applicable to receptive knowledge (Schmitt & Zimmerman, 2002). Therefore, the argument of the little effort to learn new members of the same word family might hold true in situations where productive knowledge is not required (Webb et al., 2017).

Word families seem to be the most readily operational category to assess language learners' receptive vocabulary knowledge. However, there are some situations where the use of word families as the unit of counting might be less straightforward than initially thought. Polysemy, homoforms and proper nouns are examples of single-word mismatches. Multiword nouns and verbs, as well as formulaic language might also fail to find a match in the operationalisation of the word family as the unit of counting.

2.4.2 Mismatches in word families – Single-word items

This subsection focuses on three of those special cases unable to find an exact match within the word families. It deals with polysemy, understood here as a single word with more than one meaning like 'sweet face' and 'sweet taste', homoforms – including homonyms, homographs and homophones – and proper

nouns, those of person, place, or thing in particular (Richards & Schmidt, 2002).

2.4.2.1 Polysemy

In the case of polysemic words, it might be necessary to determine what meanings are considered, especially when it comes to using them in vocabulary tests. Previous vocabulary tests and their corresponding research reports might have addressed polysemy in a slightly vague manner. For example, Schmitt et al. (2001) randomly selected word families for their Vocabulary Levels Test (VLT), checked their frequency of the family members, and included the most frequent one in the test. The items eventually selected for their vocabulary test maintained the ratio that reflects the distribution of word classes in English. Therefore, they included nouns, verbs and adjectives with this ratio: 3 (noun): 2 (verb): 1 (adjective). However, the test designers dealt with polysemy just by including the “most frequent meaning sense” (Schmitt et al., 2001, 63) for each item in the test, without presenting any evidence of how they achieved that.

Similarly, the Vocabulary Size Test (VST) from the part of speech and meaning that best reflected the “highest frequency environment” in each word family (Beglar & Nation, 2007, 12). Furthermore, the listening vocabulary size test (LVST), where subjects have to listen to the target word, first in its isolated form, and then in a sentence is equally unspecific on how to deal with polysemic words. The choices for the test takers to choose from had been translated into their L1 (Japanese), but the designers of this vocabulary test only mentioned that the correct option would have “the closest meaning to the English word being read” (McLean et al., 2015, 758).

The three tests mentioned above (VLT, VST and LVST), as well as most of the

vocabulary tests used in L2 research to assess vocabulary knowledge, select their items from vocabulary lists compiled according to the frequency of occurrence of their items in the target language (Nation, 2006; Nation, 2012, 2019). Those lists are the result of frequency analysis of large corpora like the BNC or the COCA, which comprise hundreds of millions of words from thousands of written and spoken texts. With respect to polysemy, those wordlists just present the headword (Nation, 2017), so that none of the different parts of speech of the same item, nor its possible meanings are considered. Apart from randomly selecting the items for their vocabulary tests, researchers also need to decide which of the possible meanings they are referring to.

Research reports on the assessment of receptive vocabulary should clearly inform not only about which word families or lemmas they have selected for their tests, but also about how they have decided which of the possible meanings they are testing in each of the items. Proceeding in this manner might facilitate the work of other researchers who want to compare results across investigations, or replicate studies with slight modifications.

2.4.2.2 *Homofoms*

Homofoms include three possible realisations: homonyms, homographs, and homophones. Homonyms are words that present the same spoken and written form, but with different and unrelated meanings. A 'ball' might be a round object and a formal party where a dance is involved. From a semantic point of view even expert linguists find it difficult to tell the difference between homonymy – two or more words written and pronounced alike, but with different meanings – and polysemy – a single word with more than one meaning (Richards & Schmidt, 2002). It is certainly beyond the scope of this research study to

address this question in depth and therefore, for the sake of comprehension, we might need to assume that the difference lies in the proximity of the different meanings (Parent, 2012). In this respect, the different meanings of 'sweet' depending on its object – 'sweet face', 'sweet voice', 'sweet taste' – might indicate polysemy. On the other hand, the realisations of 'can' – modal verb and a cylindrical container for drinks – might lead us to consider them a clear case of homonymy.

Homographs are words that are written in the same way, but pronounced in a different manner and with a different meaning. A 'bow' /bau/ refers to the show of respect by means of bending forward the head or the upper part of the body. However, when it is pronounced /bəʊ/ it refers to the weapon for shooting arrows (Summers & Gadsby, 1987). On the other hand, homophones are words that, although being written differently, are pronounced in the same manner, but show unrelated meaning, as it happens in 'pie' and 'pi'.

As word frequency is based on the written form of words, homophones are easily dealt with because they are assigned to two different word families. On the other hand, the actual realisations of homonyms and homographs are grouped as one single word. This decision might imply an extra learning burden as the learner is supposed to know not only the different forms included in a word family, but also the different meanings each of the forms might show.

From the point of view of vocabulary test designers, homonyms and homographs imply a similar challenge to the one posed by polysemic words. Test designers might need to specify which part of speech they are using ('work' as a noun or as a verb), and the meaning they are referring to ('play a role', 'play football', 'play the guitar', etc.). By clearly stating their policy on polysemy, homonyms and homographs, and by subsequently following it, they

might exclude any possible bias in the selection of items, or in the overall test design. In this respect the VLT (Schmitt et al., 2001), the VST (Beglár & Nation, 2007), and the LVST (McLean et al., 2015) failed to explain how they would deal with polysemy, homonyms, and homographs.

2.4.2.3 Proper nouns

Proper nouns like ‘Richard’, ‘Newcastle’, or the ‘National Healthcare System’ differ from common nouns, which are directly associated to a class or to a specific entity of a given class. They are usually excluded from text analyses based on frequency not only because they are not considered part of the vocabulary any language user should learn, but also because of feasibility reasons. Proper nouns, including those to name people (patronyms) and places (toponyms), might be too varied to be included in frequency lists. For example, a mere list of English last names yields a result of more than 14,000 different surnames (Family Education, 2019). In the case of toponyms, the list will certainly be longer. Furthermore, the possible inclusion of those proper nouns in lists based on corpora raises the issue of whether to include only those associated to the English-speaking countries, or extend the inclusion to all of them. These proper nouns are assumed to be easily recognized and understood when they are encountered in a text because they usually refer to a particular instance of reality. This reference helps the language user differentiate that instance from the rest of similar entities in their class. Furthermore, the English language distinguishes the written form of proper nouns from common nouns by means of capitalizing their first letter, unlike other modern languages like German for example, where all nouns are capitalized in their first letter.

Moreover, the inclusion of proper nouns into frequency lists might imply a huge amount of work, and its benefit might be minimal. For example, Nation (2012, 2019) has compiled so far a list of 21,662 word families for proper nouns with a total of 22,409 instances at Level 6. This list includes entries like ‘America’ (‘American’, ‘Americanisation’, ‘Americanism’, etc.), ‘Anthony’, ‘Smith’, or ‘Newyork’ (all one word). But some other proper nouns are neglected because we can find family names like ‘Dicaprio’ or ‘Jolie’ but not ‘Deniro’ or ‘Blanchett’. This compilation might be considered work in progress as new entries are added to the list. But it also needs refining, as it includes acronyms like ‘NHS’ (National Health System) or ‘HSBC’ (The Hongkong and Shanghai Banking Corporation) that should be in the acronyms list with entries like ‘BMW’ (Bayerische Motoren Werke) or ‘NATO’ (North Atlantic Treaty Organization). Consequently, some authors have implicitly supported the view of not using proper nouns in text analyses because they considered that the compilation of those nouns is an ongoing endeavour that will never end (Nation & Webb, 2011).

2.4.3 Mismatches in word families – Multiword items

The previous section has focused on single-word items featured in a text that might not find a straightforward match within a word family. Now is the turn of those words that are featured in two or more occurrences like compound nouns, phrasal verbs, and formulaic language.

2.4.3.1 *Multiword nouns*

The inclusion of multiword expressions like compound nouns and phrasal verbs

in vocabulary lists based on frequency might be a more difficult issue to deal with. According to Bauer and Nation (1993) transparent compound nouns could be added after their Level 2 category, which includes inflectional suffixes (plural, past tense, comparative, etc). Nation's compilation (2012, 2019) is a list of 3,108 word families for these compounds (e.g., 'airplane' or 'backpack') with a total of 6,044 instances at Level 6. However, the compilation of transparent compound nouns is still an ongoing process that requires further work and refinement, so their availability for use in research is still limited.

As it happened with the headword and the other members of a word family, some researchers have also claimed here that the learner already knows these compound nouns because their parts are known, and their meanings are related to the overall meaning of the compound (Nation & Webb, 2011). These researchers think then that if a learner already knows the meaning of 'back' and the meaning of 'pack', they should easily come to the meaning of the compound 'backpack'.

The term *word family* is in italics here because of the difficulty of classifying a compound noun as only one word family and not as two separate families. Thus, the compound 'yearbook' could be ascribed either to the word family 'year', or to the one for 'book', or just constitute its own category (i.e., 'yearbook'). In this last case, the potential creation of derivations and inflected forms from the headword – as in 'backpack', 'backpackers', 'backpacking', 'backpacked', etc. – might be a factor to consider the compound a category on its own. Furthermore, a compound noun might be realised in two words separated by either a hyphen or a blank space in their written form (e.g., 'passer-by', 'credit card'), with obvious repercussions on the potential derivations and inflections ('passers-by', 'credit cards').

From the point of view of implementing a research study based on frequency lists, it might be reasonable to exclude the use of multiword nouns. First, because even if we agreed on what compound nouns to include and on how to count them, manual checks might be necessary. A second reason for the exclusion of compound nouns refers to the feasibility of their use. Frequency lists based on corpora tend to exclude all compound nouns because of the question of whether to count them as one or more word families. Even when lists of compound nouns are available (Nation, 2012, 2019), most text profiling websites and software packages (Cobb, 2019).

2.4.3.2 Multiword verbs

Another instance that might be 'problematic' refers to researching multiword verbs or phrasal verbs (Capel, 2010). Although these verbs consist of two or more words, they refer to a single lexical unit as in 'look forward to', or 'put up with'. However, corpus analyses tend to consider them instances of different word families ('look', 'forward', 'to'; 'put', 'up', 'with'). Despite constituting a unit of meaning, they fail to be included into the analyses as one word family in themselves, or as part of a given word family because of the same practical reason mentioned above for compound nouns. Their inclusion would imply analysing the corpora manually to find out all the possible instances of multiword verbs, and subsequently modifying the computer programs like *Compleat* (Cobb, 2019) to include all those multiword expressions as units of meaning.

Furthermore, transparency in some of those expressions might be difficult to find. Although some authors claim that focusing on the particles of those verbs might help in understanding their overall meaning (Side, 1990; White, 2012),

they might have neglected the phenomenon of polysemy in many multiword verbs. For example, we need to look at the objects in the sentence – and not at the particle – if we want to understand the multiword verb ‘work out’ in ‘*Susan works out at the gym every Friday evening*’, and in ‘*Susan needs a bit of time to work out a solution*’. The opacity of some multiword verbs has been considered “a major source of difficulty” for second language learners of English (White, 2012, 419). For example, understanding ‘face off’ as the beginning of a confrontation, or ‘chew out’ as a synonym for reprimanding someone might imply a major challenge for some language users.

2.4.3.3 Formulaic language

One last phenomenon of multiword units is that of formulaic language, which might be described as two or more words that match a single meaning. This matching might be transparent, as in ‘more and more’ that might refer to ‘increasingly’; or fall closer to the opaque end of the continuum, as in ‘learn by heart’ (Martinez & Schmitt, 2012). Research has considered that not accounting for formulaic language is a “serious limitation of the discussion” (Schmitt & Schmitt, 2012, 484), and it has claimed that wordlists based on frequency are deficient because they only feature single words, which might be just the “tips of phraseological icebergs” (Martinez & Schmitt, 2012, 302).

Martinez and Schmitt (2012) examined the BNC to determine the most frequent phrasal expressions in English, “a particularly opaque subset of formulaic language” (299) because the unique meaning of the whole multiword expression might not be discernible from decoding each of its elements. Although the compilation of 505 phrasal expressions (Martinez & Schmitt, 2012) is a good attempt to include this type of multiword expressions in the analyses

of texts based on frequency, it presents two main limitations. First, a closer look at the items included in the list might challenge the claims made by those authors about the opacity of the phrasal expressions. Among the 15 most frequent opaque multiword expressions in English, they included items such as 'a few', 'a lot', or 'a little', which might be perfectly understandable by having a look at its individual elements.

A further criticism about this compilation of phrasal expressions is the authors checked the validity of their list by analysing one single academic article, comprising only 2,172 tokens (Axelrod, Axelrod, Jacobs, & Beedon, 2006). They concluded that, assuming that a person knows only the words in the 2000-word family level (2k level), 7.46% of the tokens in that text might be considered 'off-list' words, words that are beyond the lists employed in the comparison, and therefore likely to be unknown to that person. If the comparison of the tokens from the same text is made while taking the opaque phraseology into account, the percentage of off-list words from the text increases to 26.87%. In other words, when multiword expressions like 'fall short', 'take account of', or 'missing the boat' were considered as individual words, all of them fell within the 2k level. However, in the second comparison those expressions were considered as meaningful sets of words, and fall within the group of off-list words (Martinez & Schmitt, 2012). Moreover, a further analysis of the article employed for the compilation of opaque phraseology (Axelrod et al., 2006) reveals that expressions featured in the text like 'fall short' or 'miss the boat', are not in the list compiled by Martinez & Schmitt (2012). Consequently, more evidence is thus needed to support the validity of that phrasal expressions list.

2.4.4 Reasons for the mismatches

The previous sections have addressed cases like polysemy, homoforms, proper nouns, multiword nouns, phrasal verbs, and formulaic language. We have also discussed why in those situations further considerations and compromises on the part of the researcher might be necessary. The use of frequency wordlists based on the analysis of corpora might be really useful for vocabulary research, but it has its limitations. These limitations are heightened because specific software is used in the compilations of frequency wordlists based on corpora, and in the comparisons between texts and those frequency wordlists. The linguistic knowledge of computers in this respect is limited: they can only judge if a given string of characters in the text is different from the next one, and if it has an exact match in any of the entries stored in their databases. The use of computers in the analyses of texts has reduced the concept of 'word' to a match on a list stored in a computer, ignoring the cases of homoforms, polysemy or proper nouns, and neglecting the inclusion of multiword units in the analyses (Cobb, 2013). When dealing with those particular cases, research studies need to assume that some single words might actually be two words, and some phrases might really be single words.

We have already discussed those situations in the analysis of the words in a text where the computer might find it hard to provide a clear answer. The very use of computers in text analyses is the reason why some of those items cannot find a match in the lists (Cobb, 2013). We have seen attempts to overcome this problem by providing computer systems with new wordlists of proper nouns, compound nouns (Nation, 2012, 2019), and multiword expressions (Martinez & Schmitt, 2012). But further research is necessary until valid and reliable wordlists are readily available to account for phenomena like polysemy,

homoforms, proper nouns or multiword expressions. We have to agree with the implicit call for further and more refined research into this matter as “the measurement of the size and knowledge of formulaic language is still in its infancy” (Schmitt, Cobb, Horst, & Schmitt, 2017, 2).

Two consequences might be drawn from the discussion in this section. Firstly, frequency lists should be redefined to account for homonyms, homographs, polysemic words, multiword verbs, compound nouns and phrasal expressions. We need to go beyond the space-defined word form to stop the current inaccuracy of frequency lists, and transform text profiling based on frequency into a more useful instrument in vocabulary research (Cobb, 2013; Gardner & Davies, 2007).

A second consequence of accepting the inability of text profiling instruments to account for all those phenomena refers to research studies using vocabulary lists and other instruments to analyse the profile or lexical density of a text. These investigations have to be aware of their limitations in this respect, and clearly address them in their research reports. Until more accurate instruments are made available to the researcher interested in using frequency lists in vocabulary studies, they might need to be cautious about the accuracy of some of the instruments used in their investigations.

In the present research study, multiword instances were excluded from the vocabulary tests and subsequent analyses because of the difficulty of operationalising them within the software framework available. In the case of polysemic words and homonyms, all the realisations included in the PET Vocabulary List had an equal chance to be selected for the vocabulary tests (section 3.2.3). Appendices 3 and 4 show the vocabulary tests with 150 items employed in the preliminary study in May 2019. Appendices 5 and 6 show the

refined version of the tests (81 items) employed in the main study, whereas Appendix 7 features some elements included in the PET Vocabulary List, downloaded from Cambridge Assessment English.

2.5 – ESTIMATING VOCABULARY SIZE IN L2

This section focuses the discussion on how previous studies have estimated L2 learners' vocabulary size. Once we have shown the positive correlation between lexical knowledge and listening performance (section 2.3.2), it might seem reasonable to address the possible ways of quantifying the size of that vocabulary, and how research has matched it to the actual understanding of spoken texts. Knowing the lexical coverage necessary to understand aural texts can provide us with useful information to estimate the vocabulary size a learner needs to function in a second language (Matthews, 2018).

2.5.1 Vocabulary Size, Frequency and Lexical Coverage

The breadth or size of vocabulary among language learners has been typically estimated through vocabulary tests, and then matched to the ability to comprehend texts, either written or spoken. By quantifying the approximate number of words a learner knows, and checking the frequency of the words featured in a text, researchers have set the minimum vocabulary size to understand different types of texts. Moreover, research has claimed that word frequency is the best measure available to assess the lexical quality of a text (Crossley, Cobb & McNamara, 2013), that the actual frequency of a word in a language might correlate with other dimensions of learners' linguistic proficiency, and consequently, that frequency should guide the selection of words for learners to study (Hazenberg & Hulstijn, 1996).

A research area on word frequency has focused on analysing the influence of the vocabulary featured in a text on the ability of a language learner to understand it (Bonk, 2000; Hirsch & Nation, 1992; Stæhr, 2009; van Zeeland,

2018). The lexical density of written or spoken texts has been assessed according to the frequency of the words those texts feature, and then matched to the comprehension shown by a group of L2 learners in different tests, and to their receptive vocabulary knowledge. In this assessment of written or spoken texts, research has understood the density of a text as a synonym for its possible difficulty, based on the assumption that the more frequent words are, the more likely those items are to be known by the average language user, and the easier they render the text to be understood. The subsequent analyses of the results have led researchers to set minimum levels to achieve comprehension, either in reading or in listening.

At least three aspects might be considered when determining the lexical density of a text. The first factor is the type-token ratio, i.e., the number of separate words in a text divided by the total number of words featured in that text (Richards & Schmidt, 2002). Another approach to the assessment of the lexical density of a text will set the focus on the intrinsic density of each element featured in a text by examining issues such as polysemy, register, imaginability, tangibility, etc. For example, a word like 'sport' might be considered very frequent in English, although some realisations like 'old sport' might have certain degree of added difficulty that is not shown by the sheer frequency of the headword. Words referring to abstract concepts might be more dense – i.e., more difficult – than other words that are easier to be pictured in our minds, regardless of how many times they appear in a given text, or how frequently they are used in the language. However, in most cases research has equated lexical density to frequency and compared the words featured in a text to vocabulary lists based on frequency. In fact, both the investigations cited in this dissertation and the present study itself primarily understand the construct of

lexical density in this manner.

How much vocabulary is then necessary to understand a text? The answer to that question is not as straightforward as it might seem if we analyse the claims made by previous research. For example, some studies have suggested that knowing 95% of the words in a text is sufficient to allow “reasonable” reading comprehension – i.e., scores of 55% or higher in a test (Laufer, 1989, 321). Hirsch and Nation (1992) increased the coverage of the words in a text to 97-98% as the necessary minimum for pleasurable reading. Other subsequent studies have recommended a lexical coverage of 98% to enable enough understanding of a written text (Hu & Nation, 2000; Nation, 2006), to achieve 68% of correct answers in a reading comprehension test (Schmitt, Jiang & Grabe, 2011), to read independently in academic settings (Laufer & Ravenhorst-Kalovski, 2010), or when “very high comprehension is aimed at or more difficult text types are used” (van Zeeland, 2018, 2).

For listening comprehension, Bonk (2000) set the minimum lexical coverage at 90% of the words featured in that aural text. He claimed that listeners might be using other resources and strategies that enable them to comprehend spoken discourses far beyond what their actual vocabulary size could predict. Stæhr (2009) found that with a coverage of 94% of the words in an aural text, the mean listening comprehension scores were 60%; whereas people who knew 98% of the words were able to reach a mean score of 73% (Stæhr, 2009). In another study, van Zeeland and Schmitt (2013b, p. 457) set that minimal coverage for “adequate” comprehension of a spoken text at 90% of all its words, although they recommended knowing 95% of the words to avoid variation in the comprehension levels.

There exist some variation in the minimal percentage of words a person has to

be familiar with to operate adequately in L2 reading, or listening. Those lexical coverages vary not only depending on the language skill, but the studies also differ in their recommendations for the same skill. An increase from 90% to 95%, or from 97% to 98% might seem relatively small, but in practice it might imply learning thousands of new words. For example, increasing the coverage from 95% to 98% would imply passing from a vocabulary size of 2,000-3,000 word families to 6,000-7,000 word families (van Zeeland & Schmitt, 2013b). Such an increase would bear a clear impact on the demands placed on the learners, so percentages about the minimum knowledge necessary to achieve comprehension or function adequately in the target L2 have to be highly accurate. This accuracy relies on two separate measurements: research has to be precise when assessing the amount of vocabulary a language user has, and when analysing the lexical density of a text in terms of the frequency of its words.

Once we have explored the concepts of both lexical density and lexical coverage in the literature, the discussion will focus now on the way previous studies have estimated language learners' vocabulary size. A great deal of the following section will address the validity and reliability of those estimations, because analysing the quality of the research studies and their instruments may help "distinguish research studies from conjecture or opinion" (Heigham & Croker, 2009, 38).

2.5.2 Vocabulary Testing and Listening Comprehension

The construct of vocabulary knowledge implies several components that could be grouped into form, meaning and use (Nation, 2001; Milton, 2013). It might be

impossible to assess all those components with the same instrument, so researchers have to decide which ones are the most relevant for their study. This section addresses only the vocabulary tests used in the past to assess language learners' receptive vocabulary size with respect to their listening comprehension. Those research instruments seem to yield better correlations with listening comprehension, and explain more of its variance than other tests assessing the L2 learners' vocabulary depth, i.e., the different aspects of form, meaning or use they know (Wang & Treffers-Daller, 2017).

As well as showing a lack of consensus in the minimum figures to achieve comprehension (section 2.5.1), most studies in the past have only assessed learners' ability to recognize the link between the written form of a word and its meaning (Adolphs & Schmitt, 2003; Andringa et al., 2012; Hirsch & Nation, 1992; Nation, 2006; Stæhr, 2009). Although those tests are considered a "measure of written receptive vocabulary size" (Beglar & Nation, 2007, 11), research has employed their estimations for correlations to listening comprehension. Furthermore, those studies have made use of written vocabulary tests despite the claims that learners' ability to recognize words in their written and spoken forms might be different (Zhao & Ji, 2018), and consequently they should be assessed separately (Cheng & Matthews, 2018; van Zeeland, 2017; Zhao & Ji, 2018), so that the aural vocabulary knowledge is emphasized as the "primary construct of relevance" (Matthews, 2018, 24).

Comparatively, very few studies have employed listening vocabulary tests to estimate their participants' vocabulary size. The use of this type of vocabulary test might increase the correlation figures between vocabulary size and listening comprehension (Milton et al., 2010; Stæhr, 2008). Furthermore, it can also offer valuable perspectives into vocabulary and listening, because being able to

recognise words from speech is vital to L2 listening (Wang & Treffers-Daller, 2017).

2.5.2.1 *Unsuccessful attempts to assess the aural vocabulary size*

One of the first examples of a listening vocabulary size test – in the form of a dictation – was created by Fountain and Nation (2000). The target items for that test were selected for frequency, and included in a slightly longer text. The marking procedure only focused on the correctly spelled forms of the target words, neglecting their actual position in the sentence, and ignoring errors “with the regular -s, -es, -d, and -ed suffixes” (p. 33). This marking procedure might be indicative of confounding variables (McLean et al., 2015), as it identifies listening vocabulary knowledge with both recognizing the aural form of the words, and spelling them correctly.

Similarly, other research studies have tested the aural vocabulary knowledge by means of a dictation test with target words selected from frequency bands based on the BNC-COCA (Cheng & Matthews, 2018; Matthews, 2018). These studies attempted to avoid the possible confusion of two variables (recognition of words in connected speech and their correct spelling) by using a rubric to categorize minor spelling errors, and systematically assign marks to different levels of word recognition (Matthews, O'Toole & Chen, 2017). Based on the high levels of inter-rater reliability reported on the use of this marking scheme, it seems that the threat to the validity of the construct (word recognition) might have been avoided. In fact, research studies have claimed that this type of test might be a good instrument to assess productive phonological vocabulary knowledge (Cheng & Matthews, 2018).

Two main criticisms could be made of dictation exercises to test aural vocabulary size. First, and most importantly, this type of aural vocabulary test identifies knowing a word with just being able to recognize its aural form and produce its written form, without having to provide evidence of any link to its meaning. L2 learners with some proficiency in the target language phonology might be able to recognize and transcribe L2 words they have just encountered for the first time, particularly those words that are similar in form to their L1. However, they might fail to make any further sense of them within a broader discourse, which is the ultimate goal of listening comprehension (section 2.1.1).

The second limitation in using dictation tests for the assessment of aural vocabulary size refers to the way the answers are elicited from the test-taker. One of the reasons for the difficulty of listening when compared to reading is that there are no spaces to determine the end of one word and the beginning of the next (section 2.1.2). In those vocabulary tests (Cheng & Matthews, 2018; Fountain & Nation, 2000; Matthews, 2018), the test-takers have to write the target word within one blank, with other words before and after. Those boundaries are really helpful to the listener to anticipate when to focus their attention on the stream of words, and for how long. The ecological validity of the instrument is thus negatively affected (section 3.1.2.3), as it differs from what a listening situation actually demands from the listener.

Other research studies have used an aural vocabulary test like the Peabody Picture Vocabulary Test, where the test-taker has to recognize each target item from a series of pictures. Although this instrument to measure vocabulary knowledge is considered to be highly reliable and a “more valid measure of oral receptive vocabulary than most vocabulary tests” (Vandergrift & Baker, 2015, 401), their administration on a one-to-one basis makes it really inconvenient

and time-consuming. Furthermore, the target items in this test are pronounced in an isolated manner, without revealing what part of speech words like 'work' might refer to. Besides, among L2-populations it might be difficult to check if the learner is able to link the recognized aural form to its correct meaning.

Aural versions of word-recognition tests such as the Aural Lex or the Y_Lex test (Meara & Miralpeix, 2006) have been used in other studies (Milton & Hopkins, 2006; Milton et al., 2010; van Zeeland, 2014a). This kind of Yes-No tests present L2 learners with words pronounced in an isolated manner, and they have to decide if they *know* the word. The test also shows the learner nonwords which follow the same phonotactic rules as the target language. For every false positive, i.e., a word that a test-taker claims to know but is inexistent in the target language, a percentage is subtracted from the overall vocabulary score. The introduction of those control words aims to minimize the possible impact of carelessness and guessing.

Three aspects of this Yes-No tests might be criticized. The first refers to the fact that the test-takers themselves decide if they know the target words. Secondly, the absence of a clear criterion about what knowing the target words implies. It could be just being sure that the word exists in the target language, or it could be that they can recall their meaning, or maybe it could mean being able to use it correctly in a sentence. Since the inclusion of nonwords in the test is the only manner to control that the test-taker is being accurate in their judgements an overestimation in the results might occur (Eyckmans, 2004; van Zeeland, 2014a). A final criticism refers to the aural version of this Yes-No vocabulary test. The test is usually done on a computer, and its test-takers can play the target word as many times as they wish, and take as long as they want to answer each question (McLean et al., 2015). This might not be the case in real-

life situations, where listeners need to process the spoken text almost immediately (Field, 1999), and most of their listening success depends on not having to ask for repetitions or clarification. All these alleged flaws in the design of this type of vocabulary tests might have contributed to overestimations of learners' vocabulary size as high as 34.6% (Eyckmans, 2004; van Zeeland, 2014a).

2.5.2.2 *Listening Vocabulary Size Test (LVST)*

A possible way to avoid those errors in the estimation of L2 learners' vocabulary size might be to create the aural version of an already existing written vocabulary test which has provided evidence of its efficiency. For example, McLean et al. (2015) employed a similar format as the Vocabulary Size Test (Beglar & Nation, 2007) for their new Listening Vocabulary Size Test (LVST).

The Vocabulary Size Test (VST) presents each target word in its written form, both separately and within a short sentence to determine the part of speech the target item refers to. Test-takers have four short sentences from which they have to choose the best match for each of the target items. There are 140 items in the test, and each band with a thousand of the 14,000 most frequent words in English according to the BNC is represented by ten target items. That relative frequency is based on the wordlists compiled by Beglar and Nation (2007) from the British National Corpus (BNC).

McLean et al. (2015) created the LVST, a recorded version of a vocabulary test where the items were pronounced individually, and then repeated within a short sentence, simply to determine their part of speech. Furthermore, they selected the target words from a new set of wordlists compiled by Nation (2012, 2019)

from the BNC/COCA. Although these wordlists cover the 25,000 most frequent words in English, the LVST only employed items from the first 5 bands (1-5k), and from the Academic Word List (Coxhead, 2000). They also translated the four options for each of the target words into Japanese, the L1 of the participants in their investigation. Subsequent analysis of the LVST showed high levels of reliability, and clear signs of validity (McLean et al., 2015). Appendix 1 shows the first items in the test.

Employing a multiple-choice format, and translating the options into the test-takers' L1 might also raise concern about its validity and reliability (Nation, 2001). Firstly, the additional cognitive load on the test-taker, which might have an impact on their performance in the test. Learners are asked to switch between their L1s and the target language, as the prompt or question is in their L2 and the answers to choose from are presented in their L1. Secondly, presenting four options to the test-taker to choose from, and assuming that knowing a word is just being able to select the right choice has been considered "simplistic [and] questionable" (Huang, 2010, 4).

The alleged cognitive load that might derive from the use of translations in this kind of vocabulary tests has failed to be detected in qualitative analyses performed by McLean et al. (2015). Furthermore, a multiple-choice format for vocabulary size estimations has proven to be a valid and highly reliable method according to quantitative and qualitative analyses (Beglar, 2010; Silva & Otwinowska, 2019). Moreover, vocabulary tests that identify knowing a word with being able to recognize its form and match it to a meaning should be favoured instead of criticised. The reason is clear: "the form-meaning link is the first and most essential aspect which must be acquired" when studying L2 vocabulary (Schmitt 2008, 333) because it is "a fundamental first step in gaining

control over a particular word” (Cheng & Matthews, 2018, 4).

Several reasons support the use of translations in this kind of vocabulary test. Firstly, translations of the target words into the test-takers’ L1 might facilitate the creation of replication studies in other parts of the world, with different target languages. The scarcity of replication studies in applied linguistics in general, and in L2 acquisition in particular, is considered one of the most serious problems the discipline has to face, because they are crucial in the promotion of transparency and collaboration in research (Abbuhl & Mackey, 2017). Fortunately, several research studies into vocabulary testing have contributed to mitigate that scarcity of replication in the field, by implementing different bilingual versions of vocabulary tests (Karami, 2012; Nguyen & Nation, 2011; Zhao & Ji, 2018).

Secondly, as bilingual vocabulary tests present the words in the target language and the options in the test-takers’ L1, they might provide “feasible alternatives to more challenging and time-consuming monolingual tests” (Nguyen & Nation, 2011, 86). In monolingual versions of a multiple-choice vocabulary size test, the options are written in the target language in the form of a broad definition, a paraphrase, or a description. Test designers have to be extremely careful in those sentences, and use words that are actually more frequent than the target item. This precaution might be impossible to maintain when testing the knowledge of very frequent words (Beglar & Nation, 2007). Furthermore, test-takers with lower levels of proficiency in the target language, might be unfamiliar with some syntactic structures used in those definitions, which might result in testing additional aspects of that language, apart from just their vocabulary size (Nguyen & Nation, 2011). Consequently, when beginners or low level learners are included among the target population for a study, the use of bilingual tests

might be preferable (Nation, 2007; Levitzky-Aviad & Laufer, 2013), because the respondent's ability to recognize the target items – which should be the focus of vocabulary tests – is not confounded with their ability to read answer options in the L2 (Wang & Treffers-Daller, 2017).

This section has addressed different ways employed in the past to estimate the vocabulary knowledge of L2 learners, especially their aural vocabulary size. The discussion has focused on the validity and reliability of those studies because the accuracy in the estimations has a clear impact on the overall precision of those studies that focus on the lexical coverage to achieve comprehension in L2 (section 2.5.1). The following section will present a summary of the gaps detected in this literature review as well as an account of the way they have been addressed in the present study.

2.6 – BRIDGING GAPS

This literature review implicitly serves the purpose of accounting for the instruments employed in the present study to quantify how big the vocabulary of L2-English learners is, and to determine how influential this size might be in their listening performance. Chapter 3 – Methodology and Methods – will address the most relevant decisions taken in the planning, implementation and analysis of this research study, meant to investigate the relationship between vocabulary and listening among L2 learners.

The gaps detected in the literature review of L2 vocabulary and listening comprehension led to investigating these research questions:

- 1) How much of the listening performance in an exam might be attributed to knowing the words in a vocabulary list?
- 2) How much lexical coverage of a spoken text does a learner need to achieve comprehension in a listening test?
- 3) How similar are the scores in vocabulary size tests based on recognising either the aural or the written form of words?
- 4) How does the relationship between lexical knowledge and listening performance evolve over time?

The present study adds to the general body of knowledge in second language research because:

- 1) It is an empirical study with a clearly quantitative approach, which is less common in the field of applied linguistics.
- 2) It investigates a language skill that has received less attention than the rest.

- 3) It refuses to use written methods to investigate listening comprehension.
- 4) It uses a scarcely employed bilingual format to assess the form-meaning link in a receptive vocabulary test (e.g., Karami, 2012).
- 5) It intends to confirm the enhanced suitability of aural vocabulary tests in correlations with the language learners' listening performance (e.g., Stæhr, 2008).
- 6) It intends to bring more empirical evidence to the claim that there is a strong and positive correlation between the language learners' vocabulary size and their listening ability (e.g., Alderson, 2005).
- 7) It intends to bring more empirical evidence to the question of how much lexical coverage of a text is necessary to achieve listening comprehension: 90% (Bonk, 2000), 94% (Stæhr, 2009), 98% (van Zeeland & Schmitt, 2013b).

Moreover, this study aims to explore new territory in the realm of second language listening and vocabulary, and bridge several gaps detected in previous research:

- 1) It is the first one to use the same framework for the research instruments employed to study the two variables, vocabulary and listening. The validity of those instruments is enhanced with respect to previous studies, which have used receptive vocabulary tests and listening comprehension measures from different sources (e.g., Stæhr, 2009).
- 2) It is the first study to explain how mismatches in frequency lists (e.g., polysemy) are dealt with in the investigation.
- 3) It presents the first bilingual vocabulary test for L1-Spanish speakers, a potential population of 471 million (Eberhard, Simons & Fennig, 2021).

- 4) It is the first study to use two vocabulary tests – aural and written – especially created for the population under study.
- 5) It is the first study to use the same items in two vocabulary tests where the only difference is how they are delivered (orally or in writing).
- 6) It is the first study to estimate the actual differences between learners' aural and written vocabulary size. Unlike Masrai's study (2020), this investigation employs vocabulary tests with no validity and reliability issues that might lead to overestimations (Eyckmans, 2004).
- 7) It is the first study to investigate the relationship between L2 vocabulary and listening by delivering three tests to the same population at the same moment in time, and then after a period of about 35 weeks.

2.7 – CHAPTER SUMMARY

This literature review has begun with a brief introduction to the importance of second language listening and how it is perceived by learners, teachers, and researchers. This first section has also shown how a ‘comprehension approach’ (Field, 2009) has pervaded in most L2 classrooms (Vandegrift & Goh, 2012, 12), and in published methods available in the market (Siegel, 2015). A different pedagogy of listening is necessary, because many L2 learners are just being exposed to a series of recordings, and then tested in their comprehension by answering a batch of questions. Research has shown that new pedagogical approaches to listening should be based on the processes and components that entails the skill (section 2.2.1).

Section 2.2 has discussed the model proposed by Vandegrift & Goh (2012), where L2 listening is a complex skill that involves different sets of processes and information or knowledge sources (Figure 2.1). Firstly, it includes Anderson’s (2020) model of language comprehension with three steps in a highly iterative and overlapped process: perception, parsing, and utilization. Listeners can also draw on previous knowledge they may have stored in their memories to facilitate the process of comprehension. Depending on the direction of the processes involved – from the auditory input towards the representation in memory or vice versa – this listening model speaks of bottom-up processing or top-down processing. Additionally, it includes automaticity and metacognition as two key elements to predict listening success: the more automatized the decoding, lower-level, or bottom-up processes are, the more successful the listener. Furthermore, the more aware the listener is about their own cognitive processes, and how to monitor and regulate them, the more success they will have in their listening comprehension.

This listening model highlights the intrinsic importance of bottom-up processing as listening comprehension is prompted by an auditory signal that is perceived, then parsed and eventually utilized. If there are difficulties while noticing that signal (Schmidt, 1990), or if listeners struggle to perceive and parse it, the entire comprehension is affected. Although, top-down processing might facilitate the understanding and bridge those gaps, in some cases it is impossible, particularly with lower-level listeners, because their short-term memory is overwhelmed and the burden is “intolerable” (Nation, 2016; 5). Alternatively, more proficient language users are able to make an “orchestrated use of bottom-up and top-down sources of information” (Graham & Santos, 2015, 13).

Once we have shown the importance of listening in L2 learning (section 2.1) and how it is understood in the present study (section 2.2), Section 2.3 has introduced the reasons why vocabulary knowledge is the other variable under study in this investigation. In this respect, several studies have been cited to show the positive and facilitating relationship between vocabulary knowledge and listening comprehension. Among those references in the literature, the ‘noticing hypothesis’ (Schmidt, 1990) and the ‘cognitive load theory’ (Paas & Sweller, 2014) have been cited to support the inclusion of learners’ vocabulary size as a variable under study. If noticing the words in an aural text and how salient they are perceived (van Zeeland, 2014a) might affect the overall comprehension of aural messages from the very beginning of the process (Figure 2.1), learners’ vocabulary already stored in their memory should be one of the independent variables under study.

Section 2.4 has dealt with the issue of the unit of counting. This aspect is crucial in estimations of vocabulary size, particularly if we equate lexical difficulty of a text with the frequency of its words in the target language. Word families seem

to be the most suitable unit of counting for receptive vocabulary size. However, the section has also addressed those instances in texts that have no clear match in wordlists based on frequency: polysemy, homofoms, proper nouns, compound nouns, multiword verbs, and formulaic language.

The final thread in this literature review has focused on different ways to estimate the L2 receptive vocabulary size, and discussed issues with respect to word frequency in the analysis of texts. The present research study intends to assess the ability to recognize words and activate lexical matches in L2 learners by means of a receptive vocabulary test. Then, this estimation will be linked to each participant's ability to understand aural texts.

CHAPTER 3 – METHODOLOGY AND METHODS

The methodology followed in this investigation, and the methods employed within this framework is inherently related to the quality of the entire research study. This chapter will address the study validity and reliability, since none of the claims made in this research can be fully understood and eventually accepted by other researchers if they fail to be the result of a thorough and honest process of inquiry.

First, I will discuss how I see reality (ontology), and how I might apprehend it (epistemology). Then, those theoretical approaches to reality (ontology) and investigation (epistemology) will be operationalised in the form of the constructs used in this research study, the research questions to investigate those constructs, and the statistical analyses used to draw conclusions from the investigations. The second part of this chapter will deal with the methods employed in this investigation, and the decisions made in the planning, design, and implementation of its research instruments. Data from a preliminary study will also be analysed and presented to support the subsequent decisions made with respect to the research instruments employed in the main investigation.

3.1 – METHODOLOGY

The research process is sometimes compared to that of being a member of a given community, where you arrive at claims through the ‘disciplined’ process of using particular research methods. “[Y]ou follow its warrants; thus you belong to that club” (Heigham & Croker, 2009, 39). The ‘warrants’ to be followed in qualitative research tend to focus on meaning and sense making. They aim to prove that the study has captured fairly the essence of the phenomenon, and that the researcher’s findings make sense to the members of the particular research community concerned by the study. On the other hand, warrants in quantitative studies tend to be based more on numbers, and prefer to use standards drawn from statistics to establish the validity of their claims. In both cases, warrants are the core of the research enterprise, and help “distinguish research studies from conjecture or opinion because they make explicit the basis of belief for the claim” (Heigham & Croker, 2009, 38).

3.1.1 Ontological Assumptions and Epistemological Approach

In research studies using a quantitative approach, research tradition dictates a series of detailed procedures the investigators have to carry out. The stricter the procedures followed by the researchers are, the more confident they can feel of being right in the claims they make. In particular, quantitative researchers tend to include in their samples as many occurrences as necessary, but at the same time, they try to interfere as little as possible in the sampling process. Their sample will be representative if every possible occurrence of the phenomenon in reality has had an equal and fair chance to be included in it (section 3.2.3).

I agree with the attempts to overcome past paradigm wars or clear-cut

dichotomies between quantitative and qualitative approaches to the study of reality. In this respect, research objectives “can be classified as falling on a continuum from exploratory to confirmatory” (Onwuegbuzie & Leech, 2005, 277). In other words, instead of dividing research into two mutually excluding paradigms because of their essential nature (ontology) and the way they perceive reality (epistemology), we could classify research studies because of their ultimate purpose. I think this teleological way of approaching research is much more fruitful than adopting entrenched stances defending one paradigm over the other, because it focuses on a premise every researcher should bear in mind: research objectives drive studies, not the paradigm or method (Onwuegbuzie & Leech, 2005).

3.1.2 Ontology and Epistemology in the present Research Study

This research study primarily drew upon the claim that the vocabulary size and listening comprehension in a foreign language might be related (e.g., Fung & Macaro, 2019; Matthews, 2018; van Zeeland, 2014b). The methodology used here might be ascribed to the quantitative paradigm because of its clear intention of measuring and estimating a series of dimensions in second language learning. Adapting instruments that have been previously designed for similar research studies reinforces the ascription of the present investigation to the quantitative paradigm, and its research tradition.

First, this study used instruments to collect data that had been created within the same framework: a listening paper in a standardized test, and the vocabulary list published by the institution responsible for that examination. The vocabulary tests (Appendices 3, 4, 5 and 6) were based on a vocabulary list

compiled to help learners prepare for a language proficiency test (Appendix 7). Moreover, this vocabulary list was also meant for reference use for the people involved in writing question paper materials (Street & Ingham, 2007), so that they could “check whether it is permissible to test a word at a given level” (Capel, 2010, 2).

Secondly, the instruments employed to estimate the vocabulary size in most previous studies had only assessed their ability to recognize the link between the written form of a word and its meaning (section 2.5). However, many researchers in applied linguistics claim that learners’ ability to recognize words in their written and spoken forms might be different and “should be assessed separately” (van Zeeland, 2017, 144). A few studies have followed this advice and used separate tests to assess the aural vocabulary knowledge of L2 learners, although they might raise some concerns. Both dictation exercises (e.g., Cheng & Matthews, 2018; Fountain & Nation, 2000; Matthews, 2018), and aural versions of word-recognition tests (e.g., Milton & Hopkins, 2006; Milton et al., 2010; van Zeeland, 2014a) might show construct validity issues, as well as an overestimation of learners’ aural vocabulary size as big as 34.6% (van Zeeland, 2014a).

The present research study is a partial replication of the investigation carried out by McLean et al. (2015), as the vocabulary test also presents orally its items, both in an isolated manner and then, embedded in a sentence. Furthermore, it is a bilingual vocabulary test, where the items are presented in the target language – English – but the test-takers are given the four possible options to choose from in Spanish, their L1. On the other hand, the methods used in the present study differ in several aspects from the ones employed by McLean et al. (2015). Firstly, the items in the listening vocabulary test (LVT)

were also included in a written vocabulary test (WVT) to allow possible comparisons across study participants, and to increase the accuracy of the assessment by employing multiple measures (Webb, 2002). Moreover, the two versions of the vocabulary test were supplemented with a listening comprehension test (LCT) to enable possible correlation analyses between vocabulary and listening scores. Secondly, the possibility of experiencing either floor or ceiling effects that might affect the study reliability was avoided by selecting the target population through clearly stated inclusion criteria (Table 3.1). Furthermore, the reliability of the results was enhanced by drawing upon a vocabulary list – PET Vocabulary List – meant for the same linguistic level as in the inclusion criteria, and by using a listening paper from the same standardized examination.

3.1.2.1 Vocabulary and Listening: Homogeneity, Reliability and Generalizability

The use of a more homogenous sample of candidates is a relevant issue in this research study because it might have a positive impact on the reliability of the testing instruments designed to collect data from the participants. The preservation of homogeneity was applied to both samples used in this investigation: participants and target items for the tests. Furthermore, this homogeneity in the sample was intended to be ensured in the sampling of study participants by setting a series of inclusion criteria to enhance their representativeness of the target population (Table 3.1). The same criteria were preserved in the observational period of the study, with a view to increasing the reliability of the data gathered. Ultimately, the results and conclusions drawn from those data might show enhanced validity and reliability by having a more

homogenous sample than in previous studies (section 5.2).

Table 3.1 – Inclusion criteria for the target population.

(1) AGE: adult students
(2) L1: Spanish
(3) TEACHING: formal instruction at a language centre
(4) LINGUISTIC LEVEL: intermediate-level groups (B1-level according to the CEFR)

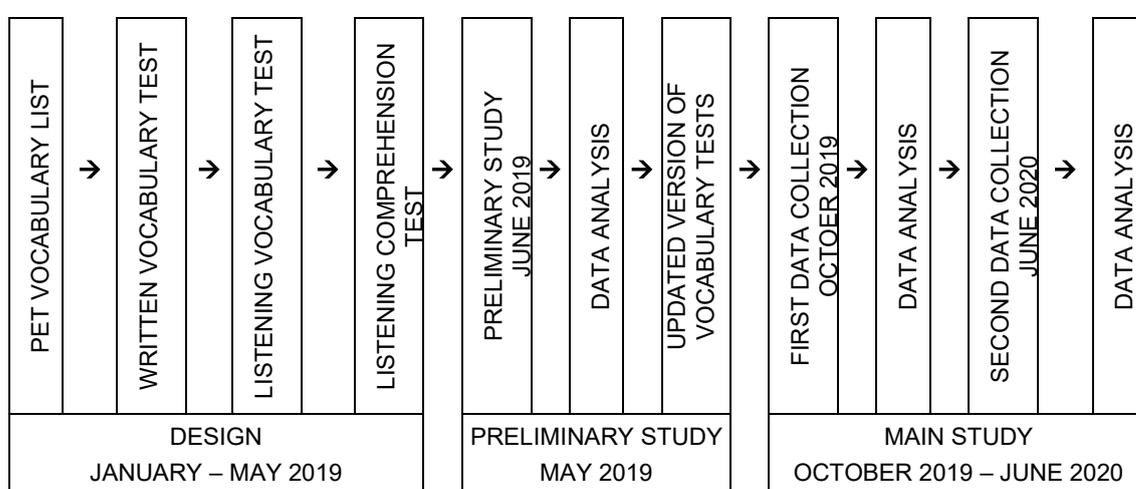
The closer and more similar the sample and the population are, the more confident we can be in attributing characteristics from the former to the latter. The statistics based on the sample used in this investigation were more likely to reflect the possible parameters because the sample was closer to the population under study. Reliability is understood here as internal consistency (Jones, 2013), where similar results will be gathered on repeated uses on the same subjects. Consequently, the more reliable the test scores are, the more generalizable the study is (Bachman, 1990).

A homogenous sample of target words was also sought for the two vocabulary tests employed in this study, as their target items were selected from a vocabulary list compiled for students with a similar level as the target population in this study (Street & Ingham, 2007). The two vocabulary tests were then delivered to a sample of students attending B1-level classes of English, and whose first language was Spanish. Their answers in the tests were analysed to determine the best performing items, so that a shorter and enhanced version was subsequently employed during the observational stage of the study.

Another major difference with previous research is the use of a longitudinal design to investigate the relationship between vocabulary and listening comprehension at two points in time. The use of the same research instruments on the same population on a second occasion was sought to enable the

corroboration of preliminary results and findings. Moreover, a standardized B1-level listening exam – based on the same framework as the vocabulary list – was used in the main study. Eventually, the participants’ scores in the vocabulary tests were matched to their results in the 25-item listening test. Once again, the research design intended to preserve high standards of both validity and reliability. Figure 3.1 shows a diagram with a timeline for the different instruments employed in thin this investigation.

Figure 3.1 – Research Instruments: Design, Implementation and Use



3.1.2.2 Vocabulary and Listening: Ecological Validity

In the research design for this study, I adopted a quasi-experimental approach to data gathering, where participants were recruited from intact classes (Dörnyei, 2007). This study was observational, and I did not intend to intervene upon, or attempt to control all the circumstances that affected the participants’ learning. In other words, the ecological validity – understood as minimal interference with the participants’ usual circumstances when learning English – took precedence in this study. Nevertheless, controlling all those factors was beyond its scope, and it would have been impossible, given its longitudinal design with intact groups of learners.

Ecological validity has been typically linked to the question of whether researchers are able to generalize from what they have observed in their laboratories to the world outside those premises (Schmuckler, 2001). Consequently, the more the laboratory conditions mimic the ones existing in reality, the more ecologically valid a study might be. Nevertheless, in social and behavioural studies this validity can only be “approximated” (Cicourel, 2007, 735), because researchers need to find compromises along their investigative journeys to make their studies operationally feasible, and to have adequate scientific control over their investigations (Schmuckler, 2001). In the present study, those research compromises might be found in the clear definition of constructs (section 3.1.2.4), the inclusion criteria for the prospective study participants (Table 3.1), and the unbiased selection of a sample as representative as possible.

Three decisions were made to keep the experimental context – especially the stimuli and the tasks – as close to real life as possible, and to approximate the ecological validity in this study (Schmuckler, 2001). First, the LVT intended to replicate what language learners might find in real-life situations: a speaker using a given lexical term embedded in a sentence that the listener has to decode in real time. Pauses were inserted between the items, so that the participants had time to read the four options from which to select the correct meaning. Translations into Spanish were used with the aim of operationalizing the idea of ‘understanding’ the meaning of the target word (section 3.1.2.4). Secondly, the target items in the vocabulary tests were selected from the PET vocabulary list published by Cambridge Assessment English (UCLES, 2012). B1-level students are referred to this compilation to prepare for that standardized language examination, including its listening paper. Thirdly, a

listening paper from the PET examination was used to assess the participants' listening comprehension. This is the same test, with the same tasks, rubrics, instructions and stimuli as thousands of B1-level students of English take every year to certify their linguistic proficiency. Moreover, this listening paper comes from the same framework as the vocabulary list employed in the random selection of the items for the vocabulary test.

Moreover, the possible variability in the data – derived from the preservation of the ecological validity of the study – was also limited by using clear inclusion criteria to have a more homogenous sample of participants than in similar research studies (McLean et al., 2015). The likely variability in the lessons and methodologies the participants were exposed to during the observation period (approximately 35 weeks) was also minimised by the fact that all of them were recruited from the same setting, a state language school in Spain (section 3.2.5.1). Although language learners from different groups participated in the main study, their teachers had to use similar materials in their classes, follow similar methodologies, and prepare them for the same end-of-course exam. Additionally, a great deal of consistency across groups was also expected because their teachers and materials were supposed to align themselves to the guidelines stated for intermediate-level language learning in the CEFR.

This section has presented the population under study, and the instruments to assess the relationship between vocabulary and listening. The first aim in the test design was to preserve the access to intact classes with learners of English as a foreign language, because the ecological validity of the study took precedence over the control for external variables that might impact on the results. At the same time, the homogeneity in the sample of participants was sought by including a series of criteria (Table 3.1). The selection and design of

the research tools to gather the data were also purposeful, as the target words for the vocabulary tests were selected from a list compiled according to the same criterion reference as the listening test. Having a homogenous sample of participants and research instruments might have helped minimise the impact of extraneous variables on the general construct of L2 vocabulary knowledge and listening comprehension. The next section in this chapter will discuss how the investigation of a topic as broad as L2 vocabulary and listening was framed by a series of research questions and the subsequent definition of the study constructs.

3.1.2.3 *Vocabulary and Listening: Research Questions*

The immediate objective of this study was to confirm the claim that the L2 vocabulary size and listening comprehension might be related (Fung & Macaro, 2019). The approach chosen for this study was a confirmatory one (Onwuegbuzie & Leech, 2005), in an attempt to find data to corroborate what other researchers had previously claimed. Therefore, a quantitative research design was employed to confirm or refute that claim. Furthermore, inferential statistics were used to analyse the data and gather evidence for the Research Questions in this study (Onwuegbuzie & Leech, 2005). A teleological approach to research might be really efficient when carrying out investigations because I do not think that research is “a philosophical exercise” but an attempt to find answers to questions (Dörnyei 2007, 207). Consequently, the decisions made within the entire research process were “highly dependent on the research question asked” (Mackey & Gass, 2015, 44). Table 3.2 shows the research questions that have guided the present study.

Table 3.2 – Research Questions

1. How much of the listening performance in an exam might be attributed to knowing the words in a vocabulary list?
2. How much lexical coverage of a spoken text does a learner need to achieve comprehension in a listening test?
3. How similar are the scores in vocabulary size tests based on recognising either the aural or the written form of words?
4. How does the relationship between lexical knowledge and listening performance evolve over time?

Research Question 1 aims to discover whether the vocabulary a person knows influences their ability to understand aural texts, and how strong that influence might be. The instruments used to assess this relationship are a listening vocabulary test (LVT) and a written vocabulary test (WVT), whereas the listening ability is examined with a listening comprehension test (LCT). The basic analysis to determine a relationship between two variables – vocabulary and listening – is the computation of Pearson product-moment correlations between the results in the vocabulary tests and in the LCT. A further way of exploring that relationship is by means of multiple regression analyses, where the variability in the dependent variable, i.e., results in the LCT, is explained with the help of the independent variables, i.e., the results in the LVT and the WVT.

Research Question 2 offers an additional perspective to RQ1 by looking at the lexical coverage necessary to achieve listening comprehension (section 2.5.1). The answers to this question might come from analysing the transcript of the LCT and determining how frequent its words are in English. The procedure is repeated with the words featured in the vocabulary list, where vocabulary tests were based. Then, results in the vocabulary tests are matched to the performance in the LCT, depending on the percentage of words in the transcript that also are featured in the vocabulary list. Further analyses might include

analysing the listening comprehension depending on the bands of lexical coverage from the results in either the LVT or the WVT. Additionally, the significance of the differences across those subdivisions, and their possible effect sizes might be examined with paired *t*-tests and Cohen's *d*.

Research Question 3 explores the possible differences that might exist between the ability to recognize words in their aural or in their written form. Descriptive statistics from both the LVT (aural form) and the WVT (written form) might help answer the question, by comparing the MIN, MAX, MEAN scores (raw data), and measures (Rasch analysis expressed in logits) from both tests. Again, the significance of the possible differences will be explored with the help of paired *t*-tests, and the size of their effects calculated with Cohen's *d*.

Research question 4 (RQ4) expands the scope of RQ1 by making the most of the longitudinal design of the present study. By comparing the answers to RQ1 obtained from two separate datasets (October 2019 vs June 2020), I intended to explore the evolution of the relationship between L2 vocabulary size and listening comprehension over time. An additional perspective on RQ4 refers to the knowledge and ability a language learner gains after a period of ± 35 weeks attending language classes. Pearson product-moment correlations and multiple regression analyses on the different datasets will be used to find possible differences. In turn, the significance of those differences and their effect sizes might also be determined. Figure 3.2 presents a summary of the instruments and analyses employed in this investigation.

Figure 3.2 – Instruments and Analyses to answer Research Questions.

RESEARCH QUESTION	RESEARCH INSTRUMENTS	ANALYSES
RQ1 – How much of the listening performance in an exam might be attributed to knowing the words in a vocabulary list?	LVT, WVT, LCT	Descriptive statistics.
	LVT, WVT, LCT	Pearson product-moment correlations LVT-LCT.
	LVT, WVT, LCT	Multiple regression LVT/WVT → LCT.
	LVT, WVT, LCT	Descriptive statistics top/bottom LCT scores.
	LVT, WVT, LCT	<i>t</i> -test for differences in measures and scores top/bottom in LVT, WVT and LCT.
	LVT, WVT, LCT	Correlations LVT/WVT with top/bottom LCT scores.
	LVT, WVT, LCT	Multiple regression LVT/WVT → top/bottom LCT scores.
RQ2 – How much lexical coverage of a spoken text does a learner need to achieve comprehension in a listening test?	PET listening transcript / PET vocabulary list / Compleat v.2	Comparison of words in transcript with words in vocabulary list (<i>Compleat</i> v.2, Cobb, 2019).
	LVT, WVT, LCT	Lexical coverage of words in LVT and WVT depending on results in LCT.
	LVT, WVT, LCT	Bands of lexical coverage based on either LVT or WVT vs results in LCT and LVT or WVT.
	LVT, WVT, LCT	Correlation and significance analyses of differences between LVT and LCT, and between WVT and LCT.
RQ3 – How similar are the scores in vocabulary size tests based on recognising either the aural or the written form of words?	LVT, WVT	Percentages of overall correct answers LVT vs WVT.
	LVT, WVT	Percentages items / persons with more correct answers in LVT or in WVT.
	LVT, WVT, LCT	MIN, MAX, and MEAN person measures for LVT, LCT and WVT in the three datasets.
	LVT, WVT	Paired <i>t</i> -tests for significance of differences in person mean measures.
	LVT, WVT	Cohen's effect size for significant differences in person mean measures.
	LVT, WVT, LCT	Paired <i>t</i> -tests for significance of differences in person mean measures, depending on scores in LCT (pass vs fail).
	LVT, WVT, LCT	Cohen's effect size for significant differences in person mean measures, depending on scores in LCT (pass vs fail).
RQ4 – How does the relationship between lexical knowledge and listening performance evolve over time?	LVT, WVT, LCT	Comparison of evolution of Pearson product-moment correlations LVT-LCT-LCT from three datasets (May'19, October'19, and June'20)
	LVT, WVT, LCT	Comparison of evolution of Pearson product-moment correlations for each test, from one dataset to the other (October'19, and June'20)
	LVT, WVT, LCT	Comparison of evolution of mean person measures: LVT vs WVT depending on scores in LCT (October'19, and June'20)
	LVT, WVT, LCT	Paired <i>t</i> -tests for significance of differences in person mean measures in LVT and in WVT (October'19 vs June'20), depending on scores in LCT (pass vs fail).
	LVT, WVT, LCT	Paired <i>t</i> -tests for significance of differences in scores in LVT and in WVT (October'19 vs June'20), depending on scores in LCT (pass vs fail).
	LVT, WVT, LCT	Cohen's effect size of significant differences in person mean measures in LVT and in WVT (October'19 vs June'20), depending on scores in LCT (pass vs fail).
	LVT, WVT, LCT	Cohen's effect size of significant differences in scores in LVT and in WVT (October'19 vs June'20), depending on scores in LCT (pass vs fail).

NOTE: LVT = Listening Vocabulary Test; WVT = Written Vocabulary Test; LCT = Listening Comprehension Test.

3.1.2.4 Vocabulary and Listening: Research Constructs

Once the research questions have been made explicit, a further step in statistical research studies consists of a clear statement about how the different constructs are understood. Instead of using a theoretical definition for them, quantitative research methodologies dictate that they have to be defined from an operational point of view (Hatch & Lazaraton, 1991; Purpura, Brown & Schoonen, 2015). These are the constructs used in this research study and their operational definitions:

Listening performance is understood as the scores obtained by participants in the listening paper of the Cambridge English: Preliminary (PET). The format of the original listening paper has been minimally adapted for the present research, and it has adopted the name of listening comprehension test (LCT).

Word is equivalent to each item featured in the PET vocabulary list. This definition considers that each part of speech or use made explicit in the list is a different 'word' (as it happens in 'work' – included as a verb and a noun –, or in 'play football', 'play the guitar', 'play the recording'). However, word is understood as 'word family' (Bauer & Nation, 1993) when references to other studies or wordlists based on frequency are made.

Knowing a word is synonymous here with *recognising a word*. Both concepts are identified as selecting the correct Spanish translation from four different options. The English target word is delivered either in its oral (LVT) or written (WVT) form, while the possible options to choose from are written in Spanish.

Lexical knowledge is the percentage of correct answers in the listening and written version of the vocabulary test.

Lexical coverage is defined as the percentage of words (understood as word

families) in a given text that are featured in wordlists based on frequency, or in other vocabulary lists.

Achieve comprehension in a listening test is understood here as reaching any of the cut-off scores for each of the bands (A, B and C) that Cambridge Assessment English considers successful performance in its PET listening test (UCLES, 2015). Those cut-off scores are 18/25 for the pass (grade C), 21/25 for the pass with merit (grade B), and 23/25 for the pass with distinction (grade A). Testing learners' listening performance through a series of discrete items and unidirectional audio files might cause a reduction in the overall validity of the language test (section 2.1.3), and it certainly correlates poorly with what research might understand under ecological validity (section 3.1.2.2). However, it is a compromise adopted in this study to operationalise the complex construct of listening success. Furthermore, this operationalisation of the construct enables comparisons with past research (e.g., McLean et al., 2015; Stæhr, 2009), and create thus a shared understanding with respect to listening success for the subsequent dissemination of the results in this study.

3.1.2.5 Vocabulary and Listening: Use of the Rasch Model

An important feature of this research study is the use of the Rasch Model for the data analysis, which implies accepting explicitly the interval nature of data, because counts “cannot replace measurement as it is known in the physical sciences” (Bond & Fox, 2015, 6). Once we have argued that the particular trait under investigation is amenable to quantification (sections 2.4 and 2.5), we need to construct a measure of it “so that the numbers indicating the variety of values of the trait *may* be subjected lawfully to the mathematical computations

that we routinely use in statistical analyses” (Bond & Fox, 2015, 297, emphasis in original).

The Rasch Model is intended to “design and revise a measurement instrument and carefully compute ‘measures’ that can be confidently used with parametric statistical tests” (Boone, Staver & Yale, 2013, 3). The model converts raw scores equivalent to counting – into linear and reproducible measurement. A unique characteristic of the Rasch model is the parameter separation, i.e., its ability to compare persons and items directly, which leads to the creation of “person-free measures and item-free calibrations, as we have come to expect in the physical sciences” (Bond & Fox, 2015, 349). By means of a probabilistic match, it conjointly analyses two factors that affect the performance in a test, the person’s ability, and the item difficulty. Georg Rasch – the Danish mathematician who first used this approach to data analysis – explained:

A person having a greater ability than another should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another one means that for any person the probability of solving the second item correctly is the greater one (Rasch, 1960; in Wright, 1997, 37).

Ability and difficulty are measured conjointly, and consequently, for quantitative analyses in the human sciences “Rasch measurement is the only game in town” (Bond & Fox, 2015, 317-318). It provides the researchers with parameters for both the participants in their investigation and the items used to quantify the variables under study, as well as the possibility of conjoint additivity (Brentari & Golia, 2007). In practical terms, the Rasch model offers the researcher a single unit of measurement called ‘logit’ which enables the comparison of items and persons on the same scale, as well as the comparison of different samples of people, or different items related to the same observed trait.

One logit is the distance along the line of the variable that increases the odds of observing the event specified in the measurement model by a factor of 2.718..., the value of "e", the base of 'natural' or Napierian logarithms used for the calculation of 'log-' odds. All logits are the same length with respect to this change in the odds of observing the indicative event (Linacre & Wright, 1989). In other words, the same way we use Fahrenheit units to compare temperatures observed at the same time, or on different moments – either in the same place, or in different locations – we can use logits to compare person abilities and item difficulties using a single unit of measurement.

Moreover, the Rasch model provides the researcher with two different measurements of reliability: one for the persons in the sample, and one for the items included in the instruments to collect the data. Those indices are “more conservative and less misleading [than Cronbach Alpha, which] overstates the reliability of the test-independent, generalizable measures the test is intended to imply” (Linacre, 1997, 581). The analyses are both conservative and reliable because the data collected in a study have to conform stochastically to the Rasch model before being able to be analysed, which makes it “preferable” to other ways to analyse data (McLean et al., 2015, 756).

Along with these two reliability figures, the Rasch model also shows the separation among items and persons in the data. In general, the bigger the separation in the items, the better they are performing in a test, as they cover all the parts along the continuum that the observed dimension might show. The same rule applies to person separation: the larger the separation, the more adequate the sample of participants is for the dimension we want to study.

A final issue that is particularly worth mentioning with respect to the Rasch model is that its reliability indices are driven primarily by N , so the performance

of 100 persons gives us more information about 30 items tested than the information that 30 items might provide about 100 persons (Bond & Fox, 2015). As both the item difficulty and the person ability are simultaneously estimated, the quantity and quality of the items will have a positive impact on the person reliability and separation estimates, and vice versa.

“If there is a single failing common to many scales, it is that the number of items is too small to support the decisions made by test users – because of the large SEs (standard errors) of the person estimates” (Bond & Fox, 2015, 94). It is precisely with respect to error measurement that Rasch analysis might be even more helpful than other theories and analyses for test data because the researcher knows how accurate the measurement has been (Bachman, 1990). Rasch analysis compares abilities (persons) and difficulties (items) in an iterative process, yielding the standard errors of means for both the persons and the items, as well as the standard error of measurement with respect to the very same persons and items. In other words, anyone reading a study report featuring those statistics is able to know exactly the amount of error that is not random, but attributable to the decisions made by the researcher while designing or implementing their study. Consequently, the trustworthiness of the claims made with respect to an investigation becomes apparent the moment its main researcher reports the precision of the research instruments through those standard errors. The more accurate the instruments used to gather data are, the fewer measurement errors, and the higher the research reliability is. Furthermore, this study also intended to preserve its reliability in the process of data gathering, which, in turn, helped support the claims about the phenomena under study. The data collection stage was actually considered an ongoing process where modifications to the original plan were added, with a view to

enhancing the study reliability. Once the overall plan for the research study and its stages was outlined (Figure 3.1), a preliminary study was implemented to decide on the accuracy of the two vocabulary tests, and to gather information about the feasibility of subsequent data gatherings. The decisions about which items from the preliminary study should be kept for successive observations were based on the possible increase in the estimates of item separation and reliability, without dramatically reducing the figures in the person separation and reliability. As the sample was the same, its impact on the item separation and reliability would increase if some poorly performing items were excluded from the analysis. However, the removal of too many items from the analysis might affect negatively the estimates for person separation and reliability, as reliability is mainly driven by N (section 3.2.5.2).

The first part of this chapter has dealt with general issues related to my positionality as a researcher with respect to reality (ontology), and the way I think we can apprehend it (epistemology). Then, a more focused discussion has been introduced by relating that positionality to actual decisions made with respect to the research topic, the population under study, its ecological validity, the research questions employed to investigate that research topic, the constructs used in those research questions, and the overall approach to analysing the data collected in the study. Now it is time to present the actual tools adopted and implemented to gather valid and reliable evidence.

3.2 – METHODS

This section will present the different research instruments employed in this study. Previous attempts to assess the relationship between L2 vocabulary size and listening comprehension might have presented flawed methodologies affecting their validity and reliability (section 2.5). Therefore, the methods employed here intended to differ from the ones employed in those studies.

All the tools used in the data collection for the present study – the listening vocabulary test, the written vocabulary test, and the listening comprehension test – are based on the examination Cambridge English: Preliminary. Three reasons account for this decision. Firstly, because Cambridge Assessment English has provided sound evidence of the criterion-related validity in its listenings (Lim & Khalifa, 2013). Secondly, because using a cohesive framework for the vocabulary and listening tests might enhance the internal reliability of the study results, although it implies accepting the operationalisation of the construct of listening comprehension as Cambridge Assessment English understands it (3.1.2.4). Thirdly, and most importantly, I decided to use this standardized language examination because it belongs to a series of exams that have become extremely popular among Spanish students learning English. Every year, more and more primary and secondary schools, as well as higher education institutions, both state and private ones, decide to externally evaluate their students' performance in English through the different options the Cambridge Assessment English offers. Consequently, the possible findings within this research framework might be particularly meaningful to a significant percentage of the target population in this study.

This section will address the adaptation, design and implementation of the research instruments. A second subsection will focus on a preliminary study

carried out to determine the best performing items in the vocabulary tests, as well as their overall validity and reliability. Lastly, the data gathered in the main study both in October 2019 and in May 2020 will be presented, with a particular focus on the reliability of the datasets.

3.2.1 Vocabulary Test – Preliminary Issues

First of all, the unit of counting for the vocabulary test was the types that appear in the PET wordlist, because its compilers failed to specify what level of grouping they intended for the words in that list, either types, lemmas or word families (section 2.4). Consequently, each entry in the PET vocabulary list – including each of the specified meanings in polysemic words like ‘play’ – is considered an independent item for its inclusion in the vocabulary test. This decision enabled the presence of items like ‘improve’ (verb, item 42) and ‘improvement’ (noun, item 65); ‘colour’ (noun, item 77; verb, item 84), in the first version of the tests.

Secondly, multiword expressions (e.g., ‘at least’, ‘bank account’, ‘find out’) were excluded from the analysis of frequency and from the design of the vocabulary test because *Compleat* (Cobb, 2019) – the software employed for text profiling does not include lists of multiword expressions compiled by other authors (section 2.4.4). Therefore, only the BNC-COCA 1-25k wordlists compiled by Nation (2012, 2019) were used in the text profiling analyses.

On the other hand, although each entry in the PET Vocabulary List was considered a potential item for the vocabulary test – including those made for different meanings in polysemic words – they were all subsequently grouped into word families to enable comparisons with other frequency lists already used

in the literature, enhancing thus the generalizability of the possible claims drawn from the present study. The BNC-COCA 1-25k implies an update because it is the most recent and complete list of words. These wordlists might be more balanced as they are based on two different and more encompassing corpora: the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), which is considered “the best corpus of general English in existence” because of its size, balance and currency (Schmitt & Schmitt, 2012, 494). Furthermore, it presents a balance between oral and written corpora closer to real-life situations (Nation, 2016). One possible negative consequence of using the BNC-COCA 1-25k wordlists is that results might be different from those obtained with the use of other lists like the BNC 1-20k (Leech, Rayson & Wilson, 2001). Nevertheless, Nation (2016, 141) claims to be “not too worried by this criticism because it is the quality of the resulting lists that matters and the adjustments have been made to improve their quality”. In any case, it seems plausible to imagine that future research on vocabulary learning and text coverage might draw on the BNC-COCA 1-25k wordlists rather than on previously compiled lists.

3.2.2 Cambridge English: Preliminary and Preliminary for Schools (PET) – Vocabulary List

The main purpose of this vocabulary list is “to give teachers a guide to the vocabulary needed when preparing students for the Preliminary and Preliminary for Schools examinations [and] to guide item writers who produce materials for [this] examination” (UCLES, 2012, 2). The PET list was used *as it is* to provide items for the vocabulary tests in the present study. Therefore, all entries in the original list, except those previously excluded because of feasibility reasons

(section 3.2.1), had an equal chance to be included in the test. The following sections will account for the necessary adjustments made on the items in the original compilation.

3.2.2.1 *Preparing the PET Vocabulary List*

The official PET Vocabulary List had 2,978 separate entries, i.e., not lemmatized or grouped into word families according to the affixation or derivation they might show. Then, it was edited to include as many entries for a word as parts of speech or word meanings had been compiled by its authors, like ‘play the guitar’ and ‘theatre play’. Moreover, items in the appendix to the PET list like the days of the week or the months of the year were added. The rest of the word sets were disregarded, either because of the difficulty of setting a limit of members to be included (ordinal and cardinal numbers), or because they are all considered proper nouns or their derivations (countries, continents, nationalities and languages). The final version of the PET list had 3,510 entries.

3.2.2.2 *The PET Vocabulary List and the BNC-COCA 1-25k*

The edited version of the vocabulary list had then 3,213 entries featuring a single word, including compound nouns like *windscreen* or *sunglasses*, and 297 entries featured multiword expressions like ‘air conditioning’, ‘carry out’ or ‘at all’. The multiword expressions were excluded from the analysis made through the online program *Compleat Web VP v.2* (section 3.2.1). Out of the 3,213 entries from the edited PET vocabulary list that had one word each, 3,089 found a match in Nation’s compilation of the 25,000 most frequent words in English (2012, 2019). Those matches made up a total of 2,168 word families, with a

ratio of 1.48 tokens per family (Table 3.3). The 124 tokens (totalling 114 types) from the PET list considered 'off-list' words by the software were all compounds like 'birthday' or 'bedroom', except for the word 'turkey' and the interjections 'oh' and 'wow'.

Table 3.3 shows a summary of the correspondences of the PET vocabulary list in the corpus formed by the 25,000 most frequent word families in English divided into bands of 1,000 words each. The first column shows the frequency band in the BNC-COCA corpora, the other columns refer to the number of word families, types or tokens in the PET vocabulary list in each of the bands in the 1-25k frequency list, with their corresponding percentages between brackets. The last column in the table features the cumulative percentage of tokens in the PET vocabulary test for each of the 1-25k bands. This figure might be the key percentage because it shows the lexical coverage of a given text that knowing the words up to that band might provide. For example, if a person knows the 3,000 most frequent words in English – according to the frequency lists based on the BNC-COCA corpora – they might be able to understand 84.6% of all words from the PET Vocabulary List. This percentage refers to the total number of tokens in that list with a match in the first three bands of the compilation, i.e., the 3,000 most frequent word families in English. This figure might be considered normal as that list includes highly frequent vocabulary “appropriate to the B1 level on the Common European Framework of Reference” (UCLES, 2012).

Table 3.3 – Items in PET Vocabulary List according to frequency bands in 1-25k BNC-COCA

Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumulative token %
K-1 WORDS	961 (44.30)	1278 (46.99)	1660 (51.68)	51.68
K-2 WORDS	634 (29.20)	726 (26.69)	797 (24.80)	76.48
K-3 WORDS	226 (10.40)	244 (8.97)	261 (8.12)	84.60
K-4 WORDS	157 (7.20)	163 (5.99)	171 (5.32)	89.92
K-5 WORDS	89 (4.10)	92 (3.38)	97 (3.02)	92.94
K-6 WORDS	48 (2.20)	48 (1.76)	49 (1.52)	94.46
K-7 WORDS	23 (1.10)	23 (0.85)	23 (0.72)	95.18
K-8 WORDS	15 (0.70)	15 (0.55)	15 (0.47)	95.64
K-9 WORDS	6 (0.30)	6 (0.22)	6 (0.19)	95.83
K-10 WORDS	2 (0.10)	2 (0.07)	2 (0.06)	95.89
K-11 WORDS	3 (0.10)	3 (0.11)	3 (0.09)	95.99
K-12 WORDS	1 (0.00)	1 (0.04)	1 (0.03)	96.02
K-13 WORDS				96.02
K-14 WORDS	3 (0.10)	4 (0.15)	4 (0.12)	96.14
OFF-LIST		114 (4.19)	124 (3.86)	100.00
Total (unrounded)	2168	2719 (100)	3213 (100)	100.00

Table 3.4 displays the results for the 222 tokens resulting from dividing the 111 one-word compound nouns in the PET Vocabulary List into their two components. Table 3.5 shows the results from the analysis of the PET Vocabulary List after including two separate elements of compound nouns. The total number of word families and types is exactly the same, despite the inclusion of 222 tokens. This implies that all of the new items incorporated to the analysis of the PET Vocabulary List were already in the previous version of the compilation. Furthermore, 207 out of the 222 tokens added to the list (93.24%) correspond to word families in the 3,000 most frequent words (bands 1-3k), which supports the argument that compound nouns are usually the result of combining very frequent words into a new instance (Nation, 2016).

Table 3.4 – Compound nouns in PET Vocabulary List according to frequency bands in 1-25k BNC-COCA

Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumulative token %
K-1 WORDS	107 (73.79)	110 (73.83)	174 (78.38)	78.38
K-2 WORDS	22 (15.17)	22 (14.77)	27 (12.16)	90.54
K-3 WORDS	4 (2.76)	4 (2.68)	6 (2.70)	93.24
K-4 WORDS	5 (3.45)	5 (3.36)	6 (2.70)	95.95
K-5 WORDS	5 (3.45)	6 (4.03)	7 (3.15)	99.10
K-6 WORDS				
K-7 WORDS	2 (1.38)	2 (1.34)	2 (0.90)	100.00
OFF-LIST	??	0 (0.00)	0 (0.00)	
Total (unrounded)	145	149 (100)	222 (100)	100.00

Table 3.5 – PET Vocabulary List including compound nouns according to frequency bands in 1-25k BNC-COCA

Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumulative token %
K-1 Words :	961 (44.30)	1278 (49.08)	1835 (55.42)	55.42%
K-2 Words :	634 (29.20)	726 (27.88)	824 (24.89)	80.31%
K-3 Words :	226 (10.40)	244 (9.37)	267 (8.06)	88.37%
K-4 Words :	157 (7.20)	163 (6.26)	177 (5.35)	93.72%
K-5 Words :	89 (4.10)	92 (3.53)	104 (3.14)	96.86%
K-6 Words :	48 (2.20)	48 (1.84)	49 (1.48)	98.34%
K-7 Words :	23 (1.10)	23 (0.88)	25 (0.76)	99.09%
K-8 Words :	15 (0.70)	15 (0.58)	15 (0.45)	99.55%
K-9 Words :	6 (0.30)	6 (0.23)	6 (0.18)	99.73%
K-10 Words :	2 (0.10)	2 (0.08)	2 (0.06)	99.79%
K-11 Words :	3 (0.10)	3 (0.12)	3 (0.09)	99.88%
K-12 Words :	1 (0.00)	1 (0.04)	1 (0.03)	99.91%
K-14 Words :				
Off-List:	3 (0.10)	3 (0.12)	3 (0.09)	100.00%
Total (unrounded)	2168	2604 (100)	3311 (100)	100.00

The compound nouns in the PET Vocabulary List might be considered transparent enough to be understood as straight combinations of their two elements like *air-port*, *bath-room*, *book-shop* (Nation, 2016; Nation & Webb, 2011). With the addition of those compound nouns, the lexical coverage rises from 84.6% to 88.37%. Nevertheless, a language user who knows the 3,000 most frequent words in English might fail to recognize about 1 in every 8 words

in a text that has been created only with elements from the PET Vocabulary List.

3.2.3 Creation of a Vocabulary Test based on the PET Vocabulary List

Recommendations found in the literature were followed in the selection of the items from the PET vocabulary list, the creation of multiple-choice options for each of them, the careful writing of non-contextualising sentences, and the recording of the items to be included in the aural version of the vocabulary test (Beglar & Nation, 2007; McLean et al., 2015; Nation, 2016).

The frequency band of each word with a match in the 1-25k compilation (Table 3.5) was recorded onto the spreadsheet. Then, I selected and wrote down the most suitable and frequent translation into Spanish for each term by drawing upon my experience in teaching languages to L1-Spanish students. This process was carried out before the random selection of the items for the test, in order to avoid possible bias. Once the translation for the items in the list was available, along with the information about the frequency bands to which each item belonged, 150 words from the list were randomly selected to be included in the vocabulary test. The possible options for the multiple-choice answers in the test were taken from the same frequency band as the target item, as well as from the same part of speech. Unlike the target items in the test, the actual selection of the four options for each of those vocabulary items was not random, but based on my intuitions and experience in language teaching and vocabulary testing. The 600 options featured in this version of the vocabulary test (150 target items * 4 options = 600 options) were used only once. The vocabulary test consisted of a series of target words that the participants in the study had to

match to their best translation into Spanish among four options. This bilingual multiple-choice format was used because its benefits outweigh its limitations (section 2.5.2.2).

Each item was presented both in an isolated manner and within a minimal sentence that only helped the test-taker determine which part of speech was tested in each case. This manner of presenting the items to the participants intended to show them where to focus their attention, in an attempt to minimise the problems derived from a possible lack of noticing (van Zeeland, 2014a). Special care was taken in the selection of the three distractors (i.e., the three incorrect options), to avoid ambiguity and confusion, and in the writing of contextualising sentences for the target word, so that no additional information about its meaning was revealed.

Nevertheless, four measures were adopted to ensure the validity and reliability of the vocabulary test with respect to the contextualising sentences and the ambiguity of the options. Firstly, a version of the vocabulary test was distributed to five native English teachers with a very high level of Spanish and extensive experience in language teaching. This version of the vocabulary test had all the target items in the test substituted by the same string of characters (XXXXXX). The English teachers were asked to write down the part of speech they thought each of the items presented in the test, as well as to try to find out the correct answer. If everyone was able to determine the part of speech, the contextualising sentence was clear enough. On the other hand, if anyone was able to select the correct answer, the contextualising sentence had to be rewritten since it was revealing too much. None of the five teachers was able to select one option in any of the questions. There were only 10 discrepancies when selecting the part of speech tested in each of the 150 items, which means

that in 98.66% of the cases, the teachers coincided in their judgements (5 raters * 150 items = 750 cases). The percentage of coincidence among raters on the part of speech for each vocabulary item was in the range 80-100%, i.e., at least four teachers selected the same part of speech to be assigned to each item. The high level of agreement among those raters (98.66% of all assessed items) showed that the context sentences and the options had been designed correctly, and that the discrepancies were due to carelessness caused by such a repetitive and taxing task.

Secondly, the written version of the vocabulary test was distributed to six English teachers whose L1 was Spanish, and who had extensive experience in teaching English to Spanish speakers. The initial hypothesis was that these external raters should be able to select the correct answer for each of the target items in the test without hesitation. Furthermore, the raters were also asked to provide feedback on the clarity of the test with respect to selecting one answer over the other distracters. 99.44% of the answers from these experienced English teachers (895 answers out of 900) were correct, and none of the teachers missed the same target words, so we can conclude that their mistakes were due to carelessness or lack of attention. Furthermore, none of the teachers raised concerns about the possible ambiguity of any of the options, or the incorrectness of any of the translations. These results clearly confirmed both the validity of the Spanish translations for each item in the test, and the overall unambiguity of the options to select the correct answer for each question.

Thirdly, a recording was made to create the listening version of the written vocabulary test (Appendices 3 and 4). A recording studio was booked and a native English speaker of Irish origin was asked to read out each of the 150 words and their context sentences in the test, as well as the introductory

instructions and examples. The only indication he received was to read the text as clearly and naturally as possible, without attempting to conceal his idiolinguistic features (i.e., accent, prosody, intonation, etc.). Once the recording session finished, the raw audio file was edited with the software Audacity®, and the questions in the test were separated 5 seconds from each other. This length was deemed sufficient for the test taker to read the four options and select the correct one in each case (van Zeeland, 2014a).

Finally, a rubric with instructions and clear examples to guide the test-takers was included. Once both versions of the test were validated, they were delivered to a group of language learners recruited with the same inclusion criteria as the main study (Table 3.1). The Rasch model was subsequently used to determine which of the 150 items were actually most effective in estimating L2 English vocabulary size (section 3.2.5).

3.2.4 Cambridge English: Preliminary – Adapting the Listening Paper

A copy of a past PET listening paper (Table 3.6) was downloaded from the official site of Cambridge Assessment English. The documents were slightly edited to reduce the number of pages in the questionnaires, while preserving their readability and the clarity of their rubrics. The audio files accompanying that listening paper were also downloaded and minimally edited with the software Audacity® to remove the parts that were not relevant to the present investigation. The total running time of the audio files was about 25 minutes.

Table 3.6 – Summary of the PET listening paper (Cambridge University Press, 2008)

PART	TASK TYPE AND FORMAT	TASK FOCUS	NUMBER OF QUESTIONS
1	Multiple choice (discrete) Short neutral or informal monologues or dialogues. Seven discrete three-option multiple-choice items with visuals, plus one example.	Listening to identify key information from short exchanges.	7
2	Multiple choice. Longer monologue or interview (with one main speaker). Six three-option multiple-choice items.	Listening to identify specific information and detailed meaning.	6
3	Gap-fill. Longer monologue. Six gaps to fill in. Candidates need to write one or more words in each space.	Listening to identify, understand and interpret information.	6
4	True/False Longer informal dialogue. Candidates need to decide whether six statements are correct or incorrect.	Listening for detailed meaning, and to identify the attitudes and opinions of speakers.	6

Several aspects in the way this standardized listening paper is delivered are meant to reduce its difficulty, although they might lower its ecological validity (section 3.1.2.2). Firstly, the audio input for each part was played twice. Secondly, the questions in parts 2, 3 and 4 are presented in the same order as their corresponding answers appeared in the recordings, and with enough distance between bits of relevant information, so that test-takers can process the input and answer the corresponding question. Finally, test-takers are given a few seconds to look at the questions in tasks 2, 3 and 4 before the auditory input is delivered.

The LCT in the present study employed the same rubrics, questions, and auditory input as the PET listening paper. Similarly the marking of the different sections followed the criteria Cambridge Assessment does (UCLES, 2019). For parts 1, 2 and 4 only one of the options received 1 mark, whereas the other choices were awarded 0 marks. For part 3, only completely correct answers

received full marks, so spelling mistakes in otherwise correct answers (for example '*elefants*'*) will mean losing all the marks for that answer. The use of a marking scheme where there is only one *correct* answer and no half-marks are allowed implies very reliable results from a psychometric point of view, higher levels of interrater agreement and a reduced allocation of resources for the marking process (Bramley, 2008). Marking is then reduced to a simple process of matching the examinees' responses to the very few possible answers that are considered the right ones, and consequently bear full marks. Furthermore, the whole marking process can be automatized, as there is no margin for interpretation, with the subsequent reduction in time before the examinees receive their scores.

However, employing such constrained items in tests might create "validity problems (e.g., construct underrepresentation), since some competences might require more complex assessments" (Lind Pantzare, 2015, 2). In the case of the PET listening paper, Cambridge Assessment might equate listening comprehension with an all-or-nothing situation where no room is left for partial understanding. Furthermore, construct validity might be particularly at risk in part 3 of the PET listening paper. Examinees are expected to understand and interpret information in aural texts (Table 3.6) but they can only show it by writing correctly spelled short answers, with little margin for interpretation. The threat to the validity of the test is clear as examiners may not be eliciting the behaviour that they intended to evaluate (Ahmed & Pollit, 2011).

Nevertheless, the benefits for the research study quality of using the PET listening paper outweigh its validity issues. In particular, the PET listening paper was employed to assess the participants' listening ability because this examination is meant to assess language proficiency among the target

population of this study (B1-level learners of English), and because it has shown evidence of criterion-related validity (Lim & Khalifa, 2013). Secondly, this paper was used because the tests employed to assess the participants' vocabulary size were created from the PET Vocabulary List. Both aspects of this research project (vocabulary and listening) were examined within the same framework, so that the validity and reliability of the study results could be enhanced. Thirdly, the face validity of this study might increase by using the PET listening paper as thousands of language learners aim to certify their proficiency in English by taking the Cambridge English: Preliminary.

3.2.5 Preliminary Study – Refining the Vocabulary Tests

This section will describe the process of evaluating the research instruments designed specifically for the data collection in the main study. It provides a detailed account of the data collection and analysis in the preliminary study, as well as discusses the validity and reliability of the preliminary version of the vocabulary tests. The main aim of this preliminary study was to determine the best performing items in the test with respect to their ability to discriminate the participants' vocabulary knowledge. Furthermore, it examined the overall validity and reliability of the vocabulary tests, and provided valuable experience about dealing with the target population and analysing their data.

3.2.5.1 *Data collection*

The state language school in Pamplona (Spain) was the initial setting for the data collection in the preliminary study. State language schools – also known as official language schools – are language centres where residents in Spain can

learn foreign languages like English, French or Russian at affordable prices, as they are subsidized by public educational authorities. The different courses and languages offered by these schools are independent from what students in primary, secondary and tertiary education find in their officially-approved curricula. There are almost 300 centres all over the country with about 400,000 students (Ministerio de Educación y Formación Profesional, 2020). Since the first school was inaugurated in Madrid in 1911, these state language schools have been the only way to certify the language learners' competence in Spain, apart from their leaving certificate in secondary education (Maza, 2020).

Most of their students, including the participants in the present investigation, attend general language courses, usually held from October to May, where they receive input about and practice the different aspects of the target language like reading, speaking or listening. In general, these courses consist of 4-5 hours of classes a week, i.e., about 130 contact hours a year. The progress and learning of all students is assessed according to the same criteria and evaluation instruments in all schools in Spain. In the case of English courses, students in A1, A2 or B1 groups are expected to show enough proficiency at the end of their academic years to begin in the next level the following October. On the other hand, the curricula for higher levels of language proficiency in English – B2, C1 and C2 – is developed within two academic years. In other words, English-language students are expected move from A1 to A2, from A2 to B1, or from B1 to B2 after 130 hours of instruction in the classroom. In order to progress from B2 to C1, and from C1 to C2 they might need to attend 260 hours of classes.

All students from the state language school in Pamplona (Spain) enrolled in the 18 different general English groups at level B1 were invited to participate in the

preliminary study. Permission was granted by the school and its teachers (Appendix 16), but the students were about to finish their classes in the academic year, and most of them were concerned about passing the end-of-course exam. As both the students and their teachers could be reluctant to spend valuable class time to participate in the study, an online version of the test was created on Google Forms®. Following standard ethics procedures when research is carried out on human participants, the first section in the form presented the basic and relevant information for the study participants and asked them to tick a box to show their agreement to voluntarily take part in the preliminary study (Appendix 22).

The second section presented the listening vocabulary tests with 150 items (Appendix 23). Participants were told to answer all the questions, including those that were totally unknown to them. This decision was based on what language users might find in real-life situations, where they have to make tentative guesses at the meaning of unknown words. Once all questions in LVT were answered, participants were led to the written version of the same test (WVT).

The same target words were employed in both the LVT and the WVT to enable comparisons and determine possible differences between aural and written vocabulary size (RQ3, Table 3.2). The order to deliver the items was thus clear: first, the items in their oral form, and then the same target words, but in their written form. Participants were unaware of this repeated testing of the same items when they began the listening vocabulary test, but once they were in the WVT, they were asked not to change any of their answers in the previous section.

The design of the present study made it impossible to set controls for practice-

of-order effects, and all participants answered all the items in the same order. However, a subsequent analysis suggested that neither the first items in the test were more difficult to answer correctly because the participants have no experience with the test format; nor were the last items in the test more difficult because the test-takers were tired. Table 3.7 shows the scores obtained by the participants ($N = 73$) in the WVT and the LVT (150 items), divided in thirds with 50 items each. The order of delivery of the items was: LVT THIRD 1

→ LVT THIRD 2 → LVT THIRD 3 → WVT THIRD 1 → WVT THIRD 2 → WVT THIRD 3. No order effects can be seen in the data as the mean scores for each of the thirds is different from the order of delivery: WVT THIRD 1 > WVT THIRD 3 > WVT THIRD 2 > LVT THIRD 1 > LVT THIRD 3 > LVT THIRD 2.

Table 3.7 – Results in consecutive thirds of items in vocabulary tests (results based on raw data)

	LISTENING VOCABULARY TEST				WRITTEN VOCABULARY TEST			
	TOTAL 150 items	THIRD 1 50 items	THIRD 2 50 items	THIRD 3 50 items	TOTAL 150 items	THIRD 1 50 items	THIRD 2 50 items	THIRD 3 50 items
Maximum correct answers (73 participants)	10,950	3650	3650	3650	10,950	3650	3650	3650
Total correct answers (73 participants)	9,569	3218	3168	3183	10,133	3461	3309	3363
Mean score	131.08	44.08	43.40	43.60	138.81	47.41	45.33	46.07
Standard Deviation	9.10	4.07	2.84	3.76	7.22	2.40	2.46	3.13
% Correct answers	87.39 %	88.16%	86.79%	87.21%	92.54%	94.82%	90.66%	92.14%

This absence of differences between the results in the first and the last answers in the tests implies that the lack of experience to answer the first items in a test, or the fatigue and boredom caused by such a demanding and repetitive task had no impact on the participants' success to select the right answer. The mean scores for the three thirds in each test show that they depend both on the difficulty of the items and on the test modality, not on the order or sequence of the items: the second third is more difficult than the other two thirds, and the

LVT is more difficult than the WVT.

One could argue that the higher mean score in the WVT was due to the familiarity of the test-takers with the target items, as they were the same as in the previously delivered LVT. We might dismiss the possible familiarity with the items in the WVT to account for the differences in the performance because the test-takers were provided with no feedback on their answers in the LVT, so they were unable to experience any possible washback effect. The differences in the mean scores might indicate a relative higher difficulty in the test format, as the items in the LVT were delivered orally, whereas in the WVT the same items appeared in their written form.

Once the online test was ready, I visited all the B1-groups to invite their students to participate in the study. 170 students at the language school agreed to take part, but only 35% of the previously approached students ($N = 60$) eventually took the test. Consequently, I decided to approach a second group of B1-level students at a different setting. The homogeneity of the sample was preserved because all these additional participants met the inclusion criteria (Table 3.1). Furthermore, these participants were in the final weeks of their English B1-level course, exactly as the participants recruited at the other setting. All in all, answers from 73 participants were gathered and then analysed to determine the best performing items in the test. Those items would thus be the basis for the final and shortened version of the vocabulary tests to be employed in the main study.

3.2.5.2 Data Analysis: Descriptive Statistics, Reliability, Separation

The data were imported onto the program Winsteps® (Linacre, 2012, 2019) to

be analysed. Given the special importance that the reliability of performance has in applied linguistics (Hatch & Lazaraton, 1991), the first steps in the data analysis focused on this issue. It showed a higher person and item reliability for the 150 items in the listening test than in the written test. Similarly, the LVT showed higher person and item separation indices than the WVT (Table 3.8).

Table 3.8 – Reliability and Separation depending on the number of items in the test (expressed in logits).

	Person Separation		Person Reliability		Item Separation		Item Reliability	
	LVT	WVT	LVT	WVT	LVT	WVT	LVT	WVT
150 ITEMS	2.46	1.93	0.86	0.79	1.64	1.09	0.73	0.54
121 ITEMS	2.46	1.93	0.86	0.79	2.21	1.36	0.83	0.65
105 ITEMS	2.44	1.94	0.86	0.79	2.59	1.55	0.87	0.71
97 ITEMS	2.41	1.94	0.85	0.79	2.66	1.70	0.88	0.74
91 ITEMS	2.40	1.94	0.85	0.79	2.77	1.75	0.88	0.75
81 ITEMS	2.34	1.94	0.85	0.79	2.87	2.12	0.89	0.82

The person reliability index reported by the Rasch model is similar to more traditional ones in test theory like KR-20 or Cronbach’s alpha (Linacre, 2012). The closer the values are to 1, the more internally consistent is the measure. However, those traditional reliability indices are considered to overstate “the reliability of the test-independent, generalizable measures the test is intended to imply. For inference beyond the test, Rasch reliability is more conservative and less misleading” (Linacre, 1997, 581), because it avoids misinterpreting raw scores as linear measures.

The main reason for the differences in the reliability and separation indices might lie in the lower number of persons or items with extreme scores in the LVT with respect to the WVT. No test-taker got a perfect score in the LVT, but in 33 items all participants ($N = 73$) chose the correct answer. In the WVT, two people answered all 150 questions correctly, and for 60 items (40%) all test-takers selected the right option. Consequently, the standard error of

measurement (SEM) was higher in the WVT than in the LVT.

A data quality analysis was undertaken to increase reliability and separation in the tests by eliminating items conveying too little information about the participants' performance, i.e., those with perfect or nearly perfect scores. At the same time, the study design – aiming at comparisons between the participants' aural and written vocabulary size – forced the exclusion of items in both tests only when there were minimal differences in scores for a particular item from one test to the other.

Table 3.9 shows the number of items with perfect scores depending on the number of items included, and their corresponding separation and reliability indices.. Among the 150 items from the original dataset, 29 presented perfect scores in the LVT and 43 in the WVT. When items with either perfect scores in both tests or 72/73 correct answers in one test and perfect scores in the other were excluded from the analysis, a total of 121 items remained in both the LVT and the WVT. The item reliability and separation indices increased because only 4 items in the LVT and 31 in the WVT still presented perfect scores, whereas the person reliability and separation varied minimally. The same process of data quality analysis continued with the exclusion of items with perfect scores in one of the tests and 71/73 correct answers in the other, leaving a total of 105 items in each of the tests. As the item and reliability indices improved while the person measures remained equally high, the next steps were to exclude the items with perfect scores in one test and 70/73 correct answers in the other (97 items per test) and then those with 69/73 right answers in either the LVT or the WVT and 73/73 in the other. Eventually, the process stopped when items perfect scores in one test and with a minimum of 68/73 correct answers in the other were discarded. None of the remaining 81

items in the LVT presented perfect scores, whereas all participants answered correctly only item 88 in the WVT, which was not eliminated as its counterpart in the LVT was correctly answered by only 64 of the participants. The selection of the best performing items stopped then, when the maximum difference in scores between the LVT and the WVT was smaller than 7%, so that no significant differences between the two versions of the test were missing, and comparisons could be made (RQ3, Table 3.2).

Table 3.9 – % of correct answers in the test depending on number of items considered and corresponding values for separation and reliability (expressed in logits)

	COUNT PERFECT SCORES		PERSON SEPARATION		PERSON RELIABILITY		ITEM SEPARATION		ITEM RELIABILITY	
	LVT	WVT	LVT	WVT	LVT	WVT	LVT	WVT	LVT	WVT
150 ITEMS	33	60	2.46	1.93	.86	.79	1.64	1.09	.73	.54
121 ITEMS	4	31	2.46	1.93	.86	.79	2.21	1.36	.83	.65
105 ITEMS	0	19	2.44	1.94	.86	.79	2.59	1.55	.87	.71
97 ITEMS	0	11	2.41	1.94	.85	.79	2.66	1.70	.88	.74
91 ITEMS	0	11	2.40	1.94	.85	.79	2.77	1.75	.88	.75
81 ITEMS	0	1	2.34	1.94	.85	.79	2.87	2.12	.89	.82

Table 3.9 clearly show how excluding the items where all or nearly all participants found the correct answer enhances the separation and reliability indices. As items with perfect scores are dismissed, the separation between participants slightly decreases or stays the same. That might be interpreted as eliminating items from the test that convey very little information about how differently people perform in a test. On the other hand, when fewer items are considered in the analysis, the separation among them logically increases because the range of difficulty – from the easiest to the most difficult item – might be slightly shorter, but the number of individuals covering that distance is certainly smaller. The person reliability index is barely affected by the number of items in the analysis, because the number of participants remains the same (*N*

= 73); whereas the item reliability dramatically increases, in particular in the written version of the test (Table 3.9).

In this section, we have presented an initial data analysis with a view to selecting the best items in the test depending on how efficiently they distinguished different abilities among a sample of persons with respect to one variable. The next section will focus on determining how well the data conform to the measurement model.

3.2.5.3 Data Analysis: Fit Statistics

The previous section has shown which items to keep. Now, the Rasch model dictates the examination of how well the items and persons included in the analysis conform to the measurement model. The motive for this fit examination is the principle of unidimensionality, which is one of the basic assumptions of the Rasch Model, although only a minority of research studies have reported it (Aryadoust, Ng & Sayama, 2021). According to this principle, the instrument has to measure one trait at a time. The data gathered in a research study meet this condition when the responses show overall characteristics that follow the Guttman pattern (Baghaei & Amrahi, 2011). This pattern implies that a person who has answered a given item in a test correctly is expected to choose the right answer for all the easier items in that test. Similarly, a person is expected to miss the more difficult items if they have been unable to answer the easier ones correctly. Since the Rasch model is a mathematical ideal, there will always be discrepancies between the data observed in reality and the expected Guttman-like pattern of responses. However, if researchers want to keep the properties of fundamental measurement, they should aim at having tolerable

deviation from the model, i.e., at having data that conform to the model enough to achieve invariant interval-level measurement (Bond & Fox, 2015).

An important consequence of the principle of unidimensionality is that the independent responses are considered without assumption of the distribution of persons (Andrich & Marais, 2019). One of the primary functions of the Rasch methodology is to transform ordinal observations into interval, linear, additive measures (Linacre, 2014). As the Rasch Model provides the researcher with measures and not raw counts, they can confidently use parametric statistical tests (Boone et al, 2013). Furthermore, the model is “robust to non-normality of the latent trait and the power and type I error are not affected by a misspecification of the distribution of the latent trait” (Guilleux, Blanchin, Hardouin & Sébille, 2014, p. 342). Consequently, on a practical level, checking that the data show a normal distribution and have equal variance is not necessary because of the intrinsic features of the Rasch Model.

The fit analysis for persons and items should be viewed as a “quality control step” (Boone et al., 2013, 176). The Winsteps® software (Linacre, 2012, 2019) provides the analyst with a table with all the items in order of misfit or inadequacy to the model. Infit and outfit values are given for each entry in the table to carry out a fit analysis. The infit value is an information-weighted sum, where each standardized residual value response is weighted by its variance and then summed. When we divide that total by the sum of the variances, we have the differential effects of those weightings in place.

On the other hand, the outfit statistics refer to unweighted values. Outfit is sensitive to outliers in the data, for example, people who have guessed the correct answer or have made thoughtless errors, whereas infit focuses less on those outliers and more on responses with respect to both item difficulty and

person ability (Boone et al., 2013). Therefore, a higher value indicates “more variation in the observed data than the Rasch model predicted” (Bond & Fox, 2015, 269). Alternatively, negative values are indicative of overfit, or less variation in the observed data than expected.

For an already existing test, where high standards of reliability and validity are sought, Wright and Linacre (1994) suggested analysing all infit and outfit mean-square (MNSQ) values greater than 1.2, because it implies an underfit bigger than 20%. As mean-squares are forced to average near 1.0, a figure above 1.2 means that there is over 20% more randomness in the data than the model predicted. Once all the items or persons underfitting the model are detected, the analysis focuses on the standardized infit and outfit of those items (ZSTD) which are outside the range ± 2.0 (Bond & Fox, 2015). Although Linacre recommended reporting only outfit statistics and not infit statistics unless “the data is heavily contaminated with irrelevant outliers” (2012, 622), the subsequent analyses comprise the close examination of both infit and outfit values for items or persons outside the range ± 2.0 .

Tables 3.10 and 3.11 show the items with the highest misfit indicators in the LVT and the WVT, respectively. Cells have been shaded for those items that meet both criteria (MNSQ > 1.2 and ZSTD outside range ± 2.0).

Table 3.10 – Preliminary Study (May 2019) – Items in the listening vocabulary test with highest misfit values

ITEM	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
L100	62	73	-.19	.34	1.09	.49	2.26	2.69
L50	65	73	-.57	.38	1.15	.63	2.19	2.11
L149	68	73	-1.11	.47	1.02	.17	2.00	1.48
L68	71	73	-2.09	.72	1.04	.29	1.97	1.11
L52	44	73	1.27	.26	1.37	3.92	1.74	4.24
L75	66	73	-.73	.41	1.14	.55	1.65	1.26
L111	62	73	-.19	.34	1.14	.68	1.49	1.29
L51	16	73	3.26	.31	1.18	1.03	1.46	1.72
L123	70	73	-1.67	.60	1.01	.19	1.36	.68
L62	29	73	2.25	.26	1.25	2.30	1.35	2.53
L70	28	73	2.32	.26	1.27	2.32	1.29	2.06
L55	71	73	-2.09	.72	1.04	.29	1.28	.59
L132	34	73	1.92	.25	1.18	1.92	1.21	1.76
BETTER FITTING ITEMS NOT SHOWN								
L60	68	73	-1.11	.47	.92	-.10	.50	-.79
L42	68	73	-1.11	.47	.90	-.14	.54	-.71
L106	50	73	.86	.27	.90	-.90	.78	-1.18
L136	49	73	.93	.27	.90	-1.00	.83	-.96
L145	45	73	1.20	.26	.88	-1.38	.79	-1.48
L59	59	73	.13	.31	.87	-.70	.72	-.92
L14	52	73	.71	.27	.86	-1.25	.75	-1.27
L104	57	73	.31	.30	.85	-.98	.72	-1.05
L48	48	73	1.00	.26	.84	-1.73	.75	-1.53
L83	56	73	.40	.29	.84	-1.11	.69	-1.29
L135	62	73	-.19	.34	.83	-.79	.61	-1.13

Table 3.11 – Preliminary Study (May 2019) – Items in the written vocabulary test with highest misfit values

ITEM	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
W149	72	73	-2.04	1.02	1.07	.39	5.10	2.11
W50	66	73	.10	.42	1.23	.80	2.74	2.51
W31	69	73	-.55	.53	.96	.06	2.30	1.57
W40	72	73	-2.04	1.02	1.06	.38	2.12	1.09
W8	66	73	.10	.42	1.02	.15	1.95	1.63
W108	67	73	-.08	.44	1.18	.61	1.77	1.30
W36	68	73	-.30	.48	1.04	.23	1.73	1.16
W68	72	73	-2.04	1.02	1.06	.38	1.54	.79
W52	43	73	2.28	.26	1.37	3.41	1.42	2.68
W62	31	73	3.11	.27	1.31	2.75	1.42	2.60
W57	42	73	2.35	.26	1.25	2.42	1.38	2.53
W94	70	73	-.87	.61	.92	.01	1.31	.61
W144	63	73	.55	.36	1.07	.38	1.25	.72
W82	69	73	-.55	.53	1.14	.45	1.20	.50
BETTER FITTING ITEMS NOT SHOWN								
W145	47	73	2.00	.27	.89	-1.02	.79	-1.36
W100	65	73	.26	.39	.88	-.34	.58	-.91
W28	57	73	1.20	.30	.86	-.89	.72	-1.07
W96	66	73	.10	.42	.86	-.39	.58	-.80
W103	71	73	-1.31	.73	.85	-.02	.46	-.30
W136	56	73	1.29	.30	.85	-1.04	.71	-1.18
W17	70	73	-.87	.61	.84	-.16	.37	-.69
W135	66	73	.10	.42	.84	-.46	.50	-1.03
W130	68	73	-.30	.48	.83	-.36	.43	-.94
W65	70	73	-.87	.61	.82	-.20	.33	-.79
W42	66	73	.10	.42	.81	-.56	.45	-1.21
W83	65	73	.26	.39	.81	-.64	.62	-.79
W85	72	73	-2.04	1.02	.81	.10	.11	-.78
W143	67	73	-.08	.44	.81	-.50	.42	-1.14
W48	60	73	.90	.33	.76	-1.33	.52	-1.69

If underfit is detected, its possible causes have to be sought because underfit

implies the degradation of the quality of the ensuing measures, as it refers to “noisy or erratic item or person performances, those that are not sufficiently predictable to make useful Rasch measures” (Bond & Fox, 2015, 271). Although some items and persons in the dataset showed clear underfit values, their impact was minimal. When the number of misfitting items was compared with the total ($N = 81$), the percentages of items with underfit to the model were 6.17% for the LVT, and 3.70% for the WVT. The misfitting person in the LVT represented 1.36% of the sample, whereas the four participants in the WVT with misfitting values amounted to 5.47% of all the participants ($N = 73$). Furthermore, the negative influence of those misfitting values onto the measurement quality of the instrument was minimal. The overall infit and outfit standardized values in the LVT and the WVT range from 0 to 0.1, for both persons and items. The ideal standardized value is 0, so the overall fit might be considered more than acceptable. With respect to the mean square values (MNSQ), they range from 0.97 to 0.99, showing closeness to the ideal, as that value is 1 (Table 3.14).

Table 3.14 – Overall fit values for the listening and written vocabulary test expressed in logits (81 items)

	PERSONS						ITEMS					
	MEAN MEASURE	SD	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	MEAN MEASURE	SD	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
LVT	1.79	.82	.99	.10	.97	.10	.00	1.20	1.00	.10	.97	.00
WVT	2.77	1.03	.99	.10	.99	.10	-.04	1.32	1.00	.20	.99	.10

A qualitative analysis was also undertaken to determine if any latent dimensions might be tested. Table 3.15 shows the answers participants chose in the questions that were qualitatively analysed. The shaded cells correspond to the items pointed out by the Rasch analysis as misfitting. The table also features the four options for each of those target items – the option in bold is the correct answer for each item – as well as the number of answers each of them elicited

in the vocabulary tests. After analysing the data featured in that table, reasons for the misfit were sought and possible solutions were implemented to reduce it. Distractors or incorrect options were rewritten either because they came from a higher band than the target word ('messy'), or because they might be confused with the correct option ('pleasant' and 'pleased'), or because they might be testing other knowledge ('have' as verb and auxiliary), or they might become confusing in the translation ('in' and 'you'). In any case, for subsequent analyses, I decided to keep a particularly closer look at all the items shown in Table 3.15.

Table 3.15 – Count of the options chosen by participants for the items with highest misfit ($N = 73$).

ITEM 36 – YOU			ITEM 62 – HAVE		
	LISTENING	WRITTEN		LISTENING	WRITTEN
A) LE	3	3	A) CONSEGUIR	0	1
B) LO	5	0	B) HABER	30	32
C) NOS	12	3	C) PODER	2	0
D) TE	53	67	D) TENER	41	40
ITEM 50 – CONFIDENT			ITEM 70 – MEND		
	LISTENING	WRITTEN		LISTENING	WRITTEN
A) CONFIADO	66	66	A) ABROCHAR	15	14
B) CRUDO	0	0	B) ARREGLAR	29	30
C) FRECUENTE	1	0	C) ORDENAR	15	18
D) PRECISO	6	7	D) SUBRAYAR	14	11
ITEM 52 – CABIN			ITEM 100 – SIDE		
	LISTENING	WRITTEN		LISTENING	WRITTEN
A) CABAÑA	47	43	A) FORMA	3	5
B) CONTABLE	3	1	B) LADO	61	61
C) LAVABO	9	14	C) PUNTO	1	1
D) TAXI	14	15	D) VISTA	8	6
ITEM 57 – PLEASANT			ITEM 149 – IN		
	LISTENING	WRITTEN		LISTENING	WRITTEN
A) AFILADO	4	2	A) AL OTRO LADO DE	2	0
B) AGRADABLE	43	40	B) DENTRO DE	70	72
C) EDUCADO	7	12	C) ENCIMA DE	1	1
D) SATISFECHO	19	19	D) FRENTE A	0	0

This section has examined the possibly misfitting items and persons in the

preliminary study, and assessed their influence on the overall test reliability. The conclusion drawn from those analyses is that the data largely fit the Rasch model. Once we have analysed the data gathered in the preliminary study and checked that they fit the probabilistic model, the following section focuses on simultaneously comparing the test takers' abilities and the item difficulties, which is one of the main features – and strengths – of the Rasch model for data analysis.

3.2.5.4 Data Analysis: Descriptive Statistics and Item Difficulty

As the percentage of correct responses in the LVT and the WVT were 77.93% and 86.35% (Table 3.9), we can conclude that the listening version of the test was consistently more difficult, and that the items in both tests were barely challenging for the participants in the preliminary study.

Three reasons might account for those results. Firstly, because all participants in this preliminary study were recruited from B1-groups that were at the end of their courses. This implies that they might already have covered the syllabus for that language level. A second is that most of those learners still attending classes might be the ones making clear progress along the level; whereas those who were struggling to advance in their learning might already have got frustrated and stopped attending classes. A final reason for those results might be that participation in the study was voluntary and its participants made the extra effort of accessing an online form outside their language classrooms. Experience tells us that good learners are usually the ones willing to be tested in their abilities, and answering all the questions.

For the listening vocabulary test, the mean person ability was 1.79 logits, with a real standard deviation of .82 logits. The items in that test showed a mean

difficulty of 0.00 logits and a standard deviation of 1.20 logits. The results for the written version of the vocabulary test showed a mean ability of 2.77 logits for the participants, with a standard deviation of 1.03 logits, and a mean item difficulty of -.04 logits, with a standard deviation of 1.32 (Table 3.16).

Table 3.16 – Descriptive statistics for the listening and written vocabulary test.

	PERSON						ITEM					
	COUNT	MEAN	SD	MEAN*	REAL SE*	SD*	COUNT	MEAN	SD	MEAN*	REAL SE*	SD*
LVT	81	63.1	8.3	1.79*	.34*	.82*	73	56.9	12.9	.00*	.39*	1.20*
WVT	81	69.9	7.1	2.77*	.47*	1.03*	73	63.0	10.8	-.04*	.55*	1.32*

NOTE: *Results expressed in logits

Figures 3.3 and 3.4 feature Wright maps drawn by the program Winsteps® (Linacre, 2012, 2019). On the left of the vertical axis, these maps show the ability of the 73 participants in the study with respect to their performance in the test, so that the ones at the top represent the best performers in the test. On the right of the axis, the items are classified in terms of difficulty, where those found most difficult to answer are located at the top of the scale. Thanks to these maps, visual comparisons of results in both tests are readily available. In the listening vocabulary test, one item (L51, ‘shut’, 3.26 logits) was clearly the most difficult one, followed by items L23 (‘wide’, 2.46 logits) and item L90 (‘handle’, 2.39 logits). The easiest items in the LVT were L35 (‘switch’, -2.81 logits) and L134 (‘glove’, -2.89 logits). For the WVT, the most difficult item was W51 (‘shut’, 3.18 logits), followed by W62 (‘have’, 3.11 logits). The easiest word in that test was W88 (‘item’, -3.27 logits), followed by a group of 12 items (e.g., ‘pig’ or ‘creature’) all of them with a difficulty of -2.04 logits. It is worth mentioning that all participants in the written test chose the correct translation for item W88 (‘item’), i.e., all participants showed perfect scores. However, it was not removed from the analysis because only 64 participants chose the correct

option for its counterpart in the listening vocabulary test (L88, -0.43 logits), which might imply a significant difference across the two versions of the vocabulary test employed in the preliminary study (section 3.2.5.2).

Furthermore, participants' abilities (marked with an 'X' along the vertical axis) are positioned comparatively higher in the map for the WVT. In fact, only one participant showed an ability slightly inferior to the average difficulty for the items in that test, marked with an 'M' on the right side of the vertical axis. In other words, that participant had an overall chance of getting a correct answer for any item in the test a bit lower than 50%. On the other hand, for the LVT, although no participant showed abilities lower than 0 logits, their mean ability with respect to that test was inferior, and consequently those abilities are positioned comparatively lower along the axis than their counterparts for the WVT.

The Wright Maps show that both the participants and the items are distributed along a continuum, so we can observe different levels of ability (persons) and difficulty (items). As the target words in both vocabulary tests are the same, we can conclude that the format of the test might be responsible for the higher difficulty of the LVT. The participants' abilities – marked with an 'X' on the left of the vertical axis – are comparatively lower in the map for the LVT than in the map for the WVT.

Figure 3.3 – Wright Map – Person abilities and item difficulties for the LVT (81 items; 73 persons)

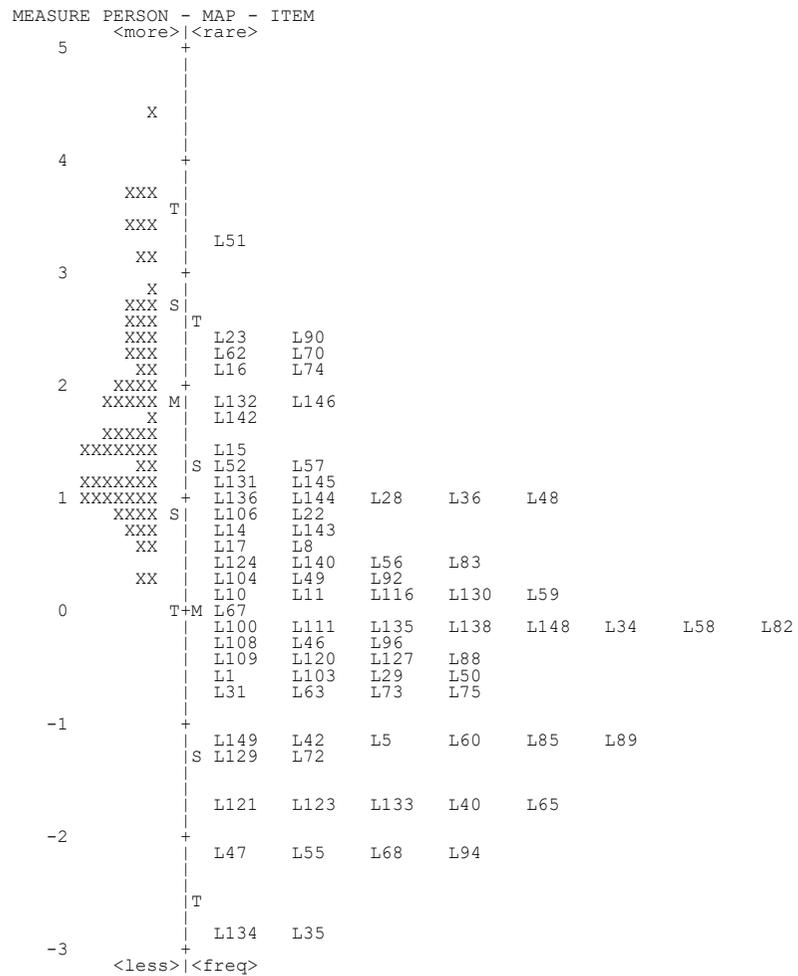
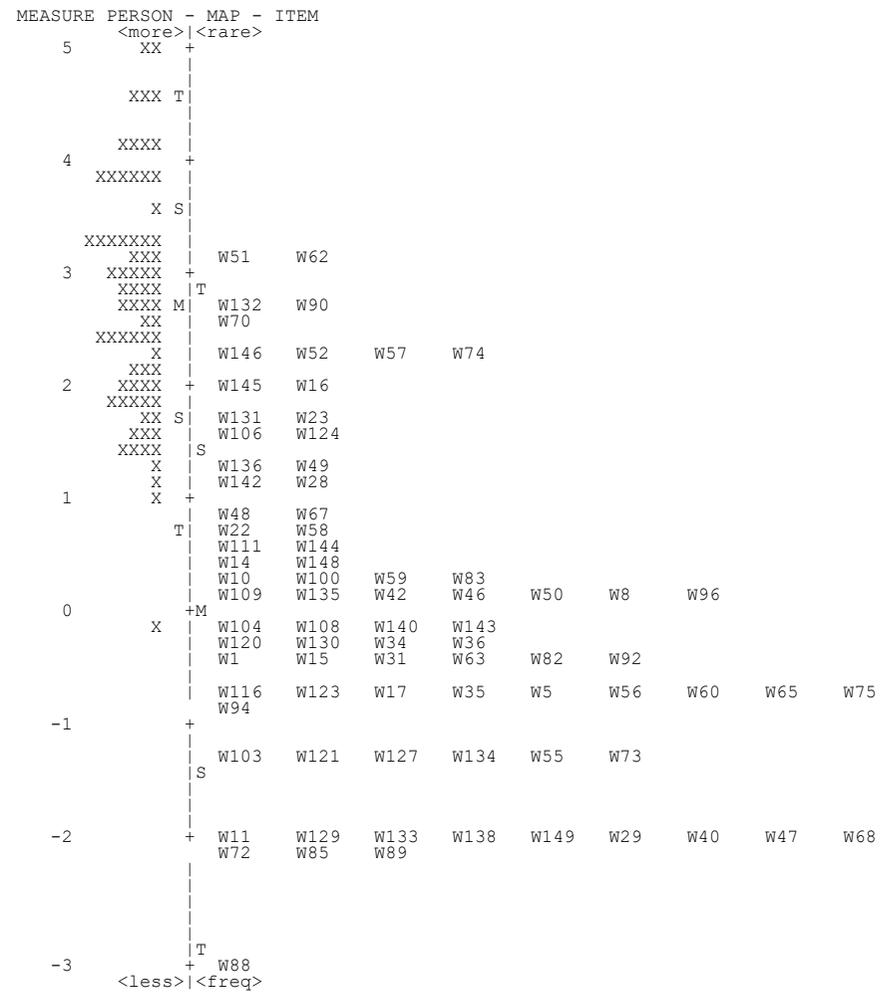


Figure 3.4 – Wright Map – Person abilities and item difficulties for the WVT (81 items; 73 persons)



This section has introduced a brief analysis of the data gathered in the preliminary study with respect to both the participants' abilities and the item difficulties. It has also presented the actual use of logits – one of the main and most useful features of the Rasch model – to compare those abilities and difficulties, and to classify them along a continuum. This classification has been clearly shown in the Wright maps. Once presented the main descriptive statistics, the following section in this chapter will deal with a thorough analysis of the validity of the two vocabulary tests employed in this investigation.

3.2.5.5 Data Analysis: Instrument Validity

Several definitions of validity can be found in the literature depending on their ontological and epistemological stance (Bachman, 1990; Chapelle, 2013; Cicourel, 2007; Dörnyei, 2007; Messick, 1995; Schmuckler, 2001). However, I am more interested in those aspects of validity that have to do with interpreting and generalizing research findings (Brown, 1997). The instruments employed to gather data about the phenomenon under study have to be as accurate as possible, and measure what they are meant to, because findings in L2 research mainly depend on the data collection measures used (Mackey & Gass, 2015).

Four different perspectives on construct validity are discussed here: the content, the substantive, the structural, and the generalizability aspect of construct validity. The discussion will thus follow Messick's idea of a more encompassing approach that includes not only the meaning of the scores in a test, but also its values when interpreting the test itself, and how it is used (Messick, 1995).

The content validity of a test refers to the relevance of its content, its representativeness, and the technical quality it shows (Messick, 1995). The

relevance of this preliminary study of the vocabulary tests with respect to the targeted construct might be seen in the random selection of all the test items from the PET vocabulary list, and in the use of the PET listening paper for the listening comprehension test (section 3.2). The representativeness of the test is understood as the sensitivity it should show towards variations in the measured construct. This representativeness might be seen in the reliability values presented by the Rasch analysis, and in the process of selecting the best performing items in the vocabulary tests (section 3.2.5.2). Furthermore, the analysis of the preliminary study showed that the data presented a clear hierarchy and sufficient spread for both persons and items (Figures 3.3 and 3.4). The representativeness of the test also derives from the sampling process for both the test items and the study participants. The inclusion criteria (Table 3.1) helped in the selection of a representative sample of the target population, the same way as the use of a PET vocabulary list contributed to the representativeness of the items included in the test. The technical quality in this preliminary study of the vocabulary tests was assessed by examining the misfit in their items. Overall fit statistics showed that the incidence of misfit in the vocabulary tests was almost inexistent (Table 3.14). Furthermore, the most misfitting items in both tests were carefully examined, the possible causes for their misfit sought, and several of them either modified or kept under closer examination in subsequent analyses (section 3.2.5.3).

The substantive aspect of construct validity was assessed through two research hypotheses from the construct of vocabulary knowledge and listening performance: the higher the frequency of a test item with respect to frequency lists, the lower its difficulty (hypothesis 1); and test scores will be higher in the written than in the listening version of the vocabulary test (hypothesis 2). I was

unable to confirm the first hypothesis as the three most difficult items in both the LVT ('shut', 'wide' and 'handle') and the WVT ('shut', 'have' and 'handle') came from the first band of most frequent words in English. Furthermore, the easiest item in the LVT was 'glove' (band 4k of frequency), and in the WVT 'item' (band 2k).

On the other hand, the hypothesis about the higher difficulty of the LVT with respect to the WVT was confirmed by the data (Table 3.16, section 3.2.5.4). Furthermore, only 9.87% of the items ($N = 8$) yielded higher scores in the LVT than in the WVT, and the number of participants who did better in the listening test than in the written one was only 4 (5.47% of all participants). A paired t-test analysis confirmed that the differences in the participants' ability in the LVT and the WVT were significant ($df = 72$, t -value = -10.52, p -value < .0001), and yielded large effect sizes, with Cohen's $d = .90$ (section 5.3).

The structural aspect of construct validity was assessed by means of the Rasch Principal Component Analysis (PCA) of item residuals, i.e., the differences between what the Rasch model expects from the items and what the items actually do in a test (Bond & Fox, 2015). The larger the item residuals the more those items have deviated from the Rasch theoretical model. Following the criteria used in similar studies (McLean et al. 2015), seven items in the LVT ('have', 'you', 'handwriting', 'improve', 'ironing', 'improvement', and 'pleased'), and four in the WVT ('recording', 'land', 'improvement', 'refuse') raised concern. A closer examination of those items failed to detect any correlations between excessive residual loadings and item difficulty (McLean et al., 2015). It also failed to discover any common traits in the items that might be indicative of a secondary dimension. Nevertheless, the options in three of those items ('you', 'have', 'pleased') were changed as it was detected that they might be testing an

additional dimension.

The final aspect of construct validity is the generalizability of the study, i.e., whether and how the scores can be applied to other populations, settings or tasks (Messick, 1995). However, the analyses focused on the consistency of the data collected because “[t]he more reliable the sample of performance, or test score is, the more generalizable it is” (Bachman, 1990, 187-188). The principle of invariance was evaluated by examining the differences in mean scores in two randomly selected halves of the exam (Half A vs Half B). If test-takers were assessed with different sets of items from the vocabulary tests, the measures for those people should be similar. There were no significant differences between the scores obtained by the same participant in the first half of the test with respect to the second ($df = 72$, t -value = -1.42, p -value = 0.16 for the LVT; and $df = 72$, t -value = 0.81, p -value = 0.42 for the WVT)

With respect to separation and reliability for persons and items, the values for each half were similar and showed no differences with respect to the overall values yielded by the 81-item test (Table 3.17). However, the person separation and reliability were considerably lower in each of the halves. The reliability indices in the Rasch model are driven primarily by N , so those lower values are the consequence of reducing the number of items to half, while keeping the same population (section 3.1.2.5).

Table 3.17 – Person and Item separation and reliability in two random halves of items in vocabulary tests

	LISTENING VOCABULARY TEST			WRITTEN VOCABULARY TEST		
	81 items	HALF A 41 items	HALF B 40 items	81 items	HALF A 41 items	HALF B 40 items
Person Separation	2.34	1.70	1.51	1.94	1.31	1.33
Person Reliability	.85	.74	.69	.79	.63	.64
Item Separation	2.87	3.13	2.69	2.12	2.17	2.07
Item Reliability	.89	.91	.88	.82	.82	.81

3.2.5.6 Conclusions

This preliminary study offered valuable data to inform the subsequent research process, particularly when deciding the most effective manners to approach the research questions. The insights gained within this study enabled the refinement of its instruments, both in terms of reliability and applicability, by reducing their size from 150 items to 81. The new version of the test was 15 minutes shorter, something of particular importance given the fact that in the longitudinal study, participants would be asked to sit for 25 additional minutes to do a listening comprehension test as well.

Furthermore, the data analysis within the preliminary study allowed the creation of a 'baseline' with respect to which new data can be compared, and new analyses performed. This extent might offer the possibility of having three datasets to be used in the analyses, unlike the single dataset in cross-sectional studies, or the usual two that longitudinal studies provide researchers with. Lastly, on a more practical level, trying out part of the research instruments on a smaller population provided priceless experience in several aspects. It offered an opportunity to become familiar with the software Winsteps® (Linacre, 2012, 2019) for data analysis, as well as valuable practice in dealing with all kinds of logistical problems that might arise when doing empirical research on human subjects.

3.2.6 Main Study – First Data Collection – October 2019

In October 2019, students from 17 B1-level English groups at a state language school were invited to participate in the present research study. This language school had already been used for the recruitment of most participants in the

preliminary study (May 2019, section 3.2.5.1). A total of 284 people agreed to answer the questions in three different tests: a listening vocabulary test (LVT), a written vocabulary test (WVT), and a listening comprehension test (LCT). Participants in the study had to answer 81 vocabulary questions delivered orally (see Appendix 5), then the 25 listening comprehension questions from the exam *Cambridge English: Preliminary* (Appendices 8-11), and finally the same 81 vocabulary questions, but delivered in writing (Appendix 6). 282 participants completed the three tests, whereas one person failed to finish the last part of the WVT, and another participant provided no answers in the WVT.

3.2.6.1 Descriptive statistics, reliability, and separation

Once the data collection finished and all the tests were manually marked, the results were imported onto the program Winsteps® (Linacre, 2012, 2019) to be analysed. The overall reliability of the data showed a slightly higher person reliability for the 81 items in the LVT than in the WVT, and identical reliability for the items in both tests. Person and item separations were also higher in the LVT than in the WVT (Table 3.18).

Table 3.18 – Person and Item reliability and separation in logits – Preliminary Study vs First Data Gathering

	Person Separation		Person Reliability		Item Separation		Item Reliability	
	MAY'19	OCT'19	MAY'19	OCT'19	MAY'19	OCT'19	MAY'19	OCT'19
LISTENING VOCABULARY TEST	2.34	2.95	0.85	0.90	2.87	6.73	0.89	0.98
WRITTEN VOCABULARY TEST	1.94	2.73	0.79	0.88	2.12	6.47	0.82	0.98
LISTENING COMPREHENSION TEST	NA	1.83	NA	0.77	NA	8.49	NA	0.99

All the values increased with respect to the previous preliminary study, particularly for the item separation and item reliability. Two reasons might

account for this increase. Firstly, the higher number of participants (73 vs 284). Secondly, unlike the preliminary study (May 2019), the first data gathering took place at the beginning of the academic year (October 2019), when the study participants had just started their B1-level courses. A more numerous sample of participants is likely to imply more heterogeneity in their spectrum of language proficiency, which might lead to both a bigger separation in the items, and a higher level in their reliability. Furthermore, as the participants were in the first weeks of their B1-level classes, ceiling effects with respect to the item difficulty might be harder to observe, when compared to the results at the end of the academic year (section 3.1.2). Consequently, the main descriptive statistics showed higher values in the preliminary study than in the data gathered in October 2019 (Table 3.19).

Table 3.19 – Comparison of MIN, MAX, MEAN and percentage of correct answers (raw data) – Preliminary Study (May 2019) vs First Data Gathering (October 2019)

	MIN		MAX		MEAN SCORE		SD		% CORRECT	
	MAY'19	OCT'19	MAY'19	OCT'19	MAY'19	OCT'19	MAY'19	OCT'19	MAY'19	OCT'19
LVT (81 items)	44	23	79	77	63.12	49.44	8.35	12.32	77.93%	61.04%
WVT (81 items)	40	25	81	79	69.94	58.37	7.37	10.92	86.35%	72.06%
LCT (25 items)	NA	1	NA	24	NA	13.26	NA	4.50	NA	53.04%

3.2.6.2 Data Quality Analysis

Tables 3.20 to 3.25 (Appendix 14) show the items or persons whose behaviour in either the LVT, the WVT, or the LCT was abnormal (section 3.2.5.3). Five items in the LVT (L51, L52, L108, L1 and L132), six items in the WVT (W52, W1, W51, W16, W58 and W62), and four items in the LCT met the criteria for further analysis of misfit (mean-square values >1.2, and standardized figures outside the range ± 2.0). With respect to the participants' behaviour in the three tests, 17 test-takers failed to conform to the Rasch model in their answers in the LVT. In the case of the WVT, 15 participants showed abnormal behaviour in

their answers, whereas 13 participants in the LCT presented unexpected patterns in the answers they had chosen. In all these cases, the abnormal behaviour implied underfit, i.e., values diverging excessively from what the probabilistic Rasch model expected.

3.2.6.3 Effect of misfit on data quality

Table 3.26 shows how there was more misfit in the outfit statistics (occurring in 6.17%-16% of the items) than in their infit counterparts (1.23%-3.70% of the total number of items). Nevertheless, the overall infit and outfit values for the LCT, the WVT and the LCT show that the negative influence of those misfitting values onto the measurement quality of the instruments was limited (Table 3.27).

Table 3.26 – First Data Collection (October 2019) – Percentage of misfitting items or persons vs overall counts

	TOTAL COUNT		MISFITTING - INFIT		MISFITTING - OUTFIT	
	ITEMS	PERSONS	ITEMS	PERSONS	ITEMS	PERSONS
LVT	81	284	1 (1.23%)	6 (2.11%)	5 (6.17%)	15 (5.28%)
WVT	81	282.2	3 (3.70%)	11 (3.89%)	6 (7.40%)	11 (3.89%)
LCT	25	284	1 (1.23%)	8 (2.42%)	4 (16%)	9 (3.17%)

Table 3.27 – First Data Gathering (October 2019) – Summary of fit values for items and persons expressed in logits.

	PERSONS						ITEMS					
	MEAN MEASURE	SD	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	MEAN MEASURE	SD	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
LVT	.60	.81	1.00	.00	1.01	.07	.00	.99	1.00	-.05	1.01	.00
WVT	1.33	.87	.99	.10	.96	.00	.00	1.16	1.00	.10	.96	-.10
LCT	.13	.97	1.00	.00	1.01	.00	.00	1.33	.99	.00	1.01	.10

The ZSTD and MNSQ figures shown in Table 3.27 deviate minimally from their respective ideals (0 for the ZSTD, and 1 for the MNSQ). As its test-specific pattern “specifies exactly how well the test can be expected to perform on any

application to any sample – past, present or future” (Wright, 1991, 157-158), we might assume that the measures shown in the table are extremely close to reality (Hatch & Lazaraton, 1991, 253). Furthermore, there is minimal variation when the fit measures from both data gatherings are compared with each other (Table 3.28).

Table 3.28 – Summary of mean measures, standard deviation and fit values for items and persons expressed in logits. Preliminary Study vs First Data Collection.

		PERSONS						ITEMS					
		MEAN MEASURE	SD	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	MEAN MEASURE	SD	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
LVT	MAY'19	1.79	.82	.99	.10	.97	.10	.00	1.20	1.00	.10	.97	.00
	OCT'19	.60	.81	1.00	.00	1.01	.07	.00	.99	1.00	-.05	1.01	.00
WVT	MAY'19	2.77	1.03	.99	.10	.99	.10	-.04	1.32	1.00	.20	.99	.10
	OCT'19	1.33	.87	.99	.10	.96	.00	.00	1.16	1.00	.10	.96	-.10

Table 3.29 shows the fit statistics of those items in the preliminary study that were discovered to behave *abnormally* with respect to their predictability within the Rasch model (section 3.2.5.3). Shaded cells present values that failed to meet the criteria for fit. Most of the items in the preliminary study that had prompted a further analysis from a qualitative perspective raised no flags with respect to their fit statistics in October 2019. They might have improved because they had benefitted both from an increase in the sample of test-takers ($N = 73$ vs 284), and from the changes in their distractors introduced after a careful qualitative analysis (section 3.2.5.3). The only exceptions were items L52 and W52 ('cabin'), L70 ('mend'), and W62 ('have').

Table 3.29 – Misfitting items in preliminary study (May'19) vs First Data Gathering (October'19)

ITEM	DATA COLLECTION	MEASURE	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
L100 ('SIDE')	MAY'19	-.19	1.09	.49	2.26	2.69
	OCTOBER'19	-.43	.99	-.21	1.02	.28
L50 ('CONFIDENT')	MAY'19	-.57	1.15	.63	2.19	2.11
	OCTOBER'19	-1.54	1.06	.55	1.14	.76
L52 ('CABIN')	MAY'19	1.27	1.37	3.92	1.74	4.24
	OCTOBER'19	.01	1.28	5.47	1.45	5.57
L62 ('HAVE')	MAY'19	2.25	1.25	2.30	1.35	2.53
	OCTOBER'19	-.41	1.07	1.16	1.10	1.07
L70 ('MEND')	MAY'19	2.32	1.27	2.32	1.29	2.06
	OCTOBER'19	2.01	1.13	1.54	1.24	1.93
W149 ('IN')	MAY'19	-2.04	1.07	.39	5.10	2.11
	OCTOBER'19	-2.68	.96	.00	.68	-.56
W50 ('CONFIDENT')	MAY'19	.10	1.23	.80	2.74	2.51
	OCTOBER'19	-.78	1.06	.56	1.02	.18
W52 ('CABIN')	MAY'19	2.28	1.37	3.41	1.42	2.68
	OCTOBER'19	1.30	1.47	9.44	1.72	9.51
W62 ('HAVE')	MAY'19	3.11	1.31	2.75	1.42	2.60
	OCTOBER'19	.19	1.11	1.69	1.27	2.27
W57 ('PLEASANT')	MAY'19	2.35	1.25	2.42	1.38	2.53
	OCTOBER'19	1.22	1.02	.48	1.08	1.32

In those items that still show underfit, either test-takers with high abilities are answering easy items incorrectly, or persons with low abilities are managing to select the right answer to difficult items. For example, item L52 had an overall difficulty of .01 logits, which means that it was almost the same as the overall difficulty of the test (.00 logits). W52 and W62 showed measures of 1.30 and .19, respectively. Consequently, we might consider items L52 and W62 as average in terms of difficulty, whereas W52 might be a difficult item for the test-takers.

Table 3.30 presents a summary of all the misfitting items in the LVT, WVT and LCT with their respective item measures. In the LCT, item L51 ('shut') was the most difficult in that test with a measure of 2.90, whereas item L108 ('fast') might be considered an easy item. In the WVT, item W51 ('shut') also had the highest measure in the test, which implies that test-takers found it the most challenging item. On the other hand, items W58 ('west') and W62 ('have') might be considered as average, in terms of difficulty, as the mean measure for the

WVT was 0 logits. In the listening comprehension test, item LISTEN13 was the least difficult item in the test, with a measure of -2.08 logits. Based on this variety of item measures we may conclude that both guessing and carelessness took place in the first data gathering, causing a certain degree of misfit. The Rasch Model is probabilistic, which implies that an able test-taker should answer the easy questions correctly whereas a *weak* test-taker should miss the tough items (Rasch, 1960 in Wright, 1997, 37).

Table 3.30 – Misfitting items in First Data Gathering (October 2019) in LVT, WVT and LCT with their item measures expressed in logits.

ITEM	MEASURE	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
L51 'shut'	2.90	1.10	.77	1.51	2.22
L52 'cabin'	.01	1.28	5.47	1.45	5.57
L108 'fast'	-1.26	1.08	.82	1.39	2.15
L1 'ticket'	-.13	1.16	3.10	1.30	3.48
L132 'cabinet'	1.00	1.17	3.44	1.24	3.62
W52 'cabin'	1.30	1.47	9.44	1.72	9.51
W1 'ticket'	-.07	1.23	2.83	1.58	3.77
W51 'shut'	3.23	1.17	1.49	1.39	2.14
W16 'term'	1.91	1.23	4.08	1.33	4.13
W58 'west'	.04	1.11	1.51	1.28	2.11
W62 'have'	.19	1.11	1.69	1.27	2.27
LISTEN13	-2.08	1.07	.62	1.64	2.41
LISTEN25	-.38	1.20	3.67	1.31	3.05
LISTEN5	.41	.99	-.14	1.27	2.87
LISTEN6	.51	1.15	2.80	1.20	2.10

Although the impact of those items and persons on the overall fit values is minimal (Table 3.28), the logical step for the test designer would be to delete those misfitting items or persons from further research or analyses. However, this research study had a longitudinal design to enable comparisons for the same items and persons at different points in time (preliminary study, baseline, and baseline + approximately 35 weeks). Dropping some items or persons from future data collections and analyses would mean losing relevant information that might help to answer Research Question 4 (Table 3.2). Having up to three datasets to draw information from might also yield valuable information for the other RQs, so none of the misfitting items or persons were excluded from

further uses or analyses.

3.2.6.4 Conclusions

Section 3.2.6 has addressed the first data collection in the main study (October 2019). When compared to the preliminary study, the dataset from October 2019 presented better statistics with respect to the reliability and separation indices (Table 3.18). The bigger sample size ($N = 284$) might have been the main contributor to the enhancement of those indices. However, a bigger sample of participants is likely to include more individual cases of *outliers*, people who have behaved abnormally, according to a probabilistic model like Rasch. The number of misfitting persons and items in the LVT, WVT and LCT was comparatively higher in the dataset from October 2019 than in the preliminary study. However, the impact of those individual cases on the overall fit statistics was minimal (Table 3.28).

3.2.7 Main Study – Second Data Collection – June 2020

The original plan of this research study was to deliver the same instruments to the same population as in the first data collection (October 2019), but after an observation period of approximately 35 weeks. This second dataset would be compared to the first one to find further evidence for the research questions (Table 3.2). One of the strengths of the research design chosen for this investigation was the use of a longitudinal approach that could enable the comparison of data collected from the same target population through the same research instruments. Unfortunately, the COVID-19 pandemic emerged in Spain in March 2020 and forced the Spanish Government to order the lockdown of the

entire population for eight weeks. All classrooms at schools, academies, or universities were closed until the following academic year in September 2020, and the classes held online. The obvious consequence for this research study was the impossibility of gathering the data as originally planned, and the necessary adaptation to a new and unexpected reality. Despite the efforts to accommodate the study participants' needs and circumstances in the second data collection, the number of participants was extremely low. Only 17 language learners from a population of 284 potential participants (5.99%) took part in the second data collection (June 2020), which impacted negatively on the data reliability and the subsequent claims of significance. Nevertheless, I have decided to present the data analysis as if nothing had happened, so that the reader is able to see the viability and research potential that a longitudinal design might have in similar investigations.

3.2.7.1 *Adapting to a new research environment caused by COVID-19*

As it happened in the preliminary study (section 3.2.5.1), an online version of the test was created on Google Forms®. Based on the experience gathered in the preliminary study, a few changes were made to facilitate the process of answering the questions in the three tests (LVT, LCT and WVT). The three tests were available online for four weeks, and all the participants in the first data collection in October 2019 ($N = 284$) were sent an invitation to answer the questions in the tests.

3.2.7.2 *Data Quality Analysis*

In general, the data in June 2020 yielded worse separation and reliability values

than in October 2019, particularly with respect to the analysis of the items. For example, the values from the WVT show no separation and reliability for the items in that test, and clearly inferior values for the separation and reliability of the person measures. Table 3.31 shows the summary of statistics for the 81 items in the vocabulary tests and in the listening comprehension test from the first and second data collection (October 2019 and June 2020), compared to those obtained with the same target words in the preliminary study (May 2019).

Table 3.31 – Person and Item reliability and separation (logits) across datasets (May'19, October'19, and June'20)

		LVT			WVT			LCT		
		MAY'19	OCT'19	JUNE'20	MAY'19	OCT'19	JUNE'20	MAY'19	OCT'19	JUNE'20
PERSON	SEPARATION	2.34	2.95	2.45	1.94	2.73	1.40	NA	1.83	1.45
	RELIABILITY	0.85	0.90	0.86	0.79	0.88	0.66	NA	0.77	0.68
ITEM	SEPARATION	2.87	6.73	1.05	2.12	6.47	0.00	NA	8.49	1.74
	RELIABILITY	0.89	0.98	0.52	0.82	0.98	0.00	NA	0.99	0.75

Person reliability values depend firstly on the sample ability variance, while the second most influential factor is the test length (Linacre, 2012). As the items in the vocabulary tests were exactly the same for all the data collections, we might conclude that the ability range in the sample of participants in June 2020 was smaller than in October 2019. On the other hand, item reliability depends primarily on the variance in the difficulty of the items, and secondly on the sample size (Linacre, 2012).

The presence of perfect scores has a negative impact on the reliability of a test because they convey very little information about their performance, and the overall standard error of measurement increases (sections 3.1.2.5 and 3.2.5.2). None of the 17 participants in the second data collection of the main study got perfect scores in any of the three tests. However, 18 items in the LVT (22.22%)

and 36 items in the WVT (44.44%) were correctly answered by all test-takers. No ceiling effects were found in the LCT because of perfect scores in the items, but all participants answered one question in that test incorrectly (floor effects). Following a similar qualitative analysis as with the perfect scores in the preliminary study (Table 3.8), I excluded from the analysis the items with perfect scores in both the LVT and the WVT to quantify their impact on the overall reliability and separation. Person indices show no differences, whereas the item reliability increases from 0.52 to 0.60 in the LVT, and from 0.00 to 0.12 in the WVT. Separation indices also improve: 1.05 vs 1.22 for the LVT, and 0.00 vs 0.38 for the WVT (Table 3.32). Although the exclusion of those items from the analysis implies better values for the item reliability and separation, I decided to use the original dataset from June 2020 to preserve the longitudinal character of this investigation. The test items work correctly on the target population – as we can see in the dataset from October 2019 – but the sample was too small ($N = 17$) to generate reliable data.

Table 3.32 – Reliability and Separation indices (logits) with and without perfect scores (June 2020)

	Person Separation		Person Reliability		Item Separation		Item Reliability	
	81 ITEMS	67 ITEMS	81 ITEMS	67 ITEMS	81 ITEMS	67 ITEMS	81 ITEMS	67 ITEMS
LVT	2.45	2.45	0.86	0.86	1.05	1.22	0.52	0.60
WVT	1.40	1.40	0.66	0.66	0.00	0.38	0.00	0.12

On the other hand, the necessary conformity to the Rasch model was ensured by checking for abnormal behaviour among the participants or the items in the tests. Appendix 15 shows that only item L8 in the listening vocabulary test behaved abnormally (sections 3.2.5.3 and 3.2.6.2). In the WVT, no indications of misfit were found, whereas in the LCT, only item LISTEN23 showed signs of abnormal behaviour. When the participants' behaviour in the tests was analysed with respect to its conformity to the Rasch model, only Person1 showed a

slightly erratic pattern when answering the questions in the LVT (Table 3.36).

3.2.7.3 Effect of misfit on data quality

Once we have detected the items or persons that might have behaved unexpectedly according to the Rasch model, it is important to analyse their relative importance within the whole. Table 3.39 shows the percentages represented by those items or persons with respect to the total. The negative influence of those misfitting values onto the measurement quality of the instruments was limited (Table 3.40). Furthermore, there is a clear reduction in the incidence of abnormal behaviour among study participants or test items, and in its relevance within the whole from October 2019 to June 2020.

Table 3.39 – Percentage of misfitting items in first and second data collection – October'19 vs June'20

	MISFITTING (INFIT)				MISFITTING (OUTFIT)			
	ITEMS		PERSONS		ITEMS		PERSONS	
	OCT'19	JUNE'20	OCT'19	JUNE'20	OCT'19	JUNE'20	OCT'19	JUNE'20
LVT	1.23%	0.00%	2.11%	0.00%	6.17%	1.23%	5.28%	5.88%
WVT	3.70%	0.00%	3.89%	0.00%	7.40%	0.00%	3.89%	0.00%
LCT	1.23%	0.00%	2.42%	0.00%	4.16%	4.00%	3.17%	0.00%

Table 3.40 – Second Data Gathering (June 2020) - Overall fit values for the items and persons (expressed in logits)

	PERSONS						ITEMS					
	MEAN MEASURE	SD	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	MEAN MEASURE	SD	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
LVT	1.80	1.09	1.00	.10	.94	-.10	-.63	1.19	1.00	.10	.94	.10
WVT	2.14	.76	.98	.10	.088	.00	-1.00	0.00	1.00	.20	.88	.20
LCT	.96	.86	1.02	.00	.99	.00	.08	1.43	1.00	.00	.99	.00

A further analysis might be the comparison of outfit and infit calculations across the three different datasets collected at three different moments in time, but from a similar population and through the same research instruments. Table

3.41 shows the mean measures, standard deviations and overall fit values for the items and persons from the three datasets, collected in May 2019, October 2019, and June 2020.

Table 3.41 – Mean measure, standard deviation, and fit statistics across datasets (expressed in logits)

		PERSONS						ITEMS					
		MEAN MEASURE	SD	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	MEAN MEASURE	SD	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
LVT	MAY'19	1.79	.82	.99	.10	.97	.10	.00	1.20	1.00	.10	.97	.00
	OCT'19	.60	.81	1.00	.00	1.01	.07	.00	.99	1.00	-.05	1.01	.00
	JUNE'20	1.80	1.09	1.00	.10	.94	-.10	-.63	1.19	1.00	.10	.94	.10
WVT	MAY'19	2.77	1.03	.99	.10	.99	.10	-.04	1.32	1.00	.20	.99	.10
	OCT'19	1.33	.87	.99	.10	.96	.00	.00	1.16	1.00	.10	.96	-.10
	JUNE'20	2.14	.76	.98	.10	.88	.00	-1.00	.00	1.00	.20	.88	.20
LCT	MAY'19	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	OCT'19	.13	.97	1.00	.00	1.01	.00	.00	1.33	.99	.00	1.01	.10
	JUNE'20	.96	.86	1.02	.00	.99	.00	.08	1.43	1.00	.00	.99	.00

The statistics indicate a clear pattern of conformity to the expected values that probability would predict (sections 3.2.5.3 and 3.2.6.3). However, a closer look at the data suggests several differences depending on the type of test and the moment when the data were collected. In general, the LVT and the LCT behaved better than the WVT in the three datasets, as their infit and outfit statistics were closer to the ideal values (1 for MNSQ, and 0 for ZSTD). With respect to the datasets, the first data collection in the main study (October 2019) yielded the best results in terms of fit statistics, whereas the data gathered in the second data collection (June 2020) deviated the most from the expected values.

A final analysis in the second data collection (June 2020) show the misfitting values in the LVT, the WVT and the LCT, along with their measures (Table 3.42). The shaded cells present the values outside the range of conformity to

the Rasch model. As the mean measure for the items in the LVT was -0.63, we might conclude that item L8 ('mug') might have shown misfit because some test-takers with abilities below that mean value answered that difficult item correctly. The other item that waved a red flag with respect to misfit was LISTEN23 in the LCT. In any case, none of previously detected misfitting items showed abnormal behaviour in the dataset collected in June 2020 (Tables 3.10, 3.11, and 3.30). Furthermore, these misfitting items had minimal impact on the overall fit statistics (Table 3.41).

Table 3.42 – Items in Second Data Gathering (June 2020) with biggest misfit in LVT, WVT and LCT with their item measures expressed in logits.

ITEM	MEASURE	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
L8	-.12	1.55	1.31	6.96	3.47
L1	-.12	1.41	1.05	2.71	1.66
L5	-.67	1.39	.83	2.28	1.20
L74	1.33	1.46	1.86	1.69	1.62
L100	-.67	1.24	.60	1.69	.88
W55	-.97	1.05	1.16	.46	1.99
W52	2.26	.53	1.50	2.48	1.62
W35	-.97	1.05	1.14	.44	1.56
W82	-.97	1.05	1.14	.44	1.56
W127	-.17	.78	1.23	.58	1.34
LISTEN23	.49	1.22	1.10	2.24	2.78
LISTEN6	-.12	1.56	1.99	1.79	1.46
LISTEN10	-.88	1.23	.68	1.52	.84
LISTEN25	-.12	1.40	1.50	1.46	.97
LISTEN5	-.12	1.33	1.29	1.38	.84

3.2.7.4 Conclusions

Section 3.2.7 has followed a similar structure as section 3.2.6, but it has included data from the two datasets in the main study to make comparisons. The availability of three different datasets collected from the same or very similar populations has enabled the comparison of statistics across samples and moments in time.

The first part of section 3.2.7 has dealt with the impact of COVID-19 on this research study. This pandemic has clearly affected the original plan, forcing its adaptation to the new circumstances. The most apparent of those changes was

the online version of the research instruments (LVT, WVT and LCT), as the participants in the second data gathering were at home.

Section 3.2.7 has also presented the main features of the second data collection in the main study (June 2020) with a sample of 17 students of L2 English, who had previously participated in the first data collection (October 2019). The data from June 2020 showed lower reliability values and smaller separation in both persons and items, when compared to the dataset from the first data collection. The differences in sample size might be the main reason to account for such low reliability and separation values in the data collected in June 2020. On the other hand, the number of misfitting persons and items in the LVT, WVT and LCT was clearly lower in this dataset than in the previous one (Table 3.39), and their relative impact on the overall fit statistics was minimal (Table 3.40).

3.3 – CHAPTER SUMMARY

In the first part of the chapter, I have presented my positionality as a researcher with respect to reality, and the possible ways to investigate it. I have also shown how the choice of methodology and methods might depend on the final purpose of the inquiry we want to undertake. Then, I have attempted to relate that theoretical stance to this investigation by explaining how my interest in the relationship between L2 vocabulary and listening led me to select one methodological approach in particular. Eventually, it helped me formulate a series of research questions, and define the operational constructs to facilitate the process of finding answers to those questions.

The second part of the chapter has dealt with the methods I have employed to investigate the research topic. In particular, I have given a detailed description of the process of design, creation, adaptation, implementation, and evaluation of the research instruments employed in the project. In this respect, the use of a preliminary study to verify the overall viability of this study, and the efficacy of the research instruments has shown to be particularly fruitful. I have also presented a detailed account of the process of refinement of two vocabulary tests, specifically designed for their use in the main longitudinal study, by selecting the best performing items in a preliminary study (section 3.2.5).

The last sections in this chapter have presented the main features of the two data collections in the longitudinal study (October 2019 and June 2020). Although no significance can be claimed from the small sample size in June 2020 ($N = 17$), the reader is presented with the original data analysis plan as if nothing had happened, so that they can see the viability – and research potential – of a longitudinal design to answer the questions posed in the investigation.

Chapter 4 will show the evidence from the data gathered at two moments in time by means of a listening vocabulary test (LVT), a written vocabulary test (WVT), and a listening comprehension test (LCT). The evidence presented in Chapter 4 will eventually support the possible answers to the Research Questions (Chapter 5), and the subsequent conclusions to be drawn (Chapter 6).

CHAPTER 4 – DATA ANALYSIS AND RESULTS

Chapter 3 has addressed the methodology and methods in this research study. The evidence from the analyses of the data gathered through a listening vocabulary test, a written vocabulary test, and a listening comprehension test is presented now to inform the answers to the research questions in this investigation.

The first section of this chapter will present the main descriptive statistics from the datasets collected within the main study in October 2019 and in June 2020. Then, different statistical analysis will be performed on the data to gain evidence that might support the answers to the research questions.

4.1 – DATA ANALYSIS – DESCRIPTIVE STATISTICS and ITEM DIFFICULTY

This section presents the main descriptive statistics for the data collected through the research instruments – the LVT, the WVT, and the LCT – in each of the datasets. The different tables and figures featured in the following pages will be the base for the further analyses presented to provide evidence for the answers to the research questions in this study (Table 3.2).

4.1.1 First Dataset – October 2019

The descriptive statistics from the dataset collected in October 2019 shows that recognizing the aural form of words might be more difficult than recognizing their written form, as the minimum and maximum number of correct answers, and the percentage of correct answers in the WVT is higher than in the LCT (Table 4.1). These differences in the values from the LVT with respect to the WVT is shown in both datasets. However, the figures in the preliminary study in both the LVT and the WVT are clearly higher when compared to the data gathered in October 2019. Nevertheless, the listening comprehension test (LCT) shows the lowest percentage of correct answers (53.04%).

Table 4.1 – Comparison of MIN, MAX, MEAN and percentage of correct answers – Preliminary Study (May 2019) vs First Data Gathering (October 2019) – Calculations based on raw data

	MIN		MAX		MEAN SCORE		SD		% CORRECT	
	MAY'19	OCT'19	MAY'19	OCT'19	MAY'19	OCT'19	MAY'19	OCT'19	MAY'19	OCT'19
LVT (81 items)	44	23	79	77	63.12	49.44	8.35	12.32	77.93%	61.04%
WVT (81 items)	40	25	81	79	69.94	58.37	7.37	10.92	86.35%	72.06%
LCT (25 items)	NA	1	NA	24	NA	13.26	NA	4.50	NA	53.04%

With respect to the study participants being challenged by the difficulty of the test, we might assume that they found the items in the tests more difficult than a similar sample of target population had in May 2019. The fact that the data

collection for the preliminary study took place at the end of the participants' academic year (May 2019), and the first data collection in the main study was carried out at the beginning of the following academic year (October 2019) might explain those differences (section 3.2.5.4).

Table 4.2 presents an additional perspective on the participants' ability and the items difficulty for both datasets. They showed lower abilities – lower person measures – in the LVT than in the WVT, and comparatively lower values in October 2019 than in May 2019 (section 3.1.2.5). The LCT presents the lowest mean measure, i.e., the lowest person ability. Interestingly, we can observe that the item mean measure – i.e., item difficulty – is exactly the same in all three tests in October 2019, and very similar for the items in the LVT and the WVT tested in May 2019.

Table 4.2 – Mean measure and standard deviation in LVT, WVT and LCT. Preliminary Study vs First Data Collection. - Results in logits.

		PERSONS		ITEMS	
		MEAN MEASURE	SD	MEAN MEASURE	SD
LVT	MAY'19	1.79	.82	.00	1.20
	OCT'19	.60	.81	.00	.99
WVT	MAY'19	2.77	1.03	-.04	1.32
	OCT'19	1.33	.87	.00	1.16
LCT	MAY'19	NA	NA	NA	NA
	OCT'19	.13	.97	.00	1.33

Finally, as it was the case with the preliminary study (section 3.2.5.4), a graphical representation of the relative difficulty of the items and the persons' ability is provided in the Wright maps (Figures 4.1-4.3). The use of Wright maps provides the reader with the possibility of performing visual comparisons between participants' abilities and the relative difficulties of the items included in a test, as items and persons are on the same scale (section 3.1.2.5). Furthermore, as the Rasch analysis provides the researcher with a unit of

measurement, the comparison can be performed with different samples of people, or different items related to the same observed trait.

For the LVT (Figure 4.1) we can see that the most difficult item in the test was L51 ('shut', 2.90 logits). This item lies more than three standard deviations away from the mean measure for the items in this test (marked with an 'M' on the right of the vertical axis). On the other hand, items L5 ('assistant') and L55 ('Hey!') are situated at the bottom of the axis – more than two standard deviations below the mean measure – because they are the easiest in the test. With respect to the participants' abilities, one test-taker clearly shows the biggest ability as their measures are situated more than three standard deviations higher than the mean measure for the persons in that test (marked with an 'M' on the left of the vertical axis). Moreover, when we compare the elements on the left of the vertical axis (participants' abilities) with the ones on the right (item difficulties), we can see that the left side of the axis is slightly skewed towards the top, and the right side towards the bottom. In other words, the test-takers ability was higher than the overall item difficulty, so the average test-taker had a higher probability than 50% of answering an average item in the test correctly.

mean measure for the participants' ability was about the same as the mean difficulty for the items in the LCT (0.13 vs 0.00 logits).

4.1.2 Second Dataset – June 2020

The overall reliability of the dataset collected in June 2020 was affected by the small sample size, so the reader is advised to handle the information from this dataset with caution (section 3.2.7.2). The figures are clearly higher in the results from the preliminary study (May 2019) and the second data collection (June 2020), when compared to the data gathered in October 2019. The percentages of correct answers in each dataset shows that the LCT is the most difficult test, followed by the LVT, and then the WVT (Table 4.3).

Table 4.3 – Comparison of MIN, MAX, MEAN and percentage of correct answers – May'19, October'19 and June'20 - Calculations based on raw data

	LVT (81 items)			WVT (81 items)			LCT (25 items)		
	MAY'19	OCT'19	JUNE'20	MAY'19	OCT'19	JUNE'20	MAY'19	OCT'19	JUNE'20
MIN	44	23	44	40	25	62	NA	1	9
MAX	79	77	79	81	79	80	NA	24	23
MEAN SCORE	63.12	49.44	66.94	69.94	58.37	73.47	NA	13.26	16.29
SD	8.35	12.32	8.88	7.37	10.92	4.50	NA	4.50	3.79
% CORRECT	77.93%	61.04%	82.64%	86.35%	72.06%	90.60%	NA	53.04%	65.18%

When the mean measures are compared, study participants showed higher values in June 2020 than in October 2019 in the three tests (Table 4.4). Furthermore, the mean person in June 2020 in the three tests were clearly more similar to the ones shown in May 2019 than in October 2019. Finally, the study participants showed higher abilities in the WVT than in the LVT, and in the LVT than in the LCT.

Table 4.4 – Mean measure and standard deviation in LVT, WVT and LCT across datasets (May'19, October'19, and June'20) – Results expressed in logits.

		PERSONS		ITEMS	
		MEAN MEASURE	SD	MEAN MEASURE	SD
LVT	MAY'19	1.79	.82	.00	1.20
	OCT'19	.60	.81	.00	.99
	JUNE'20	1.80	1.09	-.63	1.19
WVT	MAY'19	2.77	1.03	-.04	1.32
	OCT'19	1.33	.87	.00	1.16
	JUNE'20	2.14	.76	-1.00*	.00*
LCT	MAY'19	NA	NA	NA	NA
	OCT'19	.13	.97	.00	1.33
	JUNE'20	.96	.86	.08	1.43

(*) Results with poorest reliability values

The Wright maps featured below present a visual representation of the person abilities and item difficulties based on the dataset collected in June 2020. We can observe that item L51 ('shut' 5.15 logits) is still the most challenging word for the study participants in the LVT (Figure 4.4). Unlike what happened in October 2019 (Figure 4.1), none of the participants showed a greater probability of answering this item correctly than incorrectly. The 18 items at the bottom of the map were answered correctly by all study participants ($N = 17$), so they show minimum measures (-2.82 logits). Furthermore, the participants' abilities are clearly above the mean difficulty of the items, as there is only one person whose ability is below that level, marked with an 'M' on the right of the vertical axis.

Figure 4.4– Wright Map – Person abilities and item difficulties in the LVT – (Second Data Collection – June 2020)

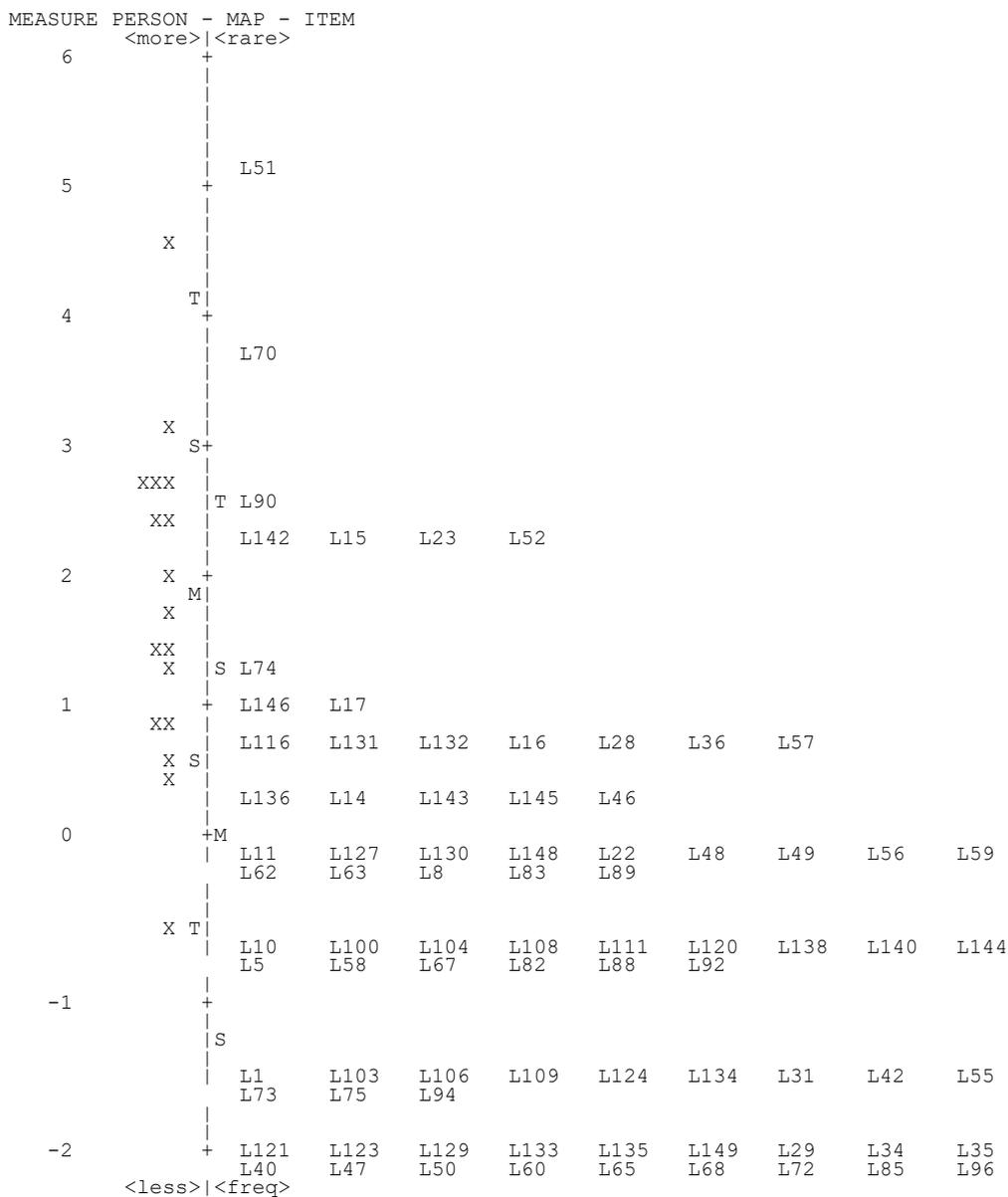
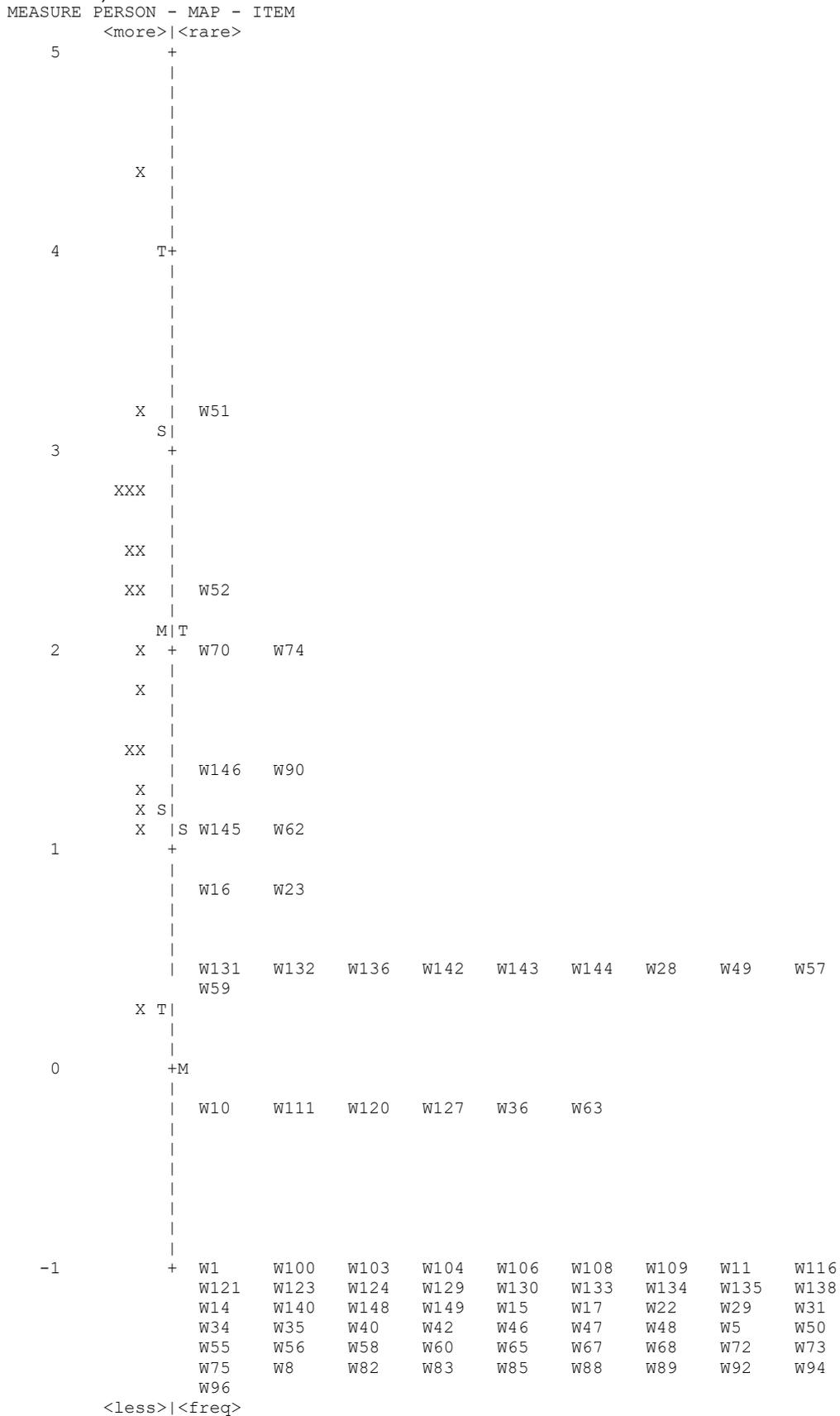


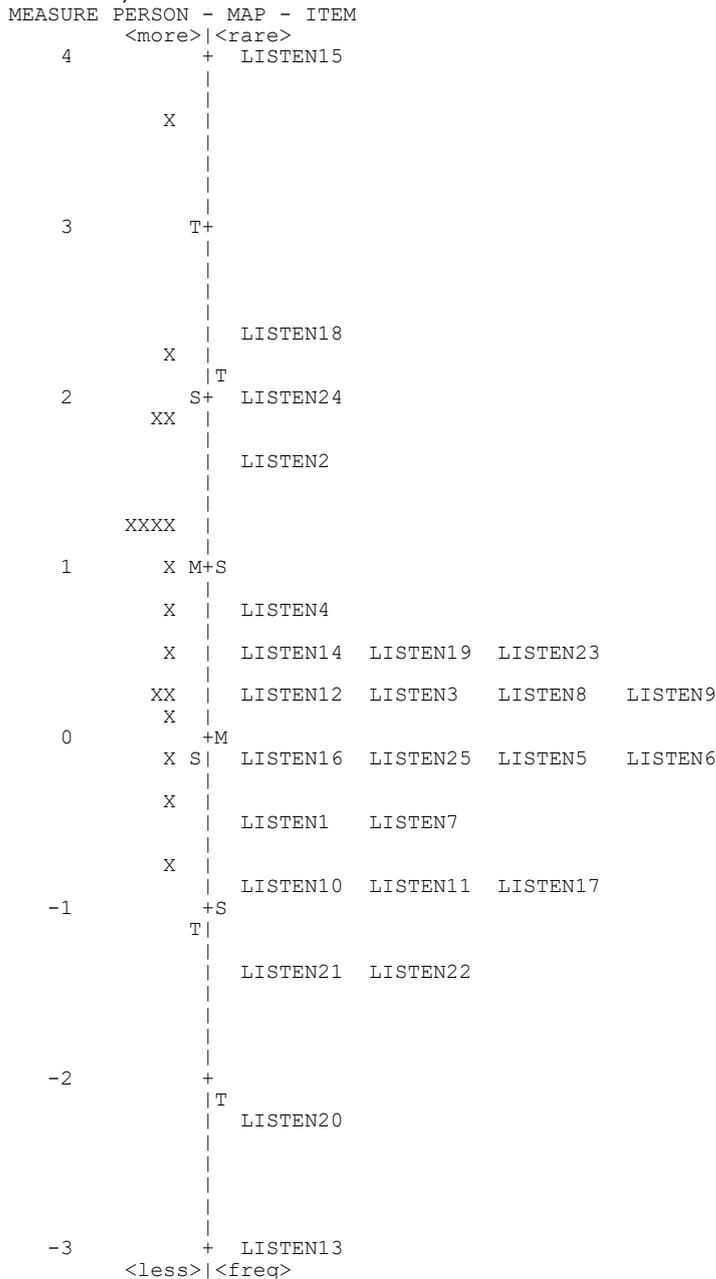
Figure 4.5 shows how the most difficult item in the WVT is W51 ('shut', 3.16 logits), although one participant had an ability higher than the difficulty of that item. Only two items (W51 and W52) are above the participants' mean ability, i.e., an average test-taker had fewer probabilities than 50% to answer them correctly. In the LVT, there were 7 items whose difficulty was above the participants' mean ability.

Figure 4.5 – Wright Map – Person abilities and item difficulties in the WVT – (Second Data Collection – June 2020)



In the LCT, item LISTEN15 is clearly more difficult than the rest, and none of the participants' ability is situated above that item difficulty (Figure 4.6). On the other hand, item LISTEN13 is featured at the bottom of the map, because all test-takers answered it correctly. Although the overall ability is higher than the item difficulty, the differences are less acute in this test than in the LVT and the WVT (Figures 4.4 and 4.5). For example, there are three participants who have fewer chances of finding the correct answer for half of the items, as their abilities are below the mean measure for the item difficulty, marked with an 'M' on the right of the vertical axis.

Figure 4.6 – Wright Map – Person abilities and item difficulties in the LCT – (Second Data Collection – June 2020)



4.1.3 Conclusions

In general, the study participants in the second data collection (June 2020) showed higher abilities in the LVT, WVT and LCT in June 2020 than in October 2019 (Table 4.4). This difference in the person mean measures might be due to the moment when the data were gathered within the participants' courses: either at the beginning of their academic year (October 2019), or at the end of their courses (June 2020). Furthermore, the participants showed their lowest

abilities in the LCT, then in the LVT, and finally in the WVT. Table 4.3 confirms these measures, as we can see in the percentages of correct answers that each dataset yielded for each of the research instruments.

In the following sections, different statistical analyses on the data gathered in the main study are presented. As research objectives and not paradigms or methods should drive studies (Onwuegbuzie & Leech, 2005), I will address my research questions one by one, and present the relevant analyses and evidence accordingly.

4.2 – RESEARCH QUESTION 1: *How much of the listening performance in an exam might be attributed to knowing the words in a vocabulary list?*

The first research question aims to discover if there is a relationship between an L2 learners' vocabulary size and their ability to understand aural texts. It also aims to determine to what extent the aural and the written vocabulary knowledge (independent variables) influence the listening ability (dependent variable).

4.2.1 Data from October 2019

Pearson product-moment correlations were computed for the participants' scores in the LVT, WVT and LCT. Instead of using the raw scores in the tests, the correlations were based on the person measures for each participant in those tests, expressed in logits (section 3.1.2.5). The principle of unidimensionality – one of the basic assumptions of the Rasch model – implies that independent responses are considered without assuming any kind of distribution of persons (Andrich & Marais, 2019). Therefore, the assumption of a normal distribution and equal variance in the data is irrelevant when the Rasch model is used, and researchers can confidently use parametric statistical tests (Boone et al., 2013).

There was a significant positive correlation between the listening vocabulary test and the written vocabulary test: $r(282) = .82$, $z = 1.18$, $p < .0001$. The correlation was also positive between the listening vocabulary test and the listening comprehension test: $r(284) = .56$, $z = .63$, $p < .0001$. A positive correlation, although slightly weaker, was also found between the WVT and the LCT: $r(282) = .41$, $z = .46$, $p < .0001$. Therefore, from a statistical point of view,

both dimensions of vocabulary knowledge, as defined in the person measures of the LVT and the WVT, might be regarded as having similar strong associations with listening comprehension. Following Cohen's typology (Cohen, 2013), the effect sizes were large for the correlation in the dyads LVT-LCT, and LVT-WVT. The correlation between the WVT and the LCT had a medium effect size.

However, Pearson's product-moment only shows possible correlations between values in the tests, without analysing the contribution of the independent variables (vocabulary knowledge in aural and written form) to the dependent variable (listening comprehension). Consequently, a multiple linear regression was calculated to predict the results in the LCT based on the results in both vocabulary tests. A significant regression equation was found ($F(2, 280) = 67.12, p < .0001$) with an $R^2 = .324$, although only the measures in the LVT were significant predictors of the results in the LCT (Table 4.5)

Table 4.5 – Multiple regression analysis of the LCT ($N=283$) – Calculations made on the person measures – (October'19)

Coefficients				
MODEL	coefficient	st.error	<i>t</i>	<i>F</i>
(Constant)	-0.13	0.11	-1.27	0.20
LVT	0.90	0.11	8.12	0.00
WVT	-0.19	0.10	-1.85	0.07
Model Summary				
MODEL	R	R ²	Adjusted R ²	Std. Error of the Estimate
	0.569	0.324	0.319	0.928

A subsequent analysis checked that all the necessary assumptions for the multiple regression analysis were met. In particular, the Durbin-Watson statistic determined the independence of residuals as d was 1.96, which is clearly within

the thresholds of 1.80 and 2.45 that delimit absence of serial correlation ($df = 280$, $CI = .99$). Furthermore, heteroscedasticity was ruled out as both the Glejser Test and the White test failed to reach levels of significance, with p -values of .88 and .93, respectively.

Moreover, when the independent variables were entered into a single linear regression, the results showed the higher predictive power of the LVT over the WVT to account for the variability in the results of the LCT. The measures from the LVT were able explain up to 31.3% of the variance in the LCT ($F(1, 281) = 129.70$, $p < .0001$), whereas the WVT could explain only 16.2% of that variance on its own ($F(1, 281) = 55.51$, $p < .0001$).

The outcomes from both the correlation and the multiple regression analysis may reflect a case of multicollinearity between listening and written vocabulary knowledge. Multicollinearity refers to the situation in which two or more independent variables in a multiple regression model are highly correlated, with values of .90 or higher (Plonsky & Ghanbar, 2018). Taken individually, each of the variables is a reliable predictor of the listening comprehension scores ($p < .0001$), and can explain a portion of the variance in listening comprehension ($R^2 = .313$, and $.162$, respectively). However, their high intercorrelation supports the idea that both variables explain much of the same variance in the listening comprehension test.

An additional insight into Research Question 1 might be gained from the view of successful comprehension as an 'all or nothing' concept. In this research study, the construct of listening comprehension is operationalized through the corresponding paper of a standardized test (section 3.1.2.4). Based on the information published by the institution responsible for that examination (UCLES, 2019), only 48 participants out of 284 (16.55%) passed the LCT as

they answered at least 72% of the questions correctly (i.e., score $\geq 18/25$).

Table 4.6 shows the mean scores and measures for the participants in each test, and the correlations between the person measures.

Table 4.6 – Comparison of scores and measures in LVT and WVT according to performance in LCT ($N = 284$)

	COUNT	LCT MEAN SCORE	LCT MEAN MEASURE	LVT MEAN SCORE	LVT MEAN MEASURE	WVT MEAN SCORE	WVT MEAN MEASURE	CORREL. LVT-LCT	CORREL. WVT-LCT
BOTTOM LCT (Scores < 18)	236	11.97	-.20	47.15	.44	56.70	1.19	.40	.20
TOP LCT (Scores ≥ 18)	48	20	1.88	60.71	1.40	66.48	2.06	.54	.54

Note: Count and scores are raw data, measures are expressed in logits. Correlations are based on person measures.

Among the top scorers in the LCT, their measures in that test were significantly different ($p < .0001$) from their measures in the LVT. Similarly, the differences between the measures of the bottom scorers in the LCT and their measures in both the LVT and the WVT were also significant ($p < .0001$). However, no significance level was reached for the differences among the top scorers in the LCT when their measures in that test and in the WVT were compared ($p = .09$).

A further step in analysing possible differences in the results in the LVT and the WVT depending on the participants' scores in the LCT consists of examining the contribution of those variables to listening comprehension. Unlike the results featured in Table 4.5, and based on all the measures in the three tests, the analysis was carried out here for two distinct groups: top measures in the LCT, bottom measures in the LCT. The overall regression model for the top scorers in the LCT was able to explain 31.5% of the variance with the help of their scores in the LVT and the WVT (Table 4.7). The values clearly reached the significance level: $F(2, 45) = 10.34$, $p < .0001$, $R^2 = .315$. However, when the p -values for each of the two independent variables were analysed, the probability of a contribution of either the LVT or the WVT to the variance in the LCT results due to chance was higher than 5% ($p = .28$; $p = .21$). Among the participants

who had fewer than 18 correct answers in the LCT (i.e., < 72% correct answers), up to 18.6% of the variance in their results could be accounted for by their results in the LVT and the WVT. The significance level was also reached here for the overall model: $F(2, 232) = 26.55, p < .0001, R^2 = .186$, and unlike what happened with the top LCT scorers, both independent variables reached the significance level in their ability to predict variability in the LCT results ($p < .0001$). Table 4.8 shows the main statistics in the multiple regression analyses for the person measures of the participants in the bottom-LCT group (scores <18).

Table 4.7 – Multiple regression analysis for top LCT scores in October 2019 ($N = 48$)

TOP LCT MEASURES				
Coefficients				
MODEL	coefficient	St.error	<i>t</i>	<i>F</i>
(Constant)	1.119	0.192	5.821	0.000
LVT	0.225	0.205	1.098	0.278
WVT	0.216	0.170	1.268	0.211
Model Summary				
MODEL	R	R ²	Adjusted R ²	Std. Error of the Estimate
	0.561	0.315	0.284	0.514

Table 4.8 – Multiple regression analysis for bottom LCT scores in October 2019 ($N = 235$)

BOTTOM LCT MEASURES				
Coefficients				
MODEL	coefficient	St.error	<i>t</i>	<i>F</i>
(Constant)	-0.17	0.09	-1.85	0.07
LVT	0.65	0.10	6.48	0.00
WVT	-0.26	0.09	-2.89	0.00
Model Summary				
MODEL	R	R ²	Adjusted R ²	Std. Error of the Estimate
	0.432	0.186	0.179	0.766

When each of the independent variables was introduced separately into the

regression, their ability to explain the variance in the results of the LCT was reduced with respect to the overall model. In particular, for the group of participants who failed the LCT (< 72% correct answers, i.e., <18/25) the ability of the WVT on its own to account for the variability of results in the LCT was clearly smaller. For this group of results, the measures in the LVT were able to explain up to 15.7% of the variability in the LCT ($F(1, 234) = 43.36, p < .0001, R^2 = .157$), whereas the WVT could account for only 3.9% of that variance ($F(1, 234) = 9.44, p = .002, R^2 = .039$). For the variability of the results in the LCT among those participants who had 18 or more correct answers in that test (i.e., 72% correct answers), both the LCT and the WVT were equally predictive. The results those participants had in the LVT were able to explain up to 27% of the variance in the LCT ($F(1, 46) = 16.68, p = .0002, R^2 = .270$), whereas their results in the WVT could account for up to 27.7% of that variance ($F(1, 46) = 17.28, p = .0001, R^2 = .277$).

4.2.2 Data from June 2020

The three tests correlated similarly in October 2019 and in June 2020 (Table 4.9). In both datasets, the LVT and the WVT correlated higher than any other combination. With respect to the listening comprehension test, its correlation statistics were higher with the LVT than with the WVT. All the correlations were significant except for the dyad WVT-LCT in June 2020.

Table 4.9 – Pearson product-moment correlations among LVT, WVT and LCT. October 2019 vs June 2020. - Calculations made on person measures

TEST	LISTENING COMPREHENSION TEST		LISTENING VOCABULARY TEST	
	October 2019	June 2020	October 2019	June 2020
LISTENING VOCABULARY TEST	.56	.51		
WRITTEN VOCABULARY TEST	.41	.30*	.82	.84

Note: p -value >.05

The multiple regression analysis carried out on the data collected in June 2020 showed that 31.2% of the variance in the results of the LCT could be explained by the independent variables of the study, i.e., the scores in the LVT and the WVT (Table 4.10). The residuals in the model were analysed and heteroscedasticity was initially ruled out as the p -values for the Glejser and the White tests were .33 and .67, respectively. However, the Durbin-Watson statistic ($d = 3.44$) indicated lack of independence in the residuals.

Although the overall model of regression for the dataset collected in June 2020 fails to reach the significance level ($p = .07$), and its small sample of participants ($N = 17$) prevents us from being confident in our conclusions (3.2.7.2), these results in the multiple regression analysis are in line with the ones drawn from the dataset collected in October 2019. Furthermore, when each of the independent variables was introduced in a single linear regression, the results were also similar. The person measures in the LVT could explain up to 25.9% of the variability of scores in the LCT scores ($F(1, 15) = 5.24, p = .04$), whereas the WVT could explain only 9.2% of that variance on its own ($F(1, 15) = 1.52, p = .24$). However, only the regression with the results from the LVT was significant ($p = .04$).

Table 4.10 – Multiple regression analysis of LCT in June 2020 ($N = 283$) – Calculations made on the person measures.person measures.

Coefficients				
MODEL	coefficient	St.error	t	F
(Constant)	0.597	0.602	0.993	0.338
LVT	0.762	0.360	2.115	0.053
WVT	-0.470	0.455	-1.033	0.319
Model Summary				
MODEL	R	R ²	Adjusted R ²	Std. Error of the Estimate
	0.558	0.312	0.213	0.952

4.2.3 Conclusions

This section has presented information to provide answers to RQ1 based on the analysis of two datasets collected from the same population in October 2019 and June 2020. Despite the low reliability of the information provided by the data collected in June 2020, similar analyses were carried out, and their statistics presented, in an attempt to show the viability and potential of a longitudinal approach to investigate the research questions in this study (section 3.2.7).

About a third of the total variance in the scores of a standardized listening comprehension test might be attributed to the test-takers' vocabulary knowledge (Tables 4.5 and 4.10). The aural vocabulary knowledge is able to explain a bigger percentage of the total variance in a listening test than the written vocabulary knowledge. When the results from the LVT were entered into a single linear regression analysis, they could explain between 25.9% and 31.6% of the variance in the LCT results. However, when only the measures from the WVT were used, they were able to account for 16.5% of the variance in the results of the LCT in October 2019 ($N = 283$) and 9.18% in June 2020 ($N = 17$). Furthermore, this comparatively higher ability of the aural vocabulary size to predict the performance in a listening comprehension test was particularly evident among weak listeners – i.e., those who had fewer than 72% correct answers in the LCT – because the LVT could account for 15.7% of the LCT variance, and the WVT for only 3.9%. Among the strong listeners – those who had 72% or more correct answers in the listening comprehension test – both the LVT and the WVT were equally able to predict results in the LCT ($R^2 = .270$ and $.277$ respectively).

Moreover, there might be collinearity between the two measures of vocabulary

knowledge employed in this investigation. Based on correlation coefficients across datasets (Table 4.9), we might assume that the LVT and WVT are testing much of the same thing. However, the unique contribution of the LVT to explaining the variance in the LCT presents a solid argument in favour of keeping this type of aural vocabulary knowledge testing in future research. This preference for the LVT over the WVT might be especially important among lower-level language learners in general, and among weak listeners in particular, as it correlates comparatively better with the LCT (Table 4.6), and explains more of its variance (Tables 4.7 and 4.8).

This section has presented some evidence about the importance of vocabulary knowledge and its contribution to listening comprehension. The following section is devoted to comparing the results in the listening and written vocabulary tests with both the words featured in the listening comprehension test, and its scores.

4.3 – RESEARCH QUESTION 2: *How much lexical coverage of a spoken text does a learner need to achieve comprehension in a listening test?*

This Research Question intends to determine how big a learner's vocabulary size should be to understand aural texts (section 3.1.2.3). In this respect, software tools to analyse the frequency of words featured in a text like *Compleat* (Cobb, 2019) are essential in the provision of evidence to answer RQ2.

4.3.1 Data from October 2019

The words featured in the LCT were compared with the items in the PET Vocabulary List (section 3.2.2) and the first result was clear: all the words employed in the LCT were present in the vocabulary compilation. Therefore, we could assume that a person knowing all the words in that list should be able to recognize all the items featured in the LCT. The second step in the analysis consisted of matching the scores obtained in the LVT and the WVT by each participant to their results in the LCT, to set a minimum lexical coverage to achieve comprehension, as it is understood by Cambridge Assessment English (section 3.1.2.4). The average raw score in the for the participants who had 18 or more correct answers in the LCT (October 2019) was 60.71 for the LVT and 66.48 for the WVT (Table 4.6). That implies that a person knowing on average 74.95% percent of the words employed in the LCT in their aural form (i.e., 60.71 correct answers out of 81 questions), or 82.07% of those words in their written form (i.e., 66.48 out of 81 questions), might be able to achieve what Cambridge Assessment English considers successful comprehension in its listening test (UCLES, 2019).

The disparity in the minimal figures depends on which vocabulary test is used, and it might be the result of the differences in difficulty of those tests. The main descriptive statistics in the LVT were comparatively lower than in the WVT across the three datasets (Table 4.3), which might lead to the conclusion that recognizing words in their aural form was more challenging than in their written form (section 4.1.2). Therefore, for a given level of comprehension in the LCT, the same participant might show lower percentages of correct answers in the LVT than in the WVT, because the former is more difficult than the latter.

Although the number of study participants in October 2019 who had at least 72% of correct answers in the LCT was relatively low ($N = 48$), the analysis of their scores in the LVT and WVT might shed light with respect to the lexical coverage necessary to achieve listening comprehension. Table 4.11 shows the mean scores in the LVT and the WVT for those participants who would pass the LCT (UCLES, 2019), and their corresponding coverage of the words featured in the LCT. This close-up focus on the top scorers in the LCT confirms that the LVT was more challenging for the participants than the WVT, even for the best performers in the LCT (section 4.1.2). Secondly, the data show that the higher the scores in the listening comprehension test, the better the results in the vocabulary tests (section 4.2.2). But most importantly, Table 4.11 shows that recognizing *only* 71.71% of the words featured in a listening comprehension test might be sufficient to answer correctly 72% of its questions.

Table 4.11 – Scores in LVT and WVT for top scorers in LCT with corresponding lexical coverage (October 2019) – Calculations based on raw scores - ($N = 48$)

SCORES IN LCT	LISTENING VOCABULARY TEST			WRITTEN VOCABULARY TEST		
	MEAN SCORE	SD	COVERAGE	MEAN SCORE	SD	COVERAGE
18-20 ($N = 32$)	58.09	8.83	71.71%	64.03	7.51	79.05%
21-22 ($N = 13$)	64.85	4.70	80.06%	71.07	4.54	87.74%
≥ 23 ($N = 3$)	70.66	7.77	87.23%	72.66	6.03	89.70%

Within this research study, passing the listening paper in a standardized test – i.e., answering correctly at least 72% of the questions – is synonymous with being a *successful listener* (section 3.1.2.4). Consequently, it might be possible to achieve that level of comprehension without recognising the aural form of almost 30% of the words featured in that paper, or without knowing the written form of more than 20% of those words. For levels of comprehension above 90%, having a lexical coverage ranging from 87.23% to 89.70% of the words employed in the listening paper might be necessary.

An alternative perspective that derives from setting different bands of lexical coverage based on the vocabulary tests, and comparing them with the results in the LCT. As all the words in that test were featured in the PET Vocabulary Test, we can assume that the results in either the LVT or the WVT are synonymous with lexical coverage of the words featured in the LCT. The mean percentage of correct answers for the 72 participants who answered less than half of the questions in the LVT correctly is 41.32%. The mean percentage of correct answers that those 72 participants had in the WVT is 57.18%, whereas in the LCT they answered an average of 40.89% of the questions correctly (Table 4.12). On the other hand, the mean percentage of correct answers for the 21 participants who failed the WVT was 43.97%, whereas they answered on average 39.68% of the questions correctly in the LVT, and 40.38% in the LCT (Table 4.13). Tables 4.12 and 4.13 show that there is a clear relationship between lexical coverage based on the vocabulary tests and the results in the LCT: the higher the scores in either the LVT or the WVT, the better the scores in the LCT. Furthermore, this positive correlation is steady throughout all the bands of results or coverage.

Table 4.12 – WVT and LCT results vs bands of lexical coverage according to results in LVT – October 2019 - (*N* = 283) – Calculations made on raw scores

LVT LEXICAL COVERAGE				WVT RESULTS		LCT RESULTS	
BAND	<i>n</i>	MEAN %	SD	MEAN %	SD	MEAN %	SD
<50%	72	41.32%	4.84	57.18%	9.27	40.89%	3.86
50-59%	59	55.33%	2.32	69.11%	6.36	47.39%	3.49
60-69%	69	64.43%	2.25	73.79%	5.86	56.64%	3.67
70-79%	49	74.96%	2.43	82.09%	4.25	63.10%	3.98
80-89%	26	84.00%	1.89	89.13%	3.87	65.38%	4.15
>=90%	8	90.90%	1.41	92.13%	2.72	74.50%	4.17

Table 4.13 – LVT and LCT results vs bands of lexical coverage according to results in WVT – October 2019 - (*N* = 283) - Calculations made on raw scores

WVT LEXICAL COVERAGE				LVT RESULTS		LCT RESULTS	
BAND	<i>n</i>	MEAN %	SD	MEAN %	SD	MEAN %	SD
<50%	21	43.97%	4.22	39.68%	5.14	40.38%	3.71
50-59%	29	55.60%	2.41	44.40%	6.58	47.72%	3.29
60-69%	67	64.90%	2.20	54.12%	7.73	49.37%	4.08
70-79%	78	75.20%	2.15	61.95%	6.97	52.67%	4.10
80-89%	67	84.39%	2.31	73.37%	8.44	58.21%	4.70
>=90%	21	93.30%	2.11	84.48%	5.73	70.86%	4.14

Moreover, Table 4.14 shows the statistics from paired *t*-tests between the scores within each band of the LVT and the percentage of correct answers obtained by the same participants in either the WVT or the LCT. It also includes the effect sizes of the differences in those percentages of correct answers. All comparisons between one band of scores in the LVT and their counterparts in the other two tests reached the significance level except for two comparisons. The differences between the percentages of correct answers for the top scorers in the LVT ($\geq 90\%$ of correct answers) and the percentages obtained by the same participants in the WVT failed to reach the significance level. Similarly, for the band of scores below 50% of correct answers no significant differences were found in the comparison with the percentages of correct answers obtained by the same people in the LCT ($df = 71$, t -value = .24, p -value .81).

Table 4.14 – Significance and Effect Size of differences between LVT vs WVT and LVT vs LCT results. Bands based on LVT results – October 2019 - ($N = 283$) – Calculations based on raw scores

LVT LEXICAL COVERAGE		WVT RESULTS				LCT RESULTS			
BAND	n	df	t -value	p -value	Cohen's effect size	df	t -value	p -value	Cohen's effect size
<50%	72	71	-12.40	<0.001	2.40	71	.24	.81	.15
50-59%	59	58	-13.17	<0.001	2.45	58	4.39	<0.001	3.08
60-69%	69	68	-11.11	<0.001	1.80	68	4.55	<0.001	2.89
70-79%	49	48	-10.39	<0.001	1.86	48	5.23	<0.001	4.07
80-89%	26	25	-6.43	<0.001	1.46	25	5.67	<0.001	6.15
>=90%	8	7	-1.28	0.38	.48	7	2.90	0.02	5.36

Table 4.15 presents similar statistics, but comparing the different bands of scores in the WVT with the corresponding results in the LVT and the LCT. Only the differences in the scores between the bottom band in the WVT (<50% correct answers) and the corresponding results in the LCT failed to reach the significance level. The differences in the other comparisons reached the significance level and showed large effect sizes, with Cohen's d values ranging from 1.14 to 7.69.

Table 4.15 – Significance and Effect Size of differences between WVT vs LVT and WVT vs LCT results. Bands based on WVT results – October 2019 - ($N = 283$) – Calculations based on raw scores

WVT LEXICAL COVERAGE		LVT RESULTS				LCT RESULTS			
BAND	n	df	t -value	p -value	Cohen's effect size	df	t -value	p -value	Cohen's effect size
<50%	21	20	3.06	0.006	1.14	20	1.28	.22	1.28
50-59%	29	28	7.49	<0.001	2.37	28	3.45	<0.001	3.21
60-69%	67	66	10.35	<0.001	1.95	66	8.05	<0.001	5.22
70-79%	78	77	13.79	<0.001	2.66	77	12.60	<0.001	7.54
80-89%	67	66	9.15	<0.001	1.83	66	11.37	<0.001	7.69
>=90%	21	20	5.97	<0.001	2.14	20	6.09	<0.001	7.41

4.3.2 Data from June 2020

Although some of the analyses and statistics featured in section 4.3.1 were not carried out on the dataset collected in June 2020 (section 4.1.2), comparing the evidence from two datasets collected at two different points in time on the same

population has evident potential. Answers found in the first dataset (October 2019) might be subsequently confirmed by the evidence collected at a subsequent moment (June 2020) on the same population and with the same instruments, which might render the research claims more robust.

The data presented in Table 4.16 confirmed the statistical results that the dataset in October 2019 had indicated (Table 4.11). The LVT was more difficult for the participants in the study than the WVT, even for those whose performance in the listening comprehension test was higher. Furthermore, the positive correlation between the scores in the LCT and the performance of the same participants in the LVT and WVT continued in the dataset from June 2020: the higher the score in the LCT, the bigger the vocabulary size shown by the participants.

Table 4.16 – Descriptive statistics and coverage in LVT and WVT according to LCT results bands (Dataset, June 2020)

SCORES IN LCT	LISTENING VOCABULARY TEST					WRITTEN VOCABULARY TEST				
	MEAN SCORE	SD	COVERAGE	MIN	MAX	MEAN SCORE	SD	COVERAGE	MIN	MAX
<18 (N = 9)	64.77	9.93	79.97%	44	74	73.33	3.39	90.53%	68	77
>=18 (N = 8)	69.38	7.41	85.65%	58	79	73.62	5.76	90.90%	62	80

Note: Values and calculations based on raw scores.

Table 4.17 features the descriptive statistics and coverage from the two datasets in the main study. The average percentage of correct answers in the vocabulary tests for those who had more than 72% correct answers in the LCT increased in June 2020 with respect to the results in October 2019. In other words, the participants showed a higher ability to recognize words in the LVT and the WVT for similar levels of comprehension in the LCT: 85.65% vs 71.71% for the LVT and 90.90 vs 79.05% for the WVT.

Table 4.17 – Descriptive statistics in LVT and WVT according to LCT results bands (October'19 vs June'20)

SCORES IN LCT		LISTENING VOCABULARY TEST				WRITTEN VOCABULARY TEST		
		N	MEAN SCORE	SD	COVERAGE	MEAN SCORE	SD	COVERAGE
<18	OCT'19	237*	47.11	11.64	58.16%	56.33	11.28	69.54%
	JUNE'20	9	64.77	9.93	79.97%	73.33	3.39	90.53%
≥18	OCT'19	47	61.17	8.16	75.52%	66.85	7.13	82.53%
	JUNE'20	8	69.38	7.41	85.65%	73.62	5.76	90.90%

Note: Values and calculations based on raw scores, not measures expressed in logits. (*) N = 236 in WVT

These results might seem logical because of the moment when the data were collected (section 3.2.7.2), and they are indicative of the learning of vocabulary within an observation period of ± 35 weeks. Interestingly, the mean scores in the WVT in June 2020 for the participants who failed the LCT and for those who passed it are very similar: 73.33 vs 73.62, which implies that knowing up to 90.53% of the words in the WVT might be insufficient to achieve comprehension in the LCT. However, we need to assume that these results have to be interpreted with caution, as the sample of participants in June 2020 is small. Unfortunately, the dataset gathered in the preliminary study (May 2019) only collected data from the LVT and the WVT, so no comparisons based on LCT scores could be made.

4.3.3 Conclusions

Section 4.3 has provided evidence for RQ2, and shown first the results of matching successful performance in a listening comprehension test to the scores in two vocabulary tests, and then, the corresponding coverage of the words featured in that listening test.

The first set of evidence derived from the analyses (section 4.3.1) points at the fact that all the words used for the listening paper in the Cambridge English:

Preliminary were featured in the official vocabulary list (UCLES, 2012, 2019). Its publisher claimed that the list was meant to help both language learners prepare for the test, and exam writers create adequate items for this examination (UCLES, 2012). The fact that all the words employed in a given listening test are to be found in that vocabulary list is a confirmation of the validity of that claim.

The second piece of evidence that we might draw confirms that there is a positive correlation between vocabulary knowledge and listening comprehension across levels of performance (Table 4.9). The more vocabulary knowledge a person shows, the higher their ability to comprehend texts delivered orally. Alternatively, a lower performance in listening comprehension is indicative of a smaller vocabulary size, both in its aural and its written form (Tables 4.11-4.13).

The final and probably most important evidence – particularly because of its implications for the classrooms – might be that being able to recognize no more than 71.71% of the words in an aural text might be enough to achieve listening comprehension in the PET examination (Table 4.11), and be considered a successful listener (UCLES, 2019).

4.4 – RESEARCH QUESTION 3: *How similar are the scores in vocabulary size tests based on recognising either the aural or the written form of words?*

The previous two sections of this report have focused on comparing the scores in the listening and the written vocabulary tests to the results in the listening comprehension test. This section presents the comparison of the scores in the LVT with the ones gathered in the WVT to determine differences between testing the aural and the written vocabulary size (section 3.1.2.3).

4.4.1 Data analysis

In all the analyses between the scores in the LVT and the WVT there were differences indicating that the listening vocabulary test was more challenging for the participants than the written vocabulary test. The values for the minimum, maximum and mean scores support the idea that presenting the target words in their aural form implied a comparatively higher difficulty for the participants than in their written form (Table 4.3). Furthermore, with respect to the dataset collected in October 2019, 266 of the participants ($N = 284$, 93.66%) showed better scores in the WVT than in the LVT. Based on the number of correct answers for both tests in that dataset, 72 items (88.89% of the total number) were easier for the test-takers when they were delivered in their written form.

The overall results of correct answers for each of the tests showed a clear difference: there were 18.05% more correct answers in the written vocabulary test than in the listening vocabulary test. Similar results had already been observed in the preliminary study carried out in May 2019, and would appear in June 2020. Table 4.18 shows the percentages of correct answers in the LVT and WVT, both in the preliminary study (May 2019) and in the two data

gatherings in the main study (October 2019 and June 2020). We can observe that the differences in terms of difficulty for the items delivered orally in the LVT are bigger in the first data collection (October 2019) than in the previous preliminary study (May 2019), or in the subsequent dataset from June 2020.

Table 4.18 – Comparing results LVT vs WVT across three datasets (May'19 – October'19 – June'20) – Calculations based on raw scores

	% CORRECT ANSWERS LVT	% CORRECT ANSWERS WVT	% DIF WVT-LVT
PRELIMINARY STUDY - MAY 2019 (81 Items, 73 persons)	77.93%	86.35%	10.80%
FIRST DATASET - OCTOBER'19 (81 Items, 284 persons*)	61.04%	72.06%	18.05%
SECOND DATASET - JUNE'20 (81 Items, 17 persons)	82.64%	90.60%	9.63%

(*) N = 282 in WVT

Table 4.19 shows the number of items with more correct answers in either the LVT or the WVT, as well as the number of participants who have more correct answers in one test or the other. Only 8 items (9.88%) presented better results in the LVT than in the WVT, i.e., they were easier to answer in their aural than in their written form. Interestingly, this result repeats itself across the three datasets employed in the analysis, but only three items ('switch', 'confident', and 'cabin') presented better results in the LVT than in the WVT in at least two of the datasets. The number of items with no differences in the results between one test format or the other is particularly high in the dataset collected in June 2020. This might be due to the fact that 14 of the items (17.28%) got perfect scores in both the LVT and the WVT (section 3.2.7.2). With respect to the percentage of participants in each of the datasets that obtained better results in the LVT than in the WVT, the figures range between 0% and 5.48%, which is in line with the assumption that the listening vocabulary test implied a higher challenge for the study participants than the written vocabulary test.

Table 4.19 – Comparing results LVT vs WVT across three datasets (May'19 – October'19 – June'20) – Calculations based on raw scores

	ITEMS CORRECT LVT vs WVT			PERSONS CORRECT LVT vs WVT		
	LVT > WVT	NO DIF	WVT > LVT	LVT > WVT	NO DIF	WVT > LVT
PRELIMINARY STUDY (81 Items, 73 persons)	8 (9.88%)	5 (6.17%)	68 (83.95%)	4 (5.48%)	5 (6.85%)	64 (87.67%)
FIRST DATASET (81 Items, 284 persons*)	8 (9.88%)	1 (1.23%)	72 (88.86%)	12 (4.24%)	6 (2.12%)	265 (93.64%)
SECOND DATASET (81 Items, 17 persons)	8 (9.88%)	26 (32.10%)	47 (58.02%)	0	0	17 (100%)

(*) N = 282 in WVT

Another perspective on the relative difficulty of one test over the others might come from the comparison of the person measures in each of the tests across the three datasets. Table 4.20 presents the minimum, maximum and mean values in the measures shown by the participants in each of the tests in May 2019, October 2019, and June 2020, respectively. The person measures – the ability shown by the same participants in the three tests – reflect the higher difficulty of the LVT over the WVT in the three datasets. Furthermore, it also reflects that the participants' ability is higher at the end of their academic year than at the beginning of the course (section 3.2.5.4).

Table 4.20 – Comparison of MIN, MAX, MEAN person measures (expressed in logits) – May'19, October'19 and June'20

	MIN PERSON MEASURE			MAX PERSON MEASURE			MEAN PERSON MEASURE (SD)		
	MAY'19	OCT'19	JUNE'20	MAY'19	OCT'19	JUNE'20	MAY'19	OCT'19	JUNE'20
LVT	.22	-1.5	-.55	4.43	3.4	4.52	1.79 (.82)	.60 (.81)	1.80 (1.09)
WVT	-.13	-1.02	.33	6.62	4.32	4.43	2.77 (1.03)	1.33 (.87)	2.14 (.76)
LCT	NA	-3.81	-.77	NA	4.11	3.61	NA	.15 (.97)	.96 (.86)

The person measures in the LVT are lower than in the WVT, which might be clearly indicative of the higher difficulty of the LVT with respect to the WVT. Paired samples *t*-tests indicated that there was a statistically significant difference between the mean measures shown by the participants in the LVT

compared to the WVT across the three datasets (Table 4.21). The effect size of those differences was medium to large. Nevertheless, the largest effect size appeared in the comparison of the person measures in the WVT with respect to their counterparts in the LCT

Table 4.21 – Comparison of *p* values and effect sizes in *t*-tests for person measures across datasets – May'19, October'19 and June'20

	<i>p</i> Value			Cohen's Effect Size		
	MAY'19 (<i>N</i> = 73)	OCT'19 (<i>N</i> =284*)	JUNE'20 (<i>N</i> = 17)	MAY'19	OCT'19	JUNE'20
LVT vs WVT	<.0001	<.0001	.05	.94	.82	.34
LVT vs LCT	NA	<.0001	.008	NA	.45	.79
WVT vs LCT	NA	<.0001	.001	NA	1.15	1.12

A similar analysis was carried out to determine if the differences in the person abilities from one test to the others varied depending on the participants' listening ability, as expressed in their scores in the LCT. Two subgroups of participants were set depending on whether they had 72% or more correct answers in the LCT. Paired samples *t*-tests based on the person mean measures indicated that all the comparisons between tests reached the significance level in the dataset gathered in October 2019, except for the comparison of the LVT with the WVT among the more proficient listeners. With respect to the dataset gathered in June 2020, only the comparisons between the LVT and the LCT, and between the WVT and the LCT among the least successful listeners showed significant differences (Table 4.22).

Table 4.22 – Significance and effect sizes for differences in mean person measures across datasets according to LCT results bands (October'19 vs June'20)

SCORES IN LCT	COMPARISON	<i>p</i> Value				Cohen's Effect Size	
		<i>N</i>	OCT'19	<i>N</i>	JUNE'20	OCT'19	JUNE'20
<18	LVT vs WVT	235	<.0001	9	.112	.80	.61
	LVT vs LCT	236	<.0001	9	.003	.69	1.50
	WVT vs LCT	235	<.0001	9	<.001	1.50	2.21
≥18	LVT vs WVT	48	<.0001	8	.429	.70	.10
	LVT vs LCT	48	<.0001	8	.268	.57	.35
	WVT vs LCT	48	.09	8	.198	.23	.47

Interestingly, the effect sizes of the differences were larger among those participants with lower scores in the LCT. These effects were moderate to large in all the subsets of participants' measures from October 2019, except for the comparison of the WVT with the LCT among the top group of performers in the LCT, which showed a small effect size (Cohen, 2013). In the dataset collected in June 2020, the two subsets that reached the significance level in the differences – LVT vs LCT, and WVT vs LCT, both within the group of less proficient listeners – presented large effect sizes. The statistics presented in Tables 4.21 and 4.22 might indicate that each of the three language dimensions under study in this research – aural vocabulary size, written vocabulary size, and listening comprehension – develop independently from each other to some extent. In other words, students fail to learn uniformly across those dimensions. Furthermore, this lack of uniformity might be particularly acute in the early stages of the learning process, and for less proficient listeners (Table 4.22).

4.4.2 Conclusions

The main conclusion we can extract from the comparison of scores between the LVT and the WVT is that the former is more difficult than the latter. There are

significant differences between the scores each participant has obtained in the LVT and in the WVT, and these differences have persisted in the three datasets used in the analysis. A possible consequence of those differences is that using a written vocabulary test might overestimate learners' aural vocabulary size by up to 18.05% (Table 4.18). Furthermore, the differences in difficulty between the two vocabulary tests were bigger at the beginning of the academic year (October 2019) than at the end of the participants' courses (section 3.2.5.4). The differences between the scores and measures in the LVT and the WVT reached the significance level ($p < .0001$), and had medium to large effect sizes with Cohen's d values between .34 and .94 (Table 4.21). Besides, those differences between the LVT and the WVT might correlate negatively with the students' linguistic ability, as they are more significant and with greater effect sizes at the beginning of the students' academic year, and among less proficient listeners (Table 4.23).

4.5 – RESEARCH QUESTION 4: *How does the relationship between lexical knowledge and listening performance evolve over time?*

Research Question 4 might be seen as a corollary to the investigation of the relationship between vocabulary knowledge and listening comprehension, as it implies the joint analysis of the two datasets in the main study (October 2019 and June 2020). Despite the low overall reliability of the dataset collected in June 2020 (section 3.2.7.2), the data analysis for both datasets was carried out as if nothing had happened. The ultimate aim of such decision was to show the reader the viability and research potential that a longitudinal design might have in similar investigations, especially when answering inquiries like RQ4.

The evolution of the relationship between lexical knowledge and listening comprehension might be observed by comparing the correlation figures for the three tests across the two datasets in the main study (October 2019 vs June 2020). A further set of evidence might come from the comparison of the results in the multiple regression analyses performed on both datasets, as well as according to the linguistic level shown by the participants in the listening comprehension test. Finally, by making the most of the longitudinal design we might also help to determine how much language learners are able to expand their vocabulary knowledge and improve their listening ability within a period of approximately 35 weeks.

4.5.1 Data analysis

The order of Pearson product-moment correlations is the same in October 2019 and June 2020: the highest correlation is to be found in the dyad LVT-WVT, then in the LVT with the LCT, and finally the WVT and the LCT correlated the

lowest (Table 4.23). Those correlations between both vocabulary tests and the listening comprehension test are higher at the beginning of the academic year than at the end of the language courses (section 3.2.5.4). This might be indicative of higher correlation values between vocabulary knowledge and listening comprehension when the overall linguistic level is lower. In other words, among low-level students lexical knowledge and listening ability might be more related than among students with higher language proficiency. The correlation between the two versions of the vocabulary tests (LVT with WVT) also showed an increase at the end of the academic year (June 2020).

Table 4.23 – Pearson product-moment correlations among LVT, WVT, and LVT based on person mean measures across three datasets – May'19 - October'19 - June'20.

	LISTENING COMPREHENSION TEST			LISTENING VOCABULARY TEST		
	MAY'19	OCTOBER'19	JUNE'20	MAY'19	OCTOBER'19	JUNE'20
LISTENING VOCABULARY TEST	N/A	.56	.51			
WRITTEN VOCABULARY TEST	N/A	.41	.30*	.73	.82	.84

(*) p -value $>.05$

The comparison of the multiple regression analysis carried out on each of the datasets shows that the variance in the LCT measures that could be explained by the scores in the vocabulary tests was 32.4% in October 2019, and 31.2% in June 2020 (Table 4.24). The decrease in the amount of variance accounted for by the scores in the vocabulary tests from the first to the second dataset might suggest that the lower the overall language proficiency, the more related are vocabulary knowledge and listening comprehension (section 3.2.5.4).

Table 4.24 – Comparison of multiple regression analysis of the LCT based on person measures - October'19 vs June'20 (N = 283).

Coefficients								
MODEL	coefficient		st.error		t		F	
	OCT'19	JUNE'20	OCT'19	JUNE'20	OCT'19	JUNE'20	OCT'19	JUNE'20
(Constant)	-0.13	0.597	0.11	0.602	-1.27	0.993	0.20	0.338
LVT	0.90	0.762	0.11	0.360	8.12	2.115	0.00	0.053
WVT	-0.19	-0.470	0.10	0.455	-1.85	-	0.07	0.319
Model Summary								
	R		R ²		Adjusted R ²		Std. Error of the Estimate	
	OCT'19	JUNE'20	OCT'19	JUNE'20	OCT'19	JUNE'20	OCT'19	JUNE'20
	0.569	0.558	0.32	0.312	0.319	0.213	0.928	0.952

The possible correlation of low language proficiency with a closer relationship between lexical knowledge and listening comprehension might be observed from a different angle. Table 4.25 presents the mean person measures in the LVT and the WVT according to the student' success in listening comprehension as expressed in their LCT score. The upper rows of the table present the results in the LVT and WVT for those participants who failed the LCT. The lower rows show the mean measures in the vocabulary tests of participants who had at least 72% correct answers in the LCT. The mean measures for both vocabulary tests increased from October 2019 to June 2020 for both groups of performance in the LCT. These values might support the assumption that the participants in the study learned aural and written vocabulary during the observation period of approximately 35 weeks.

Table 4.25 – Descriptive statistics in LVT and WVT based on person measures, according to LCT scores – October 2019 vs June 2020 – (*) N = 236 in WVT

SCORES IN LCT	DATASET	N	LISTENING VOCABULARY TEST		WRITTEN VOCABULARY TEST	
			MEAN MEASURE	SD	MEAN MEASURE	SD
<18	OCT'19	237*	.44	.79	1.19	.88
	JUNE'20	9	1.47	1.10	2.01	.67
≥18	OCT'19	47	1.40	.73	2.07	.88
	JUNE'20	8	2.17	1.30	2.29	1.24

Nevertheless, the evolution of the relationship between vocabulary knowledge and listening comprehension might be seen more clearly if we focus only on those participants whose data were recorded both in October 2019 and in June 2020. There was a clear increase in the scores of the three tests from one dataset to the other, which indicates that those students learned vocabulary and improved their listening ability within the observation period. The percentage of correct answers increased by 17.68% in the LVT, 14.48% in the WVT, and 17.87% in the LCT. Accordingly, the mean person measures in the three tests showed higher figures in the dataset from June 2020 than in October 2019 (Table 4.26). Furthermore, the differences in the scores in the three tests from one dataset to the subsequent are significant, with large effect sizes (Table 4.27). Similarly, if we compare the percentage of correct answers achieved by the same participants in one test, with their scores in another test from the same dataset, we can observe that the differences in those percentages reached the significance level in both datasets, although their effect sizes were small to medium (Table 4.28).

Table 4.26 – Descriptive statistics in LVT, WVT and LCT based on person measures in both data collections (October 2019 and June 2020) – ($N = 17$)

DATASET	LVT		WVT		LCT	
	MEAN MEASURE	SD	MEAN MEASURE	SD	MEAN MEASURE	SD
OCT'19	1.18	1.03	1.82	1.16	.41	1.00
JUNE'20	1.80	1.21	2.14	.96	.96	1.07

Table 4.27 – Statistics for paired *t*-tests on differences between raw scores in LVT, WVT and LCT from October 2019 to June 2020

<i>df</i>	LISTENING VOCABULARY TEST				WRITTEN VOCABULARY TEST				LISTENING COMPREHENSION TEST			
	<i>t</i> -value	<i>p</i> -value	Pearson	Effect size	<i>t</i> -value	<i>p</i> -value	Pearson	Effect size	<i>t</i> -value	<i>p</i> -value	Pearson	Effect size
16	-4.53	.0003	.76	-1.48	-4.80	.0002	.71	-2.47	-4.07	.0009	.82	-.86

Table 4.28 – Statistics for paired *t*-tests on differences between raw scores in LVT-WVT, LVT-LCT and WVT-LCT (October 2019 and June 2020)

	<i>df</i>	LVT vs WVT				LVT vs LCT				WVT vs LCT			
		<i>t</i> -value	<i>p</i> -value	Pearson	Effect size	<i>t</i> -value	<i>p</i> -value	Pearson	Effect size	<i>t</i> -value	<i>p</i> -value	Pearson	Effect size
OCTOBER'19	16	-5.03	.0001	.92	-.13	3.07	.0073	.32	.21	4.777	.0002	.09	.33
JUNE'20	16	-4.01	.0010	.68	-.11	5.01	.0001	.43	.24	6.80	<.0001	.12	.36

Finally, we can divide the participants whose data were recorded both in October 2019 and in June 2020 ($N = 17$) according to their results in the LCT, and compare their measures in both datasets (Table 4.29). Compared to the data presented in Table 4.25, the differences in person measures from the first dataset to the second are smaller in the LVT and the WVT, for both levels of performance in the LCT. The main reason for those smaller differences is that the mean measures in October 2019 for those 17 participants were higher than the overall measures for the entire sample in that dataset ($N = 284$). This might be indicative of relatively higher abilities within the 17 participants when compared to their classmates. The language students with the highest abilities in the class are usually those more willing to be tested, particularly when it involves making the effort of accessing an online form (section 3.2.5.4).

Table 4.29 – Descriptive statistics in LVT and WVT based on person measures, according to LCT scores for participants in both data collections (October 2019 and June 2020) – ($N = 17$)

SCORES IN LCT	DATASET	LISTENING VOCABULARY TEST		WRITTEN VOCABULARY TEST	
		MEAN MEASURE	SD	MEAN MEASURE	SD
<18 ($N = 9$)	OCT'19	.81	1.08	1.79	1.27
	JUNE'20	1.47	1.10	2.01	.67
≥18($N = 8$)	OCT'19	1.59	.91	2.03	.90
	JUNE'20	2.17	1.30	2.29	1.24

When the scores obtained by those 17 participants in either the LVT or the WVT in October 2019 were compared to the ones they had in June 2020, the differences in results reached the level of significance in all the comparisons

(Table 4.30). Furthermore, the effect sizes of those differences were larger among participants with lower levels of listening proficiency. These results might indicate that students expanded both their vocabulary knowledge and their listening ability within their academic year.

Table 4.30 – Statistics for paired *t*-tests on differences between raw scores in LVT, WVT and LCT from October 2019 to June 2020, according to bands of LCT results

	BOTTOM LCT (<i>df</i> = 8)				TOP LCT (<i>df</i> = 7)			
	<i>t</i> -value	<i>p</i> -value	Pearson	Effect size	<i>t</i> -value	<i>p</i> -value	Pearson	Effect size
LISTENING VOCABULARY TEST	-4.03	.004	.76	-1.62	-4.38	.003	.86	-1.13
WRITTEN VOCABULARY TEST	-3.22	.012	.62	-2.80	-4.58	.003	.91	-1.56
LISTENING COMPREHENSION TEST	-3.75	.006	.61	-1.38	-2.07	.077	.29	-1.27

The differences across datasets in the LVT and the WVT might be in line with the claim that learning vocabulary is more apparent within lower levels of linguistic ability, as determined by the scores in a listening comprehension test (sections 4.2.1 and 4.4.1). Furthermore, this relatively bigger expansion of vocabulary knowledge among students with lower proficiencies is experienced in both their aural and written vocabulary size (Table 4.30).

4.5.2 Conclusions

Section 4.5 has focused on the evolution of vocabulary knowledge and listening comprehension throughout time. Several comparisons have been made between the data collected in October 2019 and in June 2020, for the scores and measures shown by participants in each of the tests (LVT, WVT, and LCT). Additional comparisons have been made depending on the scores the participants obtained in the LCT. The data presented in this section has confirmed that the order of correlation in the dyads of tests remains unaltered: LVT-WVT > LVT-LCT > WVT-LCT (Table 4.23). Furthermore, both the LVT and

the WVT correlated better with the listening comprehension test at the beginning of the participants' academic year (October'19) than at the end of their courses (June'20).

A second piece of evidence about the evolution of the relationship between vocabulary knowledge and listening comprehension came from the analysis of the explained variance in the LCT by means of the scores in the vocabulary tests. More variance in the LCT was explained with those results in the first dataset than in the second dataset (Table 4.24). This might be in line with the evolution of the correlation values discussed above, indicating a possibly deeper relationship between vocabulary knowledge and listening comprehension at the beginning of the academic year. In other words, the relationship between vocabulary knowledge and listening comprehension might be closer for language students with lower abilities (section 3.2.5.4). The overall measures in the dataset from June 2020 were higher than the measures collected in October 2019 in both groups of performers (those who had failed the LCT, and those who had passed it), in both the LVT and the WVT (Table 4.25).

A final and clearer perspective on Research Question 4 can be drawn from the comparison of the scores obtained by the 17 participants whose data were collected both in October 2019 and in June 2020 (Table 4.26). Their scores in all the tests improved significantly from the first dataset to the second, with large effect sizes (Table 4.27). Furthermore, the differences in the percentages of correct answers in one test with the other two were significant, although they yielded small to medium size effects (Table 4.28). Moreover, when those 17 participants were divided into bands of performance in the LCT (i.e., weak and strong listeners), and the scores obtained by the participants

within the same band in October 2019 and in June 2020 were compared, the differences were significant, and yielded large effect sizes, especially among weaker listeners (Table 4.30).

4.6 – CHAPTER SUMMARY

Chapter 4 has presented evidence to support answers for each research question posed by this study (section 3.1.2.3). After presenting the main descriptive statistics that each of the two datasets yielded (section 4.1), evidence for every research question has been presented.

Firstly, learners' vocabulary size shows a strong positive correlation with their listening performance in a test. This correlation remains strong and positive over time, and regardless of their listening ability: the wider the vocabulary, the better their listening performance. Alternatively, the higher the listening ability, the wider the vocabulary size.

Secondly, the scores in a vocabulary test can explain a great deal of the variability in the scores in a listening test. About a third of the total variance in a listening test can be accounted for by the aural and written vocabulary size learners have. Furthermore, the contribution of the aural vocabulary size to explaining that variance on its own is clearly bigger than the one made by the written vocabulary size. This is particularly apparent among weaker listeners.

Thirdly, the aural vocabulary size is generally a better predictor of listening success than the written vocabulary size. Among weak listeners its ability is particularly high, whereas among listeners who achieve enough comprehension, both vocabulary dimensions are equally predictive.

A fourth finding presented in this chapter is that the listening vocabulary test and the written vocabulary test are testing much of the same thing. This possible collinearity between the two tests might support a preference for the LVT over the WVT among beginners in second language learning in general, and among weak listeners in particular, as it correlates better with the LCT and

explains more of its variance.

The quest for answers to RQ2 has provided us with our fifth finding: all the words employed in the listening paper of Cambridge English: Preliminary are also featured in the official vocabulary list published by UCLES. But most importantly, this search for answers to RQ2 has shown that knowing just 71.71% of the words in a listening test might be enough to achieve successful comprehension – which is clearly at odds with what the literature has suggested so far (section 2.5).

With respect to RQ3, we have presented evidence that testing the aural vocabulary is more challenging than using the written form of words. Furthermore, language learners might find a listening vocabulary test comparatively more challenging when their overall language level is lower. The differences in difficulty between one test and the other might support the argument of testing learners' aural vocabulary size, particularly if the scores are subsequently matched to their listening performance.

The evidence to support answers to RQ4 confirms several of the previous findings in this study, which adds to the confidence and soundness they might present. The strong positive correlation, and its order in the dyads of tests remains unaltered across datasets. Furthermore, the multiple linear regression analyses show that the amount of variance in the scores of a listening test explained by both the aural and written vocabulary size was similar before and after the observation period: $R^2 = .324$ in October 2019 and $.312$ in June 2020. Additionally, the regression analyses performed on the dataset collected in June 2020 confirm that the aural vocabulary size is a better predictor of results in a listening comprehension test than the written vocabulary size, particularly among weaker language learners. Nevertheless, the most important finding with

respect to the evolution of the relationship between L2 vocabulary and listening is that a learner clearly increases their vocabulary size and improves their listening performance in a test after approximately 35 weeks attending language classes. This increase has been demonstrated to be particularly acute among those learners with a lower language level.

Chapter 5 will contextualise all these research findings by drawing on the relevant literature, and by carrying out a thorough discussion. Then, Chapter 6 will close this dissertation report with a detailed account of the possible implications of those findings both in the language classroom, and in the realm of L2 theory and research.

CHAPTER 5 – GENERAL DISCUSSION

This chapter will address the study findings in an attempt to contextualize the main answers to the research questions presented in Chapter 4. The relevance of those findings will be highlighted by drawing on what previous studies have found, and by stressing the need to investigate some unsolved questions.

The present study has focused on the relationship of learners' aural and written vocabulary size on their listening comprehension ability over time. It has been the first one to use a longitudinal study to investigate both vocabulary dimensions and their influence on listening comprehension at the same time, with the same research instruments, and on the same population, at two points in time. Unfortunately, this novelty in the study design has prevented the possible contextualization of some of its findings by comparing them with previous research studies.

5.1 – RESEARCH QUESTION 1: *How much of the listening performance in an exam might be attributed to knowing the words in a vocabulary list?*

The main finding we can draw from the correlation analyses is that aural vocabulary is a better predictor of listening success than written vocabulary, particularly among weaker listeners. The scores and measures in the listening vocabulary test correlate better with the scores and measures in the listening comprehension test than with the ones in the written vocabulary test, in both datasets (Table 4.6). In the dataset from October 2019, the Pearson product-moment correlation coefficient for the LVT and the LCT was .56 ($N = 284$), whereas the WVT and the LCT correlated at .41 ($N = 282$). Similar values, although lower, could be observed in the dataset gathered in June 2020 ($N = 17$): .51 for the correlation between the LVT and the LCT, and .30 for the correlation between the WVT and the LCT. All those values reached the significance level ($p < .0001$), except for the correlation between the WVT and the LCT in June 2020.

Moreover, in the dataset from October 2019, the comparison of results between participants who passed the LCT or not (cut-off point = 72% correct answers) revealed that the differences between the LVT and the WVT in their ability to predict listening achievement are particularly acute among weak listeners. On the other hand, both vocabulary tests are equally predictive for the participants who passed the test. Among the *strong* listeners both the LVT and the WVT presented identical correlation figures with the LCT (Pearson product-moment = .54). However, among the *weak* listeners, the LVT correlated with LCT at .40, whereas the WVT correlated at .20 with the LCT (Table 4.3).

These results are in line with what previous research studies have shown with respect to the correlation between aural vocabulary knowledge and listening

comprehension. Bonk (2000) established a statistically positive correlation (Kendall's tau = .446) between lexical recognition – i.e., results in dictation tests – and listening comprehension as tested in recall protocols. Similarly, Mathews and Cheng (2015), showed a positive correlation (Pearson = .73) between aural vocabulary size – i.e., ability to recognize words in a dictation test – and listening comprehension – as expressed in the results in a standardized listening test (IELTS).

However, the present study has been the first one to test both aural and written vocabulary knowledge with the same target items and on the same population. One of the unique contributions of this study to the body of knowledge in the topic of L2 vocabulary and listening is that it has enabled the comparison of two measures of vocabulary knowledge (LVT and WVT) that *only* vary in the way the items are delivered to the participants. Consequently, the comparisons between the two tests with respect to their ability to predict the listening performance are more reliable. In the present study, it was unnecessary to account for differences between samples of participants or items in the tests, as they were the same in both tests. The main difference was in the manner they were delivered, i.e., in their aural (LVT) or their written form (WVT), as no order effects were found in the scores (Table 3.7; section 3.2.5.5).

A second result from the correlation analyses refers to the fact that the listening vocabulary test (LVT) and the written vocabulary test (WVT) are testing much of the same thing. Based on a correlation coefficient of .82 (October 2019) and .84 (June 2020) between the LVT and the WVT (Table 4.9), there exists collinearity between the two measures of vocabulary knowledge employed in the present study. However, as we will discuss in the next section, the overwhelmingly unique contribution of the LVT to explaining the variance in the listening

comprehension test presents a robust argument in favour of testing learners' aural vocabulary size instead of their written vocabulary knowledge.

Moreover, regression analyses showed that vocabulary knowledge can account for almost a third of the variance in the listening ability, and that aural vocabulary is better at explaining that variability than written vocabulary, particularly among weaker learners. Between 31.2% and 32.4% of the total variance in the results of a standardized listening comprehension test can be explained by the scores in two vocabulary tests, as it was reflected in the multiple regression analyses. Due to the small sample size in June 2020 (section 3.2.7.2), we might take 32.4% as a more realistic percentage for the total variance explained in a listening comprehension test by the results in two vocabulary tests. This percentage is in line with the results presented by previous research studies, although their figures range from 23% (Bonk, 2000) to 65% (Masrai, 2020).

Bonk (2000) determined that 23% of the variance in listening comprehension might be explained by the vocabulary size a learner has. Unlike this investigation, he equated the listening comprehension ability with the person's accuracy in recalling read-out passages of about 40 seconds in length, and about 84-86 words. Then, he separated the participants' scores in that listening recall task into two groups, based on the judgement of two independent raters. Finally, he associated the results of his study participants in a dictation task with their scores in the recall task to find possible correlations.

We have already discussed the inadequacy of using dictation exercises as research instruments to estimate learners' aural vocabulary size (sections 2.5.2.1 and 3.1.2). In a similar line of criticism, we should be cautious in the use of recall protocols to assess the ability to comprehend aural texts because they

might lack validity with respect to what language users encounter in real-life situations. Being able to comprehend aural input obviously implies remembering what the speaker has just said, so that we can analyse the utterances to *decipher* them, and then *build meaning* with the help of other parts of that aural discourse, and of our previous knowledge (Field, 2009; section 2.2.3.2). However, the ability to remember small excerpts of aural texts in the process of comprehension might be different from being able to recall details from a 40-second passage. In this case, there might be an unnecessary burden upon memory which differs from what language users usually experience in their everyday listening events. Consequently, using recall protocols to assess the listening ability might imply an excessive use of one's memory. In this respect, the methodology employed in the present study to assess the listening ability of its participants might be considered more ecologically valid, as they can process the aural input online, and answer the corresponding question immediately after it is heard.

Moreover, Bonk employed in the assessment of aural comprehension “four listening passages of increased lexical difficulty” (Bonk, 2000, 19), which had been created for that particular study. However, he failed to report whether those passages had undergone a process of quality assurance to make sure that they were representative (section 3.2.5.5). Therefore, claims based on the evidence collected with those instruments should be considered with caution.

Stæhr (2009) also assessed the relationship between learners' vocabulary size and their listening ability by means of regression analyses. He determined that up to 51% of the total variance in the results in a standardized listening test might be explained by learners' vocabulary size. Although his motivation was the lack of research into vocabulary and listening, he related the size and depth

of written vocabulary to the ability to comprehend aural texts, instead of using listening vocabulary tests to investigate that relationship (Cheng & Matthews, 2018; van Zeeland, 2017).

Nevertheless, two reasons might account for the higher figure in explained variance that Stæhr (2009) shows in his study when compared to the present investigation (51% vs 32.4%). Firstly, it had lower reliability indices in the three tests employed, particularly in the listening comprehension test (Cronbach's $\alpha = 0.60$). In the present study, the LCT showed an item reliability of 0.99 and 0.75, in the first and second datasets respectively (Table 3.32). Furthermore, the use of the Rasch Model in this research study enhances its reliability as the indices it presents are "more conservative and less misleading" than other reliability measures (Linacre, 1997, 581). Moreover, the low reliability index in Stæhr's study might be attributable to the use of a C2-level listening test (Cambridge English: Proficiency) to assess the listening ability among participants who were expected to have an overall linguistic level of B2. In fact, the mean score in the listening test in Stæhr's study was 66%, whereas the mean results in the vocabulary levels test was 85%, which might reflect a clear disparity in the difficulty of the tests. Additionally, the variability in the language abilities of the participants – from B2 to C2 – might have contributed to the low reliability index in a standardized listening test that was originally designed for a population with a narrower but higher range of language abilities.

A second reason that might account for the differences with Stæhr's study is that the overall language ability of the population in the present investigation was lower, because its target population was language students attending classes at B1-level. The amount of variance in a listening comprehension test that could be explained by the vocabulary dimensions is greater among better

performers in that test (section 4.1.1; Table 4.5). Apart from being in line with the results presented by Stæhr (2009), this evidence suggests that the relationship between vocabulary knowledge and listening comprehension is stronger among those who are more advanced in their language acquisition than among low-level learners.

Alternatively, it implies that at lower levels either there are more dimensions influencing their listening performance than at higher levels, or those dimensions are more influential. Unfortunately, the lack of reliability in the dataset gathered in June 2020 prevented the confirmation of this hypothesis. Furthermore, the sample gathered in October 2019 might also be too small to be divided in two groups of performers, as the number of participants who had 72% or more correct answers in the LCT was 48, whereas there were 234 participants in the other group. Further research is thus necessary to confirm the observed increase in the amount of variance in L2 listening comprehension explained by learners' vocabulary size.

Matthews and Cheng (2015) also attempted to explain the variance in a listening comprehension test by means of the results in a test of word recognition from speech (WRS). Words from frequency lists were used in partial dictation tests, and the participants' ability to recognize them was matched to their listening performance in a standardized test. Regression analyses showed that up to 54% of the variance in the listening results might be attributable to the ability to recognise words in their aural form. Unlike Stæhr (2009), Matthews and Cheng assessed vocabulary knowledge in its aural form, but they failed to use a more valid instrument than a dictation test (sections 2.5.2.1 and 3.1.2). Unlike what language users find in real-life listening events, the participants in that study were given written sentences where they had to write the target word.

The sentences were read out and the participants simply had to fill in the blanks (e.g., “The most _____ language is South Korean”; Matthews & Cheng, 2015, 10). There is a huge advantage in a word recognition test if the listener can anticipate when the target word is coming, and which *neighbours* it has. Furthermore, this kind of tests fail to assess the ability to recognize words and link them to a meaning, as participants are only told to write down the input they have perceived, without showing understanding of its meaning. Listeners with a minimal notion of the English phonology might be able to transcribe the words they have just heard, without having to demonstrate if they are able to link them to their correct meaning.

Not surprisingly, there is a clear disparity of results between the two tests: the Word Recognition Speech test had 71.71% correct answers, whereas the mean score in the listening comprehension tests represented 36.70% of the maximum possible score (Matthews & Cheng, 2015). These differences indicate that the dictation tests employed are clearly easier than the listening comprehension tests used in that investigation.

In a very recent study, Masrai (2020) discovered that both learners’ aural and written vocabulary size, with the help of their working memory capacity, can explain up to 65% of the variance in a standardized listening test. Furthermore, aural vocabulary size contributed the most to explaining that variability. These results are in line with the ones presented in this study, although with higher figures. The reasons for those differences in the figures might lie on the fact that Masrai used Yes-No tests to assess his participants’ aural and vocabulary size, which might have impacted negatively on the validity and subsequent reliability of the findings (section 2.5.2.1).

The present study does coincide with Stæhr (2009), Matthews and Cheng (2015) and Masrai (2020) in attributing most of the explained variance to one of the variables employed in the multiple regression. In Stæhr's study up to 49% of that variance was explained solely by the vocabulary size of a language user. The other 2% of the explained variance came from the results in the Word Associates Test (Read, 1993, 1998), which Stæhr employed in his operationalisation of the construct of vocabulary depth. Matthews and Cheng (2015) were able to attribute up to 52% of the total variance in listening comprehension to the ability to recognize words in their aural form from the lowest band of frequency employed in the test (3K). The other 2% of explained variance came from the ability to recognize words from the first band (1K). In Masrai's study learners' aural vocabulary size could account for 45% of the explained variance in the listening test, whereas their written vocabulary and their working memory capacity explained 6% and 14% of that variance, respectively.

In the dataset collected in October 2019, the contribution of both aural and written vocabulary size could explain 32.4% of the variance in the scores of the LCT. Similarly, in the second dataset (June 2020), the regression model including these independent variables could explain up to 31.2% of that variance. The importance of the listening vocabulary size in explaining the variability of results in the LCT compared to the scores in the WVT is shown when single linear regression models are employed. In October 2019, the LVT results could account on their own for 31.6% of the variance in the LCT, whereas the WVT scores on their own could only explain 16.5% of that variability. In the dataset from June 2020, the scores in the LVT could also explain more variance of the LCT than the

amount of variance explained by the WVT on its own (25.9% vs 9.18%). Moreover, the bigger contribution of aural vocabulary size in explaining the variability of results in a listening comprehension test is particularly relevant among lower-level learners. For this population, with the results of the LVT, 15.7% of the variance in the LCT is explained, whereas 3.9% is accounted for if the only element in the regression model is the WVT (Table 4.8). On the other hand, among those who had passed the LCT (i.e., with at least 72% of correct answers), when only the measures in the LVT are used in the regression, they can account for 27% of the variance, whereas the results in the WVT could explain 27.7%. In other words, among those who showed lower proficiency in the LCT, their measures in the LVT explain more variance in their LCT results than their measures in the WVT. Alternatively, among higher-level listeners, both vocabulary measures on their own – either the LVT or the WVT – are equally predictive of success in the LCT, as they can explain similar percentages of its variance (27% vs 27.7%).

In this section we have presented a straightforward answer to RQ1: language learners' vocabulary size and their listening comprehension ability are clearly related, and it can explain almost a third of all the variability in the results of a listening comprehension test. Furthermore, learners' aural vocabulary size is more related to their listening ability than their written vocabulary size, particularly among weak L2 listeners.

5.2 – RESEARCH QUESTION 2: *How much lexical coverage of a spoken text does a learner need to achieve comprehension in a listening test?*

The first set of analyses carried out with respect to RQ2 involved analysing both the words featured in the PET Vocabulary List, and in the transcript of the listening comprehension test (LCT). Having compared both texts, the conclusion was that all the words featured in the LCT were also in the vocabulary list. More importantly, this match implies equating the percentage of correct answers in the vocabulary tests with the lexical coverage in the LCT. If all the words in the listening test are featured in the vocabulary list, and all the items in that list had an equal chance to become a target word in the vocabulary tests, when a participant answers 60% of the questions correctly in the vocabulary tests, we might expect that they know 60% of the words featured in the LCT.

Based on the analyses discussed in section 4.3, knowing only 71.71% of the words featured in an aural text might be enough to achieve sufficient comprehension in a standardized listening test, and be considered a successful listener. Table 4.11 has shown that those who scored between 18 and 20 correct answers in the LCT (72-80% correct answers) recognised an average of 71.71% of the words in the LVT, or 79.05% of the items in the WVT. Although the sample of participants who had passed the LCT (i.e., $\geq 72\%$ correct answers) in the dataset gathered in October 2019 was relatively small ($N = 48$), a comparison of such results with previous research studies is necessary. Overall results are in line with what research has argued (Bonk, 2000; Stæhr, 2009; van Zeeland & Schmitt, 2013b): the higher the lexical coverage a person shows (in the present study, synonymous with scores in the LVT and the WVT), the higher their listening ability (scores in the LCT).

However, the figures of minimal lexical coverage needed for listening comprehension that presented by previous studies are clearly different. Bonk (2000) set the lexical coverage at 90% of the words featured in a text to achieve “good comprehension” (Bonk, 2000, 14). The reasons for the possible differences in the results between Bonk’s study and this investigation refer to the more thorough approach to designing and implementing the research tools adopted in the present study (section 5.1). Furthermore, the Rasch model used in the analyses on the different datasets should be considered more conservative than other data analysis methods. This, in turn, has a beneficial impact on the reliability of the results in the present study (Linacre, 1997). Moreover, Bonk (2000) failed to report the use of instruments like standardized frequency lists based on the analysis of broader corpora such as the BNC or the COCA. Consequently, other researchers have to be cautious when using his figures because they are only representative of the four texts he used in his investigation. Therefore, the range 71.71-79.05% of lexical coverage of a text to achieve listening comprehension levels of 72-80% seems a more reasonable and reliable figure than the values suggested by Bonk (2000).

Stæhr (2009) matched the frequency of the words in the transcript of a listening test with the scores obtained by his study participants in both that listening test and a Vocabulary Levels Test (Schmitt et al., 2001). For a text coverage of 94%, the mean listening comprehension scores were 60%; whereas people who knew 98% of the words in the transcript had a mean score of 73% (Stæhr, 2009). These percentages for lexical coverage are clearly above the levels determined for comprehension in the present study. For similar mean scores in a listening comprehension test (73% vs 72%), Stæhr (2009) recommended language learners to know 98% of the words featured in the transcript, whereas

the present study has shown that knowledge of only 71.71% of them might be sufficient.

Unlike Bonk (2000), Stæhr made use of a standardized listening test (Cambridge English: Proficiency) that had already undergone a process of quality assurance to preserve their validity and reliability (Lim & Khalifa, 2013). Additionally, the research instrument employed in his study for the assessment of the vocabulary size – the VLT developed by Schmitt et al. (2001) – had been subject to scrutiny, and has been considered the “closest thing to an accepted vocabulary test in English” (Stæhr, 2009, 587). Furthermore, unlike the vocabulary test used by Bonk (2000), the instrument employed in Stæhr’s study was based on frequency lists compiled from a sufficiently large corpus of the English language (i.e., the BNC). Therefore, the main criticism does not refer to its lack of generalizability – as it employed standardized, valid, reliable, and generalizable research instruments – but to the actual use of those research instruments. Stæhr (2009) used a standardized C2-level exam to assess the listening ability of L2 learners whose linguistic level ranged from B2 to C2. The listening comprehension test (LCT) used in the present study was based on a B1-level exam (Cambridge English: Preliminary), and was delivered to a group of students attending B1-classes. Furthermore, the use of research instruments based on sources within the same framework (i.e., the listening paper from Cambridge English: Preliminary, and the official Vocabulary List to prepare that examination) might have had an impact in setting those minimal lexical coverages. The original information sources were created for the same target population, i.e., candidates of a B1-level standardized exam (section 3.1.2). Consequently, the variability in the results from one research instrument to the other (LVT, WVT and LCT) could be minimized.

On the other hand, if a listening test meant for C2-level learners is used on a population whose linguistic level ranges from B2 to C2, the reliability of its results might be compromised, and they might correlate inadequately with other tests used on the same population. A piece of evidence to support this disparity in the tests results presented by Stæhr's study (2009) is that 50% of the participants showed a vocabulary size of no more than 3,000 words, whereas only 5.35% of its participants were able to master the highest band of frequency in the vocabulary test (i.e., the 10,000 most frequent words in English). Interestingly, knowing 98% of the words featured in that C2-level listening test would imply being able to recognize the 6,000 most frequent words in English.

Moreover, the assessment of the participants' vocabulary size was done through a written vocabulary test, although the results were meant to be matched to the performance in a listening test. In this respect, the claim made by researchers in applied linguistics that learners' ability to recognize words in their written and spoken forms might be different and should be tested separately (e.g., Cheng & Matthews, 2018; van Zeeland, 2017; Zhao & Ji, 2018) has been supported by several sets of evidence presented in Chapter 4 (e.g., Tables 4.3, 4.8, and 4.17). Furthermore, the present investigation has shown that aural vocabulary explains a greater percentage of variability in the results of a listening comprehension test than written vocabulary knowledge, particularly among lower-level listeners (section 5.1).

Van Zeeland & Schmitt (2013b) set the minimal coverage for adequate comprehension of a spoken test at 90% of all its words. Nevertheless, they recommended knowing 95% of the words to avoid variation in the comprehension levels. In order to determine the minimal coverage necessary for comprehension of a spoken text, they followed the same procedure Hu and

Nation (2000) had used in their study, and inserted non-words within listening passages to check lexical coverages at 90%, 95%, 98% and 100%. They also made sure that the passages only employed items from the 2,000 most frequent words in English. In their conclusions, they claimed that if “adequate comprehension” is set at 70% of correct answers in a listening test, the majority of listeners (75%) would achieve it if they knew 90% of the words featured in that test (van Zeeland & Schmitt, 2013b, 471).

At first sight, the figures proposed by the present study in terms of lexical coverage for adequate comprehension are considerably lower (71.70-79.05%), for a similar level of 72% of correct answers in a listening test. However, van Zeeland & Schmitt (2013b) used a listening text with words from the 2,000 most frequent words in English, and recommended knowing a minimum of 90% of the words featured in aural texts to understand them. In the present study, 80.31% of the words used in the vocabulary tests came from the first two bands of frequency (Table 3.5). A 90-per-cent coverage of 80.31% means that knowing 72.28% of all the words used in the tests would be sufficient for comprehension, very similar to the figures proposed in this section.

An answer to RQ2 has been presented in this section: a learner knowing no more than 71.71% of the words featured in a listening comprehension test is able to answer at least 72% of its questions correctly. This percentage is considered by some examining organizations as acceptable comprehension, because it implies passing the listening test in their standardized language exam (UCLES, 2019). Furthermore, that minimum percentage to comprehend aural texts is clearly lower than in previous studies, and it certainly depends on the accuracy and validity of the instruments employed in assessing both learners' vocabulary knowledge, their listening ability, and the frequency of the

words featured in the aural text.

5.3 – RESEARCH QUESTION 3: *How similar are the scores in vocabulary size tests based on recognising either the aural or the written form of words?*

The present study is the first attempt in the field of second language learning (SLL) to compare the impact of assessing learners' vocabulary size both orally and in its written form, with the same instruments, on the same population, but at two different moments in time (section 2.6). One unwanted effect of this novelty is the difficulty of contextualizing results in RQ3 with similar research from the past.

The analyses of the vocabulary test scores show that aural vocabulary tests are more difficult than written vocabulary tests, as reflected in a lower percentage of correct answers in the LVT compared to the WVT. It can be observed that language learners might find an aural vocabulary test up to 18.05% more difficult than a test with the same target words presented in their written form (Table 4.18). Moreover, the differences in the person measures between the LVT and the WVT reached the significance level in the three datasets ($p = .05$ or lower), and yielded medium to large effect sizes, with Cohen's d values between .34 and .94 (Table 4.21). The results are in line with what Masrai (2020) found out. His participants' mean aural vocabulary size was 2,688 words, whereas they showed a mean written vocabulary size of 4,334 words, i.e., 61.23% difference. Such different percentages – 18.05% vs 61.23% - might be due to the enhanced validity and reliability of the present study with respect to other research studies (section 2.5.2.1), which makes this investigation more conservative in its figures. Furthermore, Masrai's population consisted 130 participants with Arabic, Japanese, Chinese, Farsi and Brazilian as their first languages, unlike the population in the present study, whose L1 was Spanish.

Although more research is necessary to study the differences between learners'

aural and written vocabulary size, these data provide sound evidence to support the use of listening vocabulary tests to assess learners' aural vocabulary (Zhao & Ji, 2018). As written vocabulary tests might overestimate the aural vocabulary size by as much as 18.05%, aural vocabulary tests should be the preferred method, particularly if the ultimate purpose of the lexical assessment is to relate the results to their listening ability (Milton et al., 2010).

Secondly, the mean percentage of correct answers – both in the LVT and in the WVT – is lower in October 2019 than in May 2019, or in June 2020. One possible reason to account for the lower percentage of correct answers in October 2019 is that the data collection occurred at the beginning of the participants' academic year, unlike the datasets from May 2019 or June 2020. This difference in the time of collecting the data might also have an impact on the differences between the results in the LVT and the WVT, so that they are bigger at the beginning of the academic year.

Two possible reasons can account for the different results, depending on when the data is gathered. Firstly, that learning has occurred, so learners' vocabulary size – both aural and written – increases after attending classes for a period of about 35 weeks. The second explanation is that at the beginning of the academic year – i.e., when the learner is likely to show a lower linguistic proficiency – the disparities between aural and written vocabulary size are bigger. Therefore, aural and written vocabulary size might be considered as two different dimensions of vocabulary size that evolve differently. At first, learners might learn more about the written form of words, then they keep on expanding their written vocabulary size, but at a slower pace than their aural vocabulary size, until sizes are similar in both dimensions.

There were 18.05% more correct responses in the WVT than in the LVT in

October 2019, and only 9.63% more correct answers in the written vocabulary test in June 2020 with respect to the aural vocabulary test (Table 4.18). Furthermore, the differences between the scores in the LVT or in the WVT in those participants whose data were recorded both in October 2019 and June 2020 ($N = 17$) were all significant, but the effect sizes were slightly bigger at the beginning of the academic year than at the end (Table 4.28). Unfortunately, we need to view these figures with caution because the data collected in June 2020 yielded low reliability values due to the small sample size (Table 3.31). However, the differences between the scores in the LVT and the WVT in the same dataset (either in October 2019 or in June 2020) were significant, with slightly bigger – although small – effect sizes in the first dataset (Table 4.28).

An additional perspective to support the view that a person might learn the aural and the written form of words in a second language differently is provided by Table 4.22. It features the significance levels and effect sizes of the differences in participants' scores when the three dyads of tests are compared – LVT vs WVT, LVT vs LCT and WVT vs LCT – according to the participants' scores in the LCT. In the dataset collected in October 2019, the significance levels were reached for the differences in all the comparisons between tests, except for the comparisons between the WVT and the LCT among the top listeners, i.e., those who had at least 72% of correct answers in the listening comprehension test. In the case of the comparisons between the LVT and the WVT, the effect sizes of those differences in scores were higher among the weak listeners, which resulted in comparatively bigger effect sizes (Cohen's d 0.70 vs 0.80). The dataset collected in June 2020 indicated a similar trend, as all the comparisons of the percentages of correct answers between dyads of tests – LVT-WVT, LVT-LCT, WVT-LCT – showed significant differences, with small to medium

effect sizes (Table 4.28). The evidence drawn from both datasets, particularly from the data collected in October 2019, reinforce the view that testing the aural form of words is more challenging than presenting the same words in their written form, particularly among weaker learners.

Moreover, the claim that L2 learners might gain aural and written vocabulary at different paces can be supported by the data featured in Tables 4.12 and 4.13. They show different bands of coverage of the transcript of the LCT, depending on the results obtained in either the LVT (Table 4.12), or the WVT (Table 4.13). The other columns present the results obtained by the participants in the other two tests. For example, those learners who, according to their results in the LVT in October 2019, knew less than 50% of the words featured in the LCT transcript showed an average of 57.18% correct answers in the WVT, and 40.89% in the LCT (Table 4.12).

However interesting those descriptive statistics might seem, the relevant analyses are featured in Tables 4.14 and 4.15, where the differences across tests within the same band are examined in terms of significance and effect size. All the differences between the LVT and the WVT were significant, and yielded large effect sizes with Cohen's d in the range 1.14 to 2.66. The only comparison that failed to reach the level of significance was the comparison of the results the top scorers had in the LVT (i.e., $\geq 90\%$ correct answers; $N = 8$) with their results in the WVT. The absence of significant differences between those results in particular might indicate that learning the vocabulary meant for that language level – i.e., the items in the PET Vocabulary List – is almost complete for the top scorers in the LVT. However, the pattern is different when the results of the top scorers in the WVT ($N = 21$) are compared to their results in the LVT, as the differences are significant ($df = 20$, t -value = 5.97, p -value

<.0001), and yield a large effect size (Cohen's- $d = 2.14$). In this respect, we could assume that the significant differences when the top scorers in the WVT are considered might indicate that their learning of the written forms at B1-level is complete, but there are still words to be learned in their aural form. Table 4.12 shows that a person who knows an average of 90.90% of the words in the LVT has also learned the written form of 92.13% the words featured in the PET vocabulary list. Alternatively, a person showing a mean coverage of 93.30% of the words in the WVT only knows the aural form of 84.48% of the items in the vocabulary list. Consequently, they still need to learn the aural form of almost 16% more words (Table 4.13). Nevertheless, these explanations might be considered tentative, and further research is necessary to confirm that at later stages in the process of learning vocabulary, the differences between aural and written vocabulary size might be smaller than at earlier stages.

On the other hand, the rest of the data featured in Tables 4.12-4.14 clearly confirm the claim that aural and written vocabulary are two distinct dimensions in second language learning. The scores obtained by participants grouped in one band of performance in either the LVT or the WVT were significantly different from the scores those participants had in the other vocabulary test, and yielded large effect sizes. Furthermore, the variability in the levels of significance and the size of the effects depending on learners' language level might support the claim that learning vocabulary fails to follow a uniform path and rhythm, but evolves differently depending on the language level learners have. Again, a call for further research is necessary to investigate how vocabulary – both aural and written – is learned depending on the person's overall language level.

5.4 – RESEARCH QUESTION 4: *How does the relationship between lexical knowledge and listening performance evolve over time?*

The small sample of participants in June 2020 ($N = 17$) forces us to present all the evidence to answer RQ4 with extreme caution, and consider the possible findings as tentative, and indicative of trends that need confirmation (section 4.5). Nevertheless, the first set of evidence is in line with the results discussed in Section 5.1 with respect to the correlation figures between the vocabulary tests and the listening comprehension test (Bonk, 2000; Matthews & Cheng, 2015). In both datasets (October 2019 and June 2020), the LVT and the WVT showed the highest correlation values, followed by the LVT and the LCT, and with the written vocabulary test correlating with the LCT at the lowest (Table 4.23). Based on the correlation values presented by the LVT and the WVT, we might conclude that both vocabulary tests are testing much of the same thing, with Pearson product-moment coefficients of .82 in October 2019, and .84 in the dataset collected in June 2020.

On the other hand, the design employed in the present study enables the further discussion of those values. Perfect collinearity would mean correlation values of 1 between the two tests, so the gap shown by the actual values should account for any differences between the two tests. Since no order effects were found in the results (section 3.2.5.1), we could assume that the differential element is the manner the target words are presented and tested: either in their aural or in their written form. As those items are exactly the same in both tests, presented in the same order, at the same moment in time and – most importantly – delivered to the same participants, we might assume that aural and written vocabulary differ between .16 and .18, depending on whether the Pearson correlation values date back to October 2019 or June 2020.

Unfortunately, and to the best of my knowledge, no other research study has assessed the aural and written vocabulary size of second language learners at two different moments in time, preventing thus the possible comparison of these results with previous research.

The second set of evidence presented in section 4.5 shows that more variance in the results of the listening comprehension tests (LCT) was explained by the scores in the two vocabulary tests in October 2019 than in June 2020 (32.4% vs 31.2%). This difference in the explained variance might contradict the result presented in Section 5.1 for the data collected in October 2019. According to the multiple regression analysis performed on that dataset, more variance in the LCT was explained among higher-level learners than among less proficient ones: 31.5% of the variance in the LCT was explained for those who had more than 72% correct answers in that test (October 2019), whereas only 18.6% of the variance in the LCT results could be explained by the scores in the vocabulary tests obtained by participants with fewer than 72% of correct answers. It seems reasonable to think that learners attending classes for a period of about 35 weeks will be less proficient at the beginning of their academic years than at the end of their courses (June 2020). All the descriptive statistics confirm that the scores in the tests were higher in June 2020 than in October 2019 (see for example, Table 4.17). Although the participants in June 2020 showed a higher language level, their scores in the vocabulary tests could explain less variance in their listening comprehension results than at the beginning of their courses, with a lower overall language level. The most straightforward explanation for those mixed results is the lack of reliability in the data gathered in June 2020. Therefore, further research is necessary to determine whether the ability of learners' vocabulary size to predict results in

their listening tests is higher when they are less proficient in the language.

Nevertheless, these results confirm the higher importance of learners' aural vocabulary size in explaining their ability to understand spoken texts when compared to the actual contribution of their written vocabulary size. In two different datasets, collected from the same target population at either the beginning or the end of an academic year, the results in the LVT were able to predict more variance in the results of a listening comprehension assessment than the results from the WVT, particularly among weaker listeners. Furthermore, these results present a higher degree of reliability as the most relevant difference between the two tests was the actual manner in which the target words were delivered (Section 5.1). The comparison of the unique contribution of two dimensions in explaining the variance in the results of a third dimension, without having to draw on two separate sets of data gathered on different populations at different moments in time, provides another solid piece of evidence to advocate for the assessment of aural and written vocabulary separately (Cheng & Matthews, 2018; van Zeeland, 2017; Zhao & Ji, 2018). In any case, more research studies with a similar longitudinal design are necessary to confirm that the unique contribution of learners' aural vocabulary size when compared to their written vocabulary knowledge remains unaltered throughout time.

A final strand of findings refers to the differences in vocabulary scores across datasets (October 2019 vs June 2020), depending on the performance in a listening comprehension test. First, the study participants, regardless of their language level, clearly improve their results in both vocabulary tests within the 35 weeks of the observation period (Table 4.25). In particular, the scores in October 2019 and June 2020 obtained by the same participants ($N = 17$) in the

same tests at different points in time showed a moderate to strong positive correlation across datasets (Table 4.27). Furthermore, the comparisons of results obtained by those 17 participants in each test either in October 2019 or in June 2020 yielded significant differences, with large effect sizes ranging from 1.13 to 2.80. The effect sizes of those differences between one dataset and the other were comparatively higher among lower-level participants in all the tests (Table 4.30). This evidence supports the claim that lower-level language learners – classified thus according to their performance in a listening comprehension test – improved more than their classmates with a higher language level within an observation period of approximately 35 weeks. Although the analysis was performed on data from a limited sample ($N = 17$), and consequently each of the two bands of performance include few participants, its reliability might be higher than previous studies. This is the first study in the literature that attempts to compare results obtained by the same participants on both an aural and a written vocabulary test at two different points in time. Although we can be relatively confident in claiming that attending language classes for a period of about 35 weeks has a clearly beneficial effect, more research is necessary to confirm these findings.

5.5 – CONTEXTUALIZATION OF RESULTS

The previous sections in this chapter have contextualized the answers to each of the different research questions in the present study with respect to previous investigations into L2 vocabulary and listening. This section will present those findings within the broader context of L2 theory and research. Firstly, the study results are the empirical confirmation of the positive influence that learners' vocabulary might have on their ability to understand aural texts. The listening model proposed by Vandergrift and Goh (2012) highlighted the importance of the linguistic knowledge in perceiving, decoding and parsing the speech. The model included the lexical knowledge of the target language within the linguistic knowledge the listener may use to understand aural texts (Figure 2.1). Furthermore, the fact that almost a third of the results in a listening test might be predicted by the listeners' vocabulary size might confirm the validity of both the noticing hypothesis and the cognitive load theory (section 2.2.3.2). The more vocabulary language users have stored in their memories, the easier it is for them to perceive, parse and utilize the speech input, and the more orchestrated is the use of all resources at their disposal (Graham & Santos, 2015).

Secondly, the relatively higher ability of aural vocabulary tests to predict listening success corroborates the claims made by previous research (e.g., Milton et al., 2010), and confirms the higher importance of aural vocabulary when it comes to predicting listening success, so that it should be the "primary construct of relevance" (Matthews, 2018, 24).

Thirdly, the differences between aural and written vocabulary size align with Nation's taxonomy of different dimensions of what knowing a word implies (2001), and with his theory of learning burden (Nation, 2001, 2005). For the population under study – L1-Spanish speakers learning English at B1-level –

the learning burden of the aural form of words is comparatively heavier than that of their written form. Based on Nation's theory (2001, 2005), we could assume that English is so phonologically different from Spanish that it makes comparatively harder for L1-Spanish language learners to assimilate how words are pronounced – and aurally perceived – in English. On the other hand, when learning the written form of words might imply a heavier burden – like in Arabic – language learners might have comparatively better results in the aural version of vocabulary tests (Milton & Hopkins, 2006).

Moreover, the differences between aural and written vocabulary size are in line with what research has shown about acquiring L2 vocabulary dimensions: the same way that form recognition of new words is more easily acquired than their meaning recall (van Zeeland & Schmitt, 2013a), learners' written receptive vocabulary knowledge develops differently – and earlier – than the aural dimension of the same words (section 5.3). Interestingly, the differences between learners' aural and written vocabulary size were bigger at the beginning of their academic year (October 2019) than at the end (June 2020). One reason to account for this phenomenon might be that learning of both dimensions was complete for more items in the test in June 2020, so there were fewer words of which the learners had partial knowledge (section 5.3).

Nevertheless, the most relevant findings in the present investigation refer to the differences found in the results from either weak or strong listeners. Among better listeners, the two dimensions of vocabulary – aural and written – could account for more variance in the listening comprehension test than among weaker listeners: 31.5% vs 18.6%. Apart from being in line with previous empirical research (Stæhr, 2009), this result implies that for weaker listeners there are more dimensions influencing their listening performance. In other

words, those language learners draw more often on other factors when interpreting spoken input (Zhang & Graham, 2020), which was already suggested in previous research studies (e.g., Bonk, 2000; van Zeeland & Schmitt, 2013b).

On the other hand, the aural vocabulary size was considerably more predictive of listening success among language learners with lower listening comprehension results, whereas for more proficient listeners both vocabulary dimensions were equally predictive. We could conclude that aural and written vocabulary knowledge might be two clearly differentiated vocabulary dimensions at lower levels of proficiency that tend to merge into one single dimension – vocabulary knowledge – as the overall language proficiency develops. Within this evolution, in the early stages of language learning, learners' small aural vocabulary size is able to predict their listening success much better than their comparatively larger written vocabulary knowledge. The listeners' aptitudes in language learning or even their verbal and general intelligence, and how those aptitudes interact with each other might be responsible for the differences between weaker and stronger listeners (DeKeyser & Koeth, 2011). Furthermore, certain combinations of aptitudes – like working memory and phonemic coding for example – may be more relevant at some stages of the learning process than at others (Skehan, 2002).

However, the significant differences between weaker and stronger listeners with respect to the ability of their vocabulary size to predict listening success clearly differ from what previous research has claimed (Wang & Treffers-Daller, 2017). Consequently, further research is necessary into the possibly different impact of the listeners' aural and written vocabulary size on their listening success, as well as into the influence of their language and cognitive aptitudes on their

listening ability. Furthermore, most of the variance in a listening test still need be accounted for, as listeners' receptive vocabulary is able to explain a third of that variability. Among the possible variables that might bear an influence on the listening success, learners' overall linguistic proficiency seems to be a factor worth investigating (Wang & Treffers-Daller, 2017).

5.6 – CHAPTER SUMMARY

This chapter has addressed the different pieces of evidence presented in Chapter 4, in an attempt to present possible answers to the research questions that have guided the present study. The answers to each of the RQs can be summarized as follows:

- 1) Vocabulary and listening are related, in particular aural vocabulary, and especially among lower-level language learners.
- 2) Aural and written vocabulary are two similar but different dimensions. Language learners know more words in their written than in their aural form.
- 3) Language classes are effective to expand learners' vocabulary size and their listening ability, particularly among weaker students.

In the analyses of the relationship between vocabulary and listening (Research Question 1), and of the minimal lexical coverage necessary to understand aural texts (Research Question 2), previous research studies have been used to contextualize the findings of this investigation. However, many of the methodological decisions made in those studies have been challenged, in an attempt to account for the differences in the findings of the present study with respect to comparable past research.

On the other hand, the novelty in the approach to studying at the same time the aural and the written vocabulary knowledge of second language learners has prevented the present analysis from drawing on similar research studies to contextualize the findings with respect to the differences between aural and

written vocabulary size (Research Question 3). This is also the first study in the literature to use a longitudinal design to investigate the evolution of the relationship between vocabulary knowledge – aural and written – and listening comprehension over time. Consequently, no comparable research studies have been explicitly cited in the contextualization and discussion of the answers to RQ4, whose focus is on how the relationship between aural and written vocabulary with listening comprehension (Research Question 1) evolves over time. All the findings in the present study have been presented and contextualized within the limited existing research. In the final chapter of this dissertation, I will focus on the implications of those findings for the research community in applied linguistics, but most especially for L2 teachers and learners.

CHAPTER 6 – IMPLICATIONS, CONCLUSIONS and LIMITATIONS

A great deal of the findings in this dissertation are due to the novel approach employed in investigating second language vocabulary and listening (section 3.1.2). To the best of my knowledge, this is the first time that:

- a) aural and written vocabulary is tested with the same target words, on the same population, and at the same moment in time,
- b) the assessment of both vocabulary size and listening comprehension is carried out on the same population at two moments in time, separated by an observation period,
- c) the instruments for the assessment of the dimensions under study – aural vocabulary size, written vocabulary size, and listening comprehension – are created or adapted within the same framework, i.e., the standardized language test Cambridge English: Preliminary.

Moreover, this dissertation adds to the very few studies in applied linguistics that have used the Rasch model for the design and validation of the tests employed in their data collection (Fan & Knoch, 2019). This model is particularly conservative and thorough in its judgements about the data, increasing thus the confidence in the accurate reflection in the results of the phenomena under investigation (section 3.1.2.5).

Chapter 6 intends to address the implications of the findings in the present research project for researchers, teachers, and learners. Each of those findings will be contextualized now by analysing their impact on L2 research and classrooms. But first, the limitations that the present study has shown will be addressed to help other researchers minimize them in future investigations, as well to facilitate the contextualization of the study findings.

6.1 – LIMITATIONS OF THE STUDY

The main limitation in this study has derived from the COVID-19 pandemic. The data collection in June 2020 had to be made with an online version of the three research instruments, instead of the paper-based delivery carried out in the first data gathering (October 2019). Furthermore, the attrition rate from one dataset to the other was particularly high, as the number of participants dropped from 284 to 17, which has reduced the reliability of the data collected in June 2020 (section 3.2.7.2).

A second limitation in the study has been the absence of controls for order effects in the data collection. The need for the two vocabulary tests to deliver the same target items to the same population, first aurally and then in their written form, prevented the use of a crossover repeated measures design. A counterbalanced delivery of the test items might have enhanced the study quality. However, the complexity of creating several versions of the same tests, and delivering it to up to 450 students enrolled in 18 different groups (section 3.2.5.1), as well as the need for controlling the participants' exposure to the stimuli, particularly if the tests were delivered online, led to ruling out such study design. On the other hand, although no order effects were found (Table 3.7), the analyses performed on the datasets were unable to rule out the hypothesis that unbiased estimates might have been obtained because of the absence of a counterbalanced design (Pollatsek & Well, 1995). Future research studies should operationalise the construct of lexical knowledge in a manner that includes a counterbalanced delivery of the test items (section 3.1.2.4). Furthermore, they should include counterbalancing in the subsequent analyses of the within-subject variables under study (Pollatsek & Well, 1995).

The specificity of the inclusion criteria (Table 3.1) might have been another

limitation, as the claims derived from the findings should apply primarily to adult B1-level learners of English, attending language courses, and whose first language is Spanish. Furthermore, the claims about the relationship between vocabulary and listening comprehension should actually be interpreted as the relationship between the scores in two very specific vocabulary tests, and the scores in the listening paper in a standardized language test (section 3.1.2.4). In studies like the present investigation, where most findings are based on test scores, readers should be fully aware of the limitations that the research constructs might pose before accepting the study results and, most importantly, before attempting their application to other settings and circumstances.

Alternatively, “the more reliable the sample of performance or test score is, the more generalizable it is” (Bachman, 1990, 187-188). Within this perspective, the validity of the constructs used in an investigation, and how they are operationalized might contribute to enhance the study reliability, and lead to a higher generalizability of results. In this respect, researchers should assess how relevant the research instruments and the sample are for the population under study, determine how sensitive those instruments are to apprehend the studied phenomena, rule out whether the instruments are neglecting dimensions or variables that might mask the results, and make sure that data quality is preserved despite the abnormal behaviour of outliers (Messick, 1995). Section 3.2.5.5 has highlighted how the specificity of the research instruments and the sampling criteria employed in this investigation led to an enhanced reliability of the results, and a heightened confidence in the claims and implications derived from them. In any case, the limitation of the generalizability of results and claims could be surmounted by carrying out further research, particularly in the form of replication studies with similar designs, but with slightly different inclusion

criteria for the sampling process (section 3.1.2.1).

6.2 – IMPLICATIONS FOR SECOND LANGUAGE RESEARCH

METHODOLOGY

Chapter 5 has presented previous research studies, and drawn on hypotheses and theories to contextualize the findings of this dissertation. This section focuses now on the implications of those findings for future methodological approaches to investigating second language, particularly in empirical studies.

A clear implication derives from the exploration of a few research articles on L2 vocabulary and listening, as well as from the reflection on the findings of the present study. From a methodological point of view, changes should be made in studies that attempt to determine the minimal level of learners' vocabulary size to comprehend a variety of texts. In the contextualization of the answers to the Research Questions (Chapter 5) the findings of previous studies were used in the comparisons with the present study (e.g., Bonk, 2000; Stæhr, 2009; van Zeeland & Schmitt, 2013b). Unfortunately, it was impossible to carry out those comparative analyses because the RQs were approached quite differently from an epistemological point of view. Previous research could have profited from employing more suitable instruments (Bonk, 2000; section 2.5.2.1 and 3.1.2), using the research tools more adequately (Stæhr, 2009; section 5.1), and basing research on more accurate assumptions (van Zeeland & Schmitt, 2013b; section 5.2).

Several consequences for future investigations could be drawn from the critical analysis of the methodologies used in previous research. Firstly, data and results from studies should be accepted while bearing in mind their limitations, particularly when their generalizability might be improved (Bonk, 2000; Martinez & Schmitt, 2012). Furthermore, a balance should be found between scope and sensitivity in the research instruments employed in L2 studies. In other words,

research should use instruments that are broad enough to study as large a population as possible, but without compromising how accurate those instruments are (Stæhr, 2009). In particular, standardized criterion-based tests like the Cambridge English: Proficiency should be used on populations that are within the narrow range of proficiency originally meant to be assessed by those examinations. In this respect, the present study employed a B1-level listening test – Cambridge English: Preliminary – on a target population of B1-students. Additionally, the target items in the vocabulary tests were selected from a vocabulary list published by the institution in charge of the listening test, and meant to help its B1-test candidates prepare that examination (UCLES, 2012).

The use of replication studies might be a solution to keep the right balance between sensitivity and scope (sections 2.5.2.2 and 3.1.2). Replicating a previous study with slight modifications enables the researcher to keep the focus on the phenomenon under study, and to use instruments that are sensitive enough when applied to that specific population. The broader picture might be achieved through the addition of numerous similar studies, with comparable populations or methodologies, so that an enhanced level of generalizability is achieved by the sum of their components. For example, the present study should be replicated on language students with a language proficiency different from B1, or whose L1 is other than Spanish, or who are learning the language in different settings. Replication is crucial in the promotion of transparency and collaboration in research. Unfortunately, its scarcity in L2 research has become a serious problem (Abbuhl & Mackey, 2017).

Another consequence derived from the critical analysis of previous methodologies is that researchers should use listening vocabulary tests to assess the aural vocabulary size, particularly when the results are matched to

the same learner's performance in a listening comprehension test. There are aural versions of vocabulary tests available (e.g., Karami, 2012; McLean et al., 2015; Nguyen & Nation, 2011; Zhao & Ji, 2018) that other researchers can employ in their investigations. Furthermore, the creation of a listening vocabulary test is a perfectly acceptable challenge to any researcher, who will have the help of current technology and the experience shared by other colleagues in several reports (section 3.2.3). In fact, designing and developing a vocabulary test "has probably never been easier" (Schmitt, Nation & Kremmel, 2020, 109).

A third implication refers to partially accurate assumptions with respect to the minimal lexical coverage that is necessary "for adequate listening comprehension" (van Zeeland & Schmitt, 2013b, 457). Research in applied linguistics should be careful in the assumptions it makes, particularly when they might have an immediate impact on the classrooms. The descriptors in the Common European Framework of Reference state that a language user at B1-level can "communicate essential points and ideas in familiar contexts" (Cambridge University Press, 2013). For such language users, the assumption of 2,000 words as a sufficient vocabulary size to function adequately in spoken English might be true (section 5.2). However, the CEFR also encompasses three higher language levels where users should be able to communicate "effectively, with some fluency, in a range of contexts" (B2), "fluently and flexibly in a wide range of contexts" (C1), or "very fluently, precisely and sensitively in most contexts" (C2). When assessing how much vocabulary those language users might need to function adequately, more contexts have to be taken into account. The wider the variety of the situations to be included in the assessment, the more words to be considered, and the lower their overall

frequency (section 2.5). Therefore, for the highest levels of proficiency in a language, as described in the CEFR, a clearly bigger vocabulary size is necessary for competent functioning (section 5.2; English Profile, 2015-2020).

A final aspect to be considered by methodological approaches to investigating L2 teaching in general, and the relationship between vocabulary and listening in particular, is the inclusion of longitudinal designs. Based on the experience provided by the present dissertation, the plausibility of results is enhanced when the same population is investigated through the same research instruments before and after an observation period. Furthermore, as study participants may have had different experiences of classroom-based language instruction, applying the same research methods on the same population might help to compensate for the impact of those differences on the results.

This section has addressed the broader implications of the present research study for the methodology employed to investigate L2 vocabulary and listening. Some of the gaps that have become apparent after the literature review (Section 2.6), as well as those detected while carrying out this study will be the focus of the following section.

6.3 – IMPLICATIONS FOR SECOND LANGUAGE THEORY AND RESEARCH

The literature review presented in Chapter 2 has addressed the current situation of second language listening comprehension in the language classrooms, and its impact on learners' levels of confidence and self-efficacy. Furthermore, it has also highlighted the importance of L2 vocabulary knowledge as a factor to explain the success in listening comprehension. However, the amount of research into listening is smaller when compared to other language skills. This section focuses on how the present study might influence theoretical perspectives and future research on L2 vocabulary knowledge and listening comprehension, so that some of the gaps found in the literature can be bridged.

Firstly, L2 researchers have in vocabulary a reliable factor to predict listening comprehension success among language learners. The clear and positive correlation between a learner's vocabulary size and their ability to comprehend aural texts support this claim. Furthermore, if nearly a third of the success in comprehending aural texts is due to learners' vocabulary size, the approach to investigating listening comprehension should take it into account. Moreover, the skill of listening causes anxiety and frustration among language learners (Graham, 2011), and it is considered the least researched of all (Vandergrift, 2007). Implementing investigations into the relationship between vocabulary and listening will help to reduce this lack of research into the skill, and have a clear and positive impact on the manner it is taught in the language classrooms.

Secondly, the observed trend that L2 vocabulary and listening are less related among lower-level learners implies that there should be more dimensions to account for the variability of performance in listening tests. Therefore, future research should focus on unearthing the impact of other factors on learners' listening performance – particularly among weaker learners – like memory, or

the ability to focus (Rubin, 1994).

Thirdly, there should be a clear preference among researchers for listening vocabulary tests to assess learners' aural vocabulary size, particularly if their results are to be related to the ability to understand aural texts (Milton et al., 2010). The higher correlation of listening scores with aural than with written vocabulary scores, and the clearly greater ability of a listening vocabulary test to explain variance in a listening comprehension test are solid arguments to support this choice. Additionally, this result provides robust evidence for the claim made by other researchers advocating for a separate assessment of aural and written vocabulary size (Cheng & Matthews, 2018; van Zeeland, 2017; Zhao & Ji, 2018).

Moreover, as the present study is the first one in the literature to assess both dimensions of vocabulary size using the same target items on the same population, this should be the first confirmation that both dimensions – aural and written vocabulary size – lie on common grounds. The collinearity between the two tests has an important implication for language testing: if both tests tap into similar realms, but the listening vocabulary test shows a closer and more important relationship with the listening ability, it should be the preferred standard to assess learners' vocabulary size when listening comprehension is being investigated. Another argument in favour of using listening vocabulary tests to assess the aural vocabulary size comes from the differences in difficulty detected between the LVT and the WVT, because using a written vocabulary test might overestimate the aural vocabulary size of language learners by up to 18.05%.

Nevertheless, the present investigation has focused only on the relationship of receptive aural and written vocabulary on the performance in listening

comprehension tests. Further research should include the use of a reading comprehension test in the comparison, to determine how aural and written vocabulary size tests correlate with either listening or reading comprehension, so that researchers are able to decide what type of vocabulary test is preferable in each case. Furthermore, more studies are necessary to determine the relationship of productive vocabulary size with listening comprehension, or what kind of vocabulary – aural or written, productive or receptive – is more related to productive oral skills like speaking.

A fourth implication for future research derives from the fact that the present research study has demonstrated that aural and written vocabulary size are two distinct dimensions, albeit lying on common grounds. Future research studies should aim at confirming these differences by means of similar methodologies on different populations, with other language levels, or with first languages different from Spanish. In particular, more studies – probably with a longitudinal design like in the present investigation – are necessary to confirm that at earlier stages of vocabulary learning, the differences between language learners' aural and written vocabulary size are bigger than at later stages of the process.

Moreover, researchers should carry out studies to explore the possible reasons that might account for the differences in learners' aural and written vocabulary. They might want to find out whether the differences appear because the teaching of aural vocabulary is neglected both in the classrooms and in the published materials to learn the language, or maybe because the aural form of words are intrinsically more difficult to learn. The analysis of how L2 vocabulary is actually taught, both in the classrooms and in the didactic materials available in the market, was beyond the scope of the present research study. Consequently, future studies should investigate the actual teaching of

vocabulary in L2 classrooms and materials.

Another implication for future research derives from the confirmation of the positive impact of attending L2 classes for a period of about 35 weeks in expanding the aural and written vocabulary size, and in improving the performance in a listening comprehension test. Unfortunately, it was beyond the scope of this study to investigate the relative efficacy of having language lessons in a classroom compared to other approaches to learning a language such as self-studying, or simply living in a place where the target language is spoken. Future research should look into the possible differences between those ways of gaining vocabulary and becoming a proficient listener in a second language. In particular, cohort longitudinal studies might be indicated to determine at what levels of language proficiency one approach is more suitable than the other.

A final set of findings in this investigation refer to the lexical coverage necessary to understand aural texts adequately. Presenting one figure for minimal lexical coverage or another is paramount for subsequent research. For example, if language learners have to know 95% of the 2,000 most frequent word families in English to achieve “adequate listening comprehension” (van Zeeland & Schmitt, 2013b, p 457), they should be able to recognize the meaning and form of approximately 12,563 different words at Level 6 (Bauer & Nation, 1983; Nation, 2017). If the recommendation is to know 98% of the 7,000 most frequent word families (Nation, 2006) the numbers would increase to more than 31,000 words. Future investigations should be cautious when assuming the percentages proposed by previous research, and relate those figures to the limitations of the studies in terms of generalizability (section 6.2). The present study claims that knowing 71.71% of the words featured in a listening

comprehension test is enough to achieve 72% of correct answers in that test. However, that finding might only be applied with some degree of confidence to similar populations of L2-English learners attending classes at a B1-level, and whose first language is Spanish. Replication studies are thus necessary to confirm this finding, and extend its applicability to other populations with different L1s, language levels, or instructional circumstances.

Moreover, researchers should be aware of the differences that using one vocabulary test or the other might have. The minimal lexical coverage in this investigation was 71.71% or 79.05%, depending on whether the vocabulary size was assessed with the LVT or the WVT. Further research is necessary to increase the accuracy of the estimations of lexical coverage required to achieve adequate performance in different language skills. In this respect, this research has demonstrated the particular sensitivity of listening vocabulary tests in studies about vocabulary size and listening.

Researchers should also be aware of the representativeness of the texts – aural or written – employed for the lexical profiling, and the scope of situations contemplated. For example, all the words in the listening comprehension test employed in this investigation were featured in the PET Vocabulary List, but more studies are necessary to confirm that the words in other examples of those language tests are also in the compilation. This line of research might also be seen as a good manner to audit some of the claims made by the organization in charge of such tests (i.e., Cambridge Assessment English). When research employs vocabulary tests and text profiling instruments based on the most frequent words in a language, researchers should make sure that the sample of texts selected and analysed, as well as the vocabulary bands employed in the assessment are relevant enough to encompass most of the

situations that the target population usually deal with (section 6.1).

Providing those involved in L2 teaching and learning – including publishers and materials writers – with a reliable figure based on sound research methodologies is essential. Ideally, this minimal percentage of words a language learner has to know to achieve comprehension of a text will eventually guide matters as varied as the writing and publication of teaching materials, the planning of courses, the design of novel approaches to teaching vocabulary, the intended pacing of vocabulary lessons, or something as simple as the allocation of time to study vocabulary.

The implications of the study findings for future research have been discussed in an attempt to bridge gaps that have become apparent both in the literature review and in the implementation of the present study. The ultimate goal of this dissertation was to provide second language learners with better methodologies to help them become better listeners. The following section will address the possible implications for L2 learners and teachers, as well as for designers and publishers of learning materials.

6.4 – IMPLICATIONS FOR SECOND LANGUAGE TEACHING

The fact that language learners' vocabulary size and their listening comprehension skills are clearly related should have a clear impact on the language classrooms. Teachers and publishers should explore this avenue in their teaching methodologies, and try to implement programs where learners' acquisition and consolidation of vocabulary play an important role. By doing so, they will both be making use of research findings – like the ones presented in this dissertation – to inform their decisions, as well as having a positive impact on language learners.

Furthermore, the fact that aural and written vocabulary are two separate dimensions, and that the aural vocabulary knowledge is comparatively more related to listening comprehension, particularly among weaker listeners, might have two main implications. More emphasis should be put into teaching other aspects of *knowing* a word rather than the link between its written form and its meaning (Nation, 2001; Webb, 2002). Investing time in teaching L2-students the aural form of words is certainly an advisable way to teach vocabulary, particularly if the aim is to help them be better listeners. Regardless of the reasons why L2 learners have a smaller aural vocabulary size, teachers and publishers should include a clear focus on the teaching of the aural form of words, particularly in connected speech (section 2.3.1). Furthermore, the relevance of teaching such aspects of L2 vocabulary is particularly high among lower-level learners, as the differences between their aural and written vocabulary are bigger, more significant and with larger effect sizes.

Similarly, publishers should include the teaching of the aural form of words in the materials they offer. Many books used in the language classrooms like *Straightforward* (Scrivener & Jones, 2012), or *Empower* (Doff et al., 2015)

provide glossaries, or vocabulary lists. However, those compilations of words rarely include anything apart from a definition or a translation, and a phonological transcription of the word. With the widespread use of multimedia approaches to teaching languages, it might be more than reasonable to include recordings where students can be exposed to the aural form of those words. Furthermore, the higher sensitivity of the aural vocabulary size to predict listening comprehension supports the inclusion within those textbooks of more exercises and activities where the focus lies on learning the aural form of words, and in recognizing them both in their citation forms, and in connected speech (section 2.3.1).

The results of the lexical profiling of both the PET Vocabulary List and the transcript of the PET listening paper also have implications for the language classrooms. Firstly, the use of that official vocabulary list might inform the creation of specific vocabulary activities for those preparing themselves to certify their level of English with the exam Cambridge English: Preliminary. Furthermore, teachers and publishers might use the list as a reference to make sure that the lexis used in the exercises employed in those PET preparation courses is adequate for this standardized test. Additionally, international language testing institutions like Cambridge Assessment English should provide both their exam candidates and writers with similar vocabulary lists for higher levels of language proficiency such as B2 or C1.

Another set of implications for language classrooms derives from the findings about the minimal lexical coverage necessary to achieve listening comprehension. Knowing between 71.71% and 79.05% of the words featured in a listening comprehension test might be enough to answer 72% of its questions correctly. The difference lies on what aspect of *knowing a word* is assessed –

either recognizing its aural or its written form – and provides further evidence to support the teaching of the aural form of words in the language classroom. Some learners complain about being unable to understand most of the input in a listening task, even if they can later recognize and understand the same words in the corresponding transcript of the recording (Cai & Lee, 2010; Goh, 2005; van Zeeland, 2014b). Those learners might think that knowing a word is just being able to match it with what it means and recognizing its written form (section 2.3.1), ruling out the importance of knowing how the word is pronounced, or acoustically perceived (Nation, 2001). Apart from planning the time investment to gain the necessary lexical knowledge to achieve listening comprehension, teachers – as well as designers and publishers of language teaching materials – should stop assuming that by being able to recognize the written form of a word, a learner can also do the same with its aural form.

One final conclusion that might be drawn from the findings in the present study is that the language learners who have participated in the study should feel reassured about the efficacy of their approach to learning the language. Furthermore, attending a B1-level course is more efficient for the weak students in the class. Similarly, teachers should remain confident that the benefits of the methodologies they employ will eventually become apparent, particularly among those students that might seem *lost*, or not making the most of the lessons, at first.

6.5 – IMPACT ON MY TEACHING PRACTICE

This dissertation is the culmination of an EdD journey that started five years ago. This section might be seen as supplementary to the previous one, where the implications for second language teaching were discussed. I will present some examples of how the EdD journey in general – and the findings of this investigation in particular – have impacted the way I teach my English-L2 students.

First and foremost, I have included a clear emphasis on vocabulary teaching as a way to improve my students' listening skills. In particular, I tend to follow the criteria of frequency in the explicit teaching of lexical terms, and consequently, I tell my students that one word, phrase or expression is of particular interest because it is frequently used in English. Although I always answer all my students' questions about unknown words, I try to make it clear in my students' mind that some words are more *important* than others because they are more frequent in general, or they are more relevant – and frequent – for a particular field of expertise like medicine or management, for example.

The emphasis on vocabulary teaching also implies writing down all unknown words on the whiteboard so that students can notice how the word is not only pronounced, but also written, which in turn might lead them to the reinforcement of the form-meaning link on both modalities of vocabulary knowledge (Nation, 2001). As my students' aural vocabulary size is usually smaller than their written vocabulary breadth, I emphasize the learning of the aural form of words. For that purpose, all my classes have access to Quizlet – a vocabulary learning app – where they are exposed to and practise with both the aural and written form of the words they have encountered in our lessons. Furthermore, when I prepare recycling vocabulary exercises to review previously seen vocabulary, or to

revise for an exam, I always include practice with the aural form of words. Additionally, when some listening activities have been completed, I prepare gap-filling exercises based on the transcripts, where my students have to write a few missing words they have just heard in connected speech. A final step in reinforcing the form-meaning link of words in my students' minds is to include integrated tasks where they have to listen to short sentences, transcribe them, and then record their voices while pronouncing those sentences back.

Another piece of evidence to show the impact of vocabulary and listening in my classes is that in all the courses I teach, their exams and tests include specific vocabulary exercises based on the lessons and materials covered in the classes. Generally speaking, the weight of those vocabulary exercises in the course final grade is the same as with listening, reading, speaking and writing – i.e., 20%.

Regarding the actual teaching of listening comprehension, a few changes have been introduced in my classes because of the evidence found in the literature. Firstly, more time is allocated to listening tasks in my classes than before. I tend to do the activities with my students, insisting on making them anticipate the type of input they are going to be exposed to – for example, a conversation, a monologue, an interview – and, most importantly, what the task is expecting them to do. I quite often disregard what the book says about how to proceed with the recording and the comprehension questions, and play the recording more than twice, or stop it after a relevant piece of information comes out, or repeat one specific excerpt over and over again. Furthermore, in an attempt to avoid that my students just focus on having the *right* answer, I usually tell them that they should not aim at 100% comprehension of the input, but at sufficient comprehension to complete the task in a satisfactory way. They should behave

in the same way as in real life in their L1s, when they can perfectly function in oral interactions even if they have a certain percentage of uncertainty about the exchange because they were not paying attention to some excerpts, or because the speaker was too ambiguous, for example.

Finally, the importance gained by listening in my current teaching of L2 English is evident in how all my students are given every week extra listening practice to be done outside our classrooms. Furthermore, the emphasis on listening is also apparent when I have more room for manoeuvre when deciding the specific weight of the course components and the language skills. Every time I have had the opportunity to increase the time allocated to listening, I have taken it while reducing the attention given to other aspects of language like reading or writing.

6.6 – CHAPTER SUMMARY

This chapter has discussed the implications to be drawn from the results presented in this doctoral dissertation. First of all, a listening vocabulary test is a better predictor of listening success than previous forms of written vocabulary tests, like the VLT (Schmitt et al., 2001) or the VST (Beglar & Nation, 2007). Therefore, results from previous studies which have related written vocabulary size to listening performance should be considered differently (Stæhr, 2009; van Zeeland & Schmitt, 2013b). Furthermore, as listening vocabulary tests and written vocabulary tests might tap into similar dimensions, but LVTs are better predictors of listening performance, they should be considered the preferred standard in studies relating vocabulary and listening comprehension.

Secondly, language practitioners should be extremely cautious in accepting the lexical coverage figures that previous research has proposed to understand aural texts. Those studies might have not chosen the best methodological approaches to quantifying vocabulary sizes, and relating them to listening comprehension. Therefore, more research studies with adequate methodologies are necessary to present answers to the question of how many words a language learner should know to achieve comprehension.

Moreover, the present study has confirmed that two separate dimensions of vocabulary exist: aural and written. The implications for language classrooms, designers and publishers of second language materials, and investigators on L2 teaching and learning are evident. More attention should be paid to aural vocabulary as a separate dimension, in particular among weaker students. Furthermore, the clear differences between the aural and written vocabulary shown by language learners should force us to rule out the idea that knowing the written form of a word is enough to truly know that word. Its aural form might

be as important, particularly in oral communication.

A final conclusion is that employing a longitudinal design has enabled the possibility of confirming with new evidence trends observed in previous datasets, leading to the enhancement of the reliability in the results, and in the soundness of the conclusions. The implication for research practice is clear: more L2 research studies should benefit from using longitudinal approaches in their design. This need for longitudinal designs is particularly apparent when it comes to investigating the relationship between L2 vocabulary and listening, as the present study has demonstrated.

BIBLIOGRAPHY AND REFERENCES

- Abbuhi, R., & Mackey, A. (2017). Second language acquisition research methods. In King, K. A., Lai, Y. J., & May, S. (Eds.). *Research methods in language and education* (3rd edition). (pp. 183-193).
<https://doi.org/10.1007/978-3-319-02249-9>
- Adolphs, S. and Schmitt, N. (2003). Lexical Coverage of Spoken Discourse. *Applied Linguistics* 24(4): 425-438. <https://doi.org/10.1093/applin/24.4.425>
- Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice*, 18(3), 259-278. <https://doi.org/10.1080/0969594X.2010.546775>
- Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. New York: Continuum.
- Anderson, J. (2020). *Cognitive psychology and its implications*. (9th edition). New York, NY: Worth Publishers.
- Andrich, D., & Marais, I. (2019). *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*. Singapore: Springer.
- Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). *Language Learning*, 62(Suppl. 2), 49–78. <https://doi.org/10.1111/j.1467-9922.2012.00706.x>
- Arnold, J. (2000). Seeing Through Listening Comprehension Exam Anxiety. *TESOL Quarterly*, 34(4), 777–786. <https://doi.org/10.2307/3587791>
- Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, 38(1), 6-40.
<https://doi.org/10.1177/0265532220927487>
- Axelrod, R. H., Axelrod, E., Jacobs, R. W., & Beedon, J. (2006). Beat the odds and succeed in organizational change. *Consulting to Management*, 17(2), 1-4.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

- Baghaei, P., & Amrahi, N. (2011). Validation of a Multiple Choice English Vocabulary Test with the Rasch Model. *Journal of Language Teaching & Research*, 2(5), 1052-1060. <https://doi:10.4304/jltr.2.5.1052-1060>
- Baïdak, N., Balcon, M. P., & Motiejunaite, A. (2017). Key Data on Teaching Languages at School in Europe. 2017 Edition. Eurydice Report. *Education, Audiovisual and Culture Executive Agency, European Commission*. Luxembourg: Publications Office.
- Bandura, A. (1997). *Self-efficacy : the exercise of control*. New York, NY: W.H. Freeman.
- Bauer, L., & Nation, P. (1993). Word Families. *International Journal of Lexicography*, 6(4), 253–279. <https://doi.org/10.1093/ijl/6.4.253>
- Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language testing*, 27(1), 101-118. <https://doi.org/10.1177/0265532209340194>
- Beglar, D., & Nation, P. (2007). A vocabulary size test. *The Language Teacher*, 31(7), 9-13.
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model : fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.
- Bonk, W. (2000). Second Language Lexical Knowledge and Listening Comprehension. *International Journal of Listening*, 14(1), 14–31. <https://doi.org/10.1080/10904018.2000.10499033>
- Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences*, Dordrecht, Netherlands: Springer.
- Boyle, J. (1984). Factors affecting listening comprehension. *ELT Journal*, 38(1), 34–38. <https://doi.org/10.1093/elt/38.1.34>
- Bramley, T. (2008, September). Mark scheme features associated with different levels of marker agreement. In *British Educational Research Association conference* (pp. 3-6). <https://www.cambridgeassessment.org.uk/Images/474559-mark-scheme-features-associated-with-different-levels-of-marker-agreement.pdf>
- Brentari, E., & Golia, S. (2007). Unidimensionality in the Rasch model: how to detect and interpret. *Statistica*, 67(3), 253-261. <https://doi.org/10.6092/issn.1973-2201/3508>

- Brown, J. D. (1997). Designing a Language Study. In Griffee, D. and Nunan, D., (Eds) *Classroom Teachers and Classroom Research*, (pp. 55-70). Tokyo: Japan Association for Language Teaching.
- Brown, S. (2006). *Teaching listening* (Vol. 5, No. 1, pp. 36-39). New York: Cambridge University Press.
- Brunfaut, T., & Révész, A. (2015). The Role of Task and Listener Characteristics in Second Language Listening. *TESOL Quarterly*, 49(1), 141–168. <https://doi.org/10.1002/tesq.168>
- Buck, G. (2001). *Assessing Listening*. Cambridge: CUP.
- Cai, W., & Lee, B. P. (2010). Investigating the effect of contextual clues on the processing of unfamiliar words in second language listening comprehension. *Australian Review of Applied Linguistics*, 33(2), 18.1-18.28. <https://doi.org/10.2104/aral1018>
- Cambridge University Press (2008). *Cambridge English : preliminary 5 with answers : official examination papers from University of Cambridge ESOL examinations*. (2008). Cambridge: Cambridge University Press.
- Cambridge University Press (2013). *An introductory guide to the Common European Framework*. Retrieved from <http://www.englishprofile.org/images/pdf/GuideToCEFR.pdf>
- Capel, A. (2010). A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1(e3), 1-11. <https://doi.org/10.1017/S2041536210000048>
- Chang, A. C., & Millett, S. (2014). The effect of extensive listening on developing L2 listening fluency: Some hard evidence. *ELT journal*, 68(1), 31-40. <https://doi.org/10.1093/elt/cct052>
- Chapelle, C. A. (2013). Conceptions of validity. In Fulcher, G., & Davidson, F. (Eds.). *The Routledge handbook of language testing*, (pp. 21-33). New York: Routledge.
- 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. <https://doi.org/10.1177/0265532216676851>

- Cicourel, A. V. (2007). A personal, retrospective view of ecological validity. *Text & Talk*, 27(5), 735-752. <https://doi.org/10.1515/TEXT.2007.033>
- Cobb, T. (2013). Frequency 2.0: Incorporating homoforms and multiword units in pedagogical frequency lists. In Bardel, C., Lindqvist, C., & Laufer, B. (Eds.) *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*, (pp. 79-108). EUROSLA-the European Second Language Association. Retrieved from <https://www.eurosla.org/>
- Cobb, T. (2019). Compleat Web VP v.2 [computer program]. Accessed on 16 Jan 2019 at <https://www.lex tutor.ca/cgi-bin/range/texts/index.pl>
- Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. New York: Academic Press Inc.
- Conklin, K., & Pellicer-Sánchez, A. (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research*, 32(3), 453-467. <https://doi.org/10.1177/0267658316637401>
- Conrad, L. (1985). Semantic versus syntactic cues in listening comprehension. *Studies in Second Language Acquisition* 7(1), 59-72. <https://doi.org/10.1017/S0272263100005155>
- Cop, U., Drieghe, D., & Duyck, W. (2015). Eye Movement Patterns in Natural Reading: A Comparison of Monolingual and Bilingual Reading of a Novel. *PloS One*, 10(8), e0134008–. <https://doi.org/10.1371/journal.pone.0134008>
- Costa, P., & Albergaria-Almeida, P. (2015). The European survey on language competences: Measuring foreign language student proficiency. *Procedia-Social and Behavioral Sciences*, 191, 2369-2373. <https://doi.org/10.1016/j.sbspro.2015.04.255>
- Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, 34(2), 213–238. <https://doi.org/10.2307/3587951>
- Cross, J. D. (2009). Diagnosing the process, text and intrusion problems responsible for L2 listeners' decoding errors. *Asian EFL Journal*, 11(2), 31–53.
- Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, 41(4), 965-981. <https://doi.org/10.1016/j.system.2013.08.002>

DeKeyser, R. M., & Koeth, J. (2011). Cognitive aptitudes for second language learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning*, 2, 395-406. Routledge.

Doff, A., Thaine, C., Puchta, H., Stranks, J., Lewis-Jones, P., & Burton, G. (2015). *Cambridge English Empower. Pre-intermediate, Student's book. B1*. Cambridge University Press.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.

Eberhard, D., Simons, G., Fennig C. (eds.). 2021. *Ethnologue: Languages of the World*. Twenty-fourth edition. Dallas, Texas: SIL International. Online version: <http://www.ethnologue.com>

English Profile (2015-2020) [online] Cambridge University Press.
<http://englishprofile.org/>

European Commission (2012). *First European survey on language competences. Final Report*. Publications Office of the European Union. Version 4.0, 15 June 2012. Retrieved from
https://crell.jrc.ec.europa.eu/sites/default/files/files/eslc/ESLC_Final%20Report_210612.pdf

Eyckmans, J. (2004). *Measuring Receptive Vocabulary Size: Reliability and Validity of the Yes/No Vocabulary Test for French-speaking Learners of Dutch* (Doctoral dissertation, Netherlands Graduate School of Linguistics). Utrecht: LOT

Family Education (2019). [online] 2000–2017 Sandbox Networks, Inc., publishing as Family Education. Retrieved from:
<https://www.familyeducation.com/baby-names/browse-origin/surname/english>

Fan, J., & Knoch, U. (2019). Fairness in language assessment: What can the Rasch model offer. *Papers in Language Testing and Assessment*, 8(2), 117-142.

Ferris, D. (1998). Students' Views of Academic Aural/Oral Skills: A Comparative Needs Analysis. *Tesol Quarterly* 32(2), 289-318.
<https://doi.org/10.2307/3587585>

- Field, J. (1998). Skills and strategies: towards a new methodology for listening. *ELT Journal* 52(2), 110-118. <https://doi.org/10.1093/elt/52.2.110>
- Field, J. (1999). "Bottom-up" and "top-down." *ELT Journal*, 53(4), 338–339. <https://doi.org/10.1093/eltj/53.4.338>
- Field, J. (2004). An insight into listeners' problems: Too much bottom-up or too much top-down?. *System*, 32(3), 363-377. <https://doi.org/10.1016/j.system.2004.05.002>
- Field, J. (2008a). Bricks or mortar: which parts of the input does a second language listener rely on? *TESOL quarterly*, 42(3), 411-432. <https://doi.org/10.1002/j.1545-7249.2008.tb00139.x>
- Field, J. (2008b). Emergent and divergent: A view of second language listening research. *System* 36(1), 2–9. <https://doi.org/10.1016/j.system.2008.01.001>
- Field, J. (2009). *Listening in the Language Classroom*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511575945>
- Field, J. (2010). Listening in the language classroom. *ELT journal*, 64(3), 331-333. <https://doi.org/10.1093/elt/ccq026>
- Field, J. (2012). The cognitive validity of the lecture-based question in the IELTS listening paper. In: *IELTS Collected Papers 2: Research in reading and listening assessment* (pp. 391-453). Cambridge: Cambridge University Press.
- Fountain, R. L., & Nation, I. S. P. (2000). A vocabulary-based graded dictation test. *RELC journal*, 31(2), 29-44. <https://doi.org/10.1177/003368820003100202>
- Fung, D., & Macaro, E. (2019). Exploring the relationship between linguistic knowledge and strategy use in listening comprehension. *Language Teaching Research: LTR*, 1362168819868879. <https://doi.org/10.1177/1362168819868879>
- Gardner, D., & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL quarterly*, 41(2), 339-359. <https://doi.org/10.1002/j.1545-7249.2007.tb00062.x>
- Goh, C. (2005). Second language listening expertise. In *Expertise in second language learning and teaching* (pp. 64-84). London: Palgrave Macmillan. https://doi.org/10.1057/9780230523470_4

- Goh, C. (2008). Metacognitive instruction for second language listening development: Theory, practice and research implications. *RELC journal*, 39(2), 188-213. <https://doi.org/10.1177/0033688208092184>
- Goh, C. C. (2002). Exploring listening comprehension tactics and their interaction patterns. *System (Linköping)*, 30(2), 185-206. [https://doi.org/10.1016/s0346-251x\(02\)00004-0](https://doi.org/10.1016/s0346-251x(02)00004-0)
- Goh, C. C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*. 28(1), 55-75. [https://doi.org/10.1016/s0346-251x\(99\)00060-3](https://doi.org/10.1016/s0346-251x(99)00060-3)
- Graham, S. (2002). Experiences of learning French: a snapshot at Years 11, 12 and 13. *Language Learning Journal*, 25(1), 15-20. <https://doi.org/10.1080/09571730285200051>
- Graham, S. (2006). Listening comprehension: The learners' perspective. *System (Linköping)*, 34(2), 165-182. <https://doi.org/10.1016/j.system.2005.11.001>
- Graham, S. (2011). Self-efficacy and academic listening. *Journal of English for Academic Purposes*, 10(2), 113-117. <https://doi.org/10.1016/j.jeap.2011.04.001>
- Graham, S., & Santos, D. (2015). *Strategies for second language listening: Current scenarios and improved pedagogy*. Palmgrave McMillan UK. <https://doi.org/10.1057/9781137410528>
- Graham, S., Santos, D., & Francis-Brophy, E. (2014). Teacher beliefs about listening in a foreign language. *Teaching and Teacher Education*, 40, 44-60. <https://doi.org/10.1016/j.tate.2014.01.007>
- Graham, S., Santos, D., & Vanderplank, R. (2008). Listening comprehension and strategy use: A longitudinal exploration. *System*, 36(1), 52-68. <https://doi.org/10.1016/j.system.2007.11.001>
- Graham, S., Santos, D., & Vanderplank, R. (2010). Strategy clusters and sources of knowledge in L2 listening comprehension. *Innovation in Language Learning and Teaching* 4(1), 1– 20. <https://doi.org/10.1080/17501220802385866>

Guilleux, A., Blanchin, M., Hardouin, J., & Sébille, V. (2014). Power and sample size determination in the Rasch model: evaluation of the robustness of a numerical method to non-normality of the latent trait. *PloS One*, 9(1), e83652–. <https://doi.org/10.1371/journal.pone.0083652>

Gyllstad, H. (2013). Looking at L2 vocabulary knowledge dimensions from an assessment perspective—challenges and potential solutions. Bardel, C., Lindqvist, C., & Laufer, B. (Eds.) *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*, (pp. 11-28). EUROSLA—the European Second Language Association. Retrieved from <https://www.eurosla.org/>

Hatch, E. M., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York, NY: Heinle & Heinle Publishers.

Hazenberg, S., & Hulstijn, J. H. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. *Applied linguistics*, 17(2), 145-163. <https://doi.org/10.1093/applin/17.2.145>

Heigham, J., & Croker, R. R. (2009). *Qualitative Research in Applied Linguistics*, Basingstoke: Palgrave.

Hirsch, D. and Nation, P. (1992). What Vocabulary Size is Needed to Read Unsimplied Texts for Pleasure? *Reading in a Foreign Language* 8(2), 689-696

Holzknrecht, F. (2019). *Double play in listening assessment*. Lancaster University. <https://doi.org/10.17635/lancaster/thesis/812>

Hu, H. M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a foreign language*, 13(1), 403-430.

Huang, H. T. (2010). *How Does Second Language Vocabulary Grow over Time? A Multi-Methodological Study of Incremental Vocabulary Knowledge Development* (Doctoral dissertation, University of Hawai'i).

Hulstijn, J. (2003). Connectionist Models of Language Processing and the Training of Listening Skills With the Aid of Multimedia Software. *Computer Assisted Language Learning*, 16(5), 413–425. <https://doi.org/10.1076/call.16.5.413.29488>

Jones, N. (2013). Reliability and dependability. In Fulcher, G., & Davidson, F. (Eds.). *The Routledge handbook of language testing*, (pp. 350-362). New York: Routledge.

Karami, H. (2012). The development and validation of a bilingual version of the Vocabulary Size Test. *RELC Journal*, 43(1), 53-67.

<https://doi.org/10.1177/0033688212439359>

Kelly, P. (1991). Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners. *International Review of Applied Linguistics in Language Teaching*, 29(2), 135-149.

<https://doi.org/10.1515/iral.1991.29.2.135>

Kim, J. (2002). Affective reactions to foreign language listening: retrospective interviews with Korean EFL students. *Language Research* 38(1), 117-151.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension. In Lauren, C., and Norman, M., (eds), *Special language: From humans thinking to thinking machines* (pp. 316-323). Clevedon: Multilingual Matters.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension. *Reading in a foreign language*, 22(1), 15-30.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English*. Harlow, UK: Longman.

LeLoup, J. W. & Ponterio, R. (2007). Listening: You've got to be carefully taught. *Language Learning & Technology* 11(1), 4-15. Retrieved from

<http://llt.msu.edu/vol11num1/net/>

Levitzky-Aviad, T., & Laufer, B. (2013). Lexical properties in the writing of foreign language learners over eight years of study: Single words and collocations. In Bardel, C., Lindqvist, C., & Laufer, B. (Eds.) *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*, (pp. 127-148). EUROSLA-the European Second Language Association. Retrieved from <https://www.eurosla.org/>

Lim, G. S., & Khalifa, H. (2013). Criterion-related validity. In Garanpayeh, A., & Taylor, L. (Eds.). *Examining Listening: Research and Practice in Assessing Second Language Listening* (Vol. 35), (pp. 303-321). Cambridge: Cambridge University Press.

Linacre, J. M. (1997). KR-20 / Cronbach Alpha or Rasch Person Reliability: Which Tells the "Truth"? *Rasch Measurement Transactions*, 11(3), 580-581.

Linacre, J. M. (2012). *A user's guide to Winsteps Ministeps Rasch-model computer programs* [version 3.74.0]. Retrieved from <http://www.winsteps.com/index.htm>

Linacre, J. M. (2012, 2019). Winsteps® Rasch Measurement, version 4.4.3. [Computer software] Downloaded from <http://www.winsteps.com>

Linacre, J. M. (2014, 28th February). Using Rasch Measurement and Parametric Tests. [Forum post]. Retrieved from <https://raschforum.boards.net/thread/39/using-rasch-measurements-parametric-tests>

Linacre, J. M., & Wright, B.D. (1989). The "Length" of a Logit. *Rasch Measurement Transactions*, 1989, 3(2), 54-55.

Lind Pantzare, A. (2015). Interrater reliability in large-scale assessments—Can teachers score national tests reliably without external controls?. *Practical Assessment, Research, and Evaluation*, 20(1), 9. <https://doi.org/10.7275/y2enzm89>

Lynch, T. (1997). Life in the slow lane: Observations of a limited L2 listener. *System*, 25(3), 385-398. [https://doi.org/10.1016/S0346-251X\(97\)00030-4](https://doi.org/10.1016/S0346-251X(97)00030-4)

Lynch, T. (1998). Theoretical perspectives on listening. *Annual review of applied linguistics*, 18, 3-19. <https://doi.org/10.1017/s0267190500003457>

Mackey, A., & Gass, S. M. (2015). *Second language research: Methodology and design* (Second Edition). London: Routledge.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied linguistics*, 33(3), 299-320. <https://doi.org/10.1093/applin/ams010>

- Masrai, A. (2020). Exploring the impact of individual differences in aural vocabulary knowledge, written vocabulary knowledge and working memory capacity on explaining L2 learners' listening comprehension. *Applied Linguistics Review*, 11(3), 423-447. <https://doi.org/10.1515/applirev-2018-0106>
- Matthews, J. (2018). Vocabulary for listening: Emerging evidence for high and mid-frequency vocabulary knowledge. *System (Linköping)*, 72, 23-36. <https://doi.org/10.1016/j.system.2017.10.005>
- Matthews, J., & Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System (Linköping)*, 52, 1-13. <https://doi.org/10.1016/j.system.2015.04.015>
- Matthews, J., O'Toole, J. M., & Chen, S. (2017). The impact of word recognition from speech (WRS) proficiency level on interaction, task success and word learning: design implications for CALL to develop L2 WRS. *Computer Assisted Language Learning*, 30(1-2), 22-43. <https://doi.org/10.1080/09588221.2015.1129348>
- Maza, T. L. (2020). Las escuelas oficiales de idiomas: Una perspectiva histórica. *e-CO: Revista digital de educación y formación del profesorado*, (17), 465-489. <http://revistaeco.cepcordoba.es/wp-content/uploads/2020/04/Linan2.pdf>
- McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research: LTR*, 19(6), 741-760. <https://doi.org/10.1177/1362168814567889>
- Meara, P. M., & Miralpeix, I. (2006). Y_Lex: The Swansea advanced vocabulary levels test. v2. 05. Swansea: Lognostics.
- Mendelsohn, D. (2001). Listening Comprehension: We've come a long way, but... *Contact* 27(2), 33-40. Retrieved from: <https://www.teslontario.org/uploads/publications/researchsymposium/ResearchSymposium2001.pdf#page=33>
- Mendelsohn, D. J. (2006). Learning How to Listen Using Learning Strategies. In E. Usó-Juan & A. Martínez-Flor (Eds.), *Current Trends in the Development and Teaching of the Four Language Skills* (pp. 75–89). Mouton de Gruyter; Walter de Gruyter, Inc. <https://doi.org/10.1515/9783110197778.2.75>

- Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Score Meaning. *American Psychologist*, 50(9), 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Mills, N., Pajares, F., & Herron, C. (2006). A reevaluation of the role of anxiety: Self-efficacy, anxiety, and their relation to reading and listening proficiency. *Foreign language annals*, 39(2), 276-295. <https://doi.org/10.1111/j.1944-9720.2006.tb02266.x>
- Milton, J. (2009). Measuring second language vocabulary acquisition. Bristol: Multilingual Matters.
- Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In Bardel, C., Lindqvist, C., & Laufer, B. (Eds.) *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*, (pp. 57-78). EUROSLA-the European Second Language Association. Retrieved from <https://www.eurosla.org/>
- Milton, J., & Hopkins, N. (2006). Comparing Phonological and Orthographic Vocabulary Size: Do Vocabulary Tests Underestimate the Knowledge of Some Learners. *The Canadian Modern Language Review / La Revue Canadienne Des Langues Vivantes*, 63(1), 127–147. <https://doi.org/10.1353/cml.2006.0048>
- Milton, J., Wade, J., & Hopkins, N. (2010). Aural Word Recognition and Oral Competence in English as a Foreign Language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. del M. Torreblanca-López (Eds.), *Insights into Non-Native Vocabulary Teaching and Learning* (pp. 83–98). Multilingual Matters.
- Ministerio de Educación y Formación Profesional (2020). [online] *Enseñanzas no universitarias / Alumnado matriculado / Curso 2019-2020*. Retrieved from: http://estadisticas.mecd.gob.es/EducaDynPx/educabase/index.htm?type=pcaxis&path=/no-universitaria/alumnado/matriculado/2019-2020-rd/especial_idiomas&file=pcaxis&l=s0
- Nation, I. S. P. (2001) *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- Nation, I. S. P. (2005). Teaching and learning vocabulary. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning*. 581-595. Routledge.

- Nation, I. S. P. (2006) How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1). pp. 59-82.
<https://doi.org/10.3138/cmlr.63.1.59>
- Nation, I. S. P. (2007). Fundamental issues in modelling and assessing vocabulary knowledge. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge*, 35-43. Cambridge: Cambridge University Press
- Nation, I. S. P. (2012, 2019). The BNC/COCA word family lists (17 September 2012). Unpublished paper. [online] Retrieved from
<http://www.victoria.ac.nz/lals/about/staff/paul-nation>
- Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins. <https://doi.org/10.1093/applin/amx052>
- Nation, I. S. P. (2017). The BNC/COCA Level 6 word family lists (Version 1.0.0) [Data file]. Retrieved from <http://www.victoria.ac.nz/lals/staff/paul-nation.aspx>
- Nation, I. S. P., & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle, Cengage Learning.
- Nation, P., & Coxhead, A. (2014). Vocabulary size research at Victoria University of Wellington, New Zealand. *Language Teaching*, 47(3), 398-403.
<https://doi.org/10.1017/S0261444814000111>
- Nguyen, L.T.C., & Nation, P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC journal*, 42(1), 86-99.
<https://doi.org/10.1177/0033688210390264>
- Nunan, D. (2002). Listening in language learning. In Richards, J. C., & Renandya, W. A. (Eds.) *Methodology in language teaching: An anthology of current practice*, (pp. 238-241). Cambridge University Press.
- Ohata, K. (2006). Auditory short-term memory in L2 listening comprehension processes. *Journal of Language and Learning*, 5(1), 21-27.
- Onwuegbuzie, A. J., & Leech, N. L. (2005). Taking the “Q” out of research: Teaching research methodology courses without the divide between quantitative and qualitative paradigms. *Quality and Quantity*, 39(3), 267-295.
<https://doi.org/10.1007/s11135-004-1670-0>

- Paas, F., & Sweller, J. (2014). Implications of cognitive load theory for multimedia learning. In Mayer, R. E. (Ed.), *The Cambridge handbook of multimedia learning*, 2nd edition, (pp. 27–42). Cambridge : Cambridge University Press. <https://doi.org/10.1017/CBO9781139547369.004>
- Pan, Y. C., Tsai, T. H., Huang, Y. K., & Liu, D. (2018). Effects of expanded vocabulary support on L2 listening comprehension. *Language Teaching Research*, 22(2), 189-207. <https://doi.org/10.1177/1362168816668895>
- Parent, K. (2012). The most frequent English homonyms. *RELC Journal*, 43(1), 69-81. <https://doi.org/10.1177/0033688212439356>
- Plonsky, L., & Ghanbar, H. (2018). Multiple Regression in L2 Research: A Methodological Synthesis and Guide to Interpreting R² Values. *The Modern Language*, 102(4), 713–731. <https://doi.org/10.1111/modl.12509>
- Pollatsek, A., & Well, A. D. (1995). On the use of counterbalanced designs in cognitive research: A suggestion for a better and more powerful analysis. *Journal of Experimental psychology: Learning, memory, and Cognition*, 21(3), 785. <https://doi.org/10.1037/0278-7393.21.3.785>
- Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research. *Language Learning*, 65(s1), 37-75. <https://doi.org/10.1111/lang.12112>
- Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language testing*, 10(3), 355-371. <https://doi.org/10.1177/026553229301000308>
- Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41 –60). Mahwah, NJ : Erlbaum .
- Read, J. (2013). Second language vocabulary assessment. *Language Teaching*, 46(1), 41-52. <https://doi.org/10.1017/S0261444812000377>
- Richards, J. C. (2008a). *Moving Beyond the Plateau. From Intermediate to Advanced Levels in Language Learning*. New York: CUP
- Richards, J. C. (2008b). *Teaching listening and speaking from theory to practice*. New York: Cambridge University Press.

Richards, J., & Schmidt, R. (2002). *Longman dictionary of language teaching and applied linguistics*. (3rd ed. / Jack C. Richards and Richard Schmidt ; with Heidi Kendricks and Youngkyu Kim.). Longman.

Ridgway, T. (2000). Listening Strategies - I Beg Your Pardon? *ELT Journal*, 54(2), 179–185. <https://doi.org/10.1093/elt/54.2.179>

Rost, M. (2005). L2 Listening. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 503-527). Mahwah, NJ: Lawrence Erlbaum Associates.

Rost, M. (2006). Areas of research that influence L2 listening instruction. In E. Usó-Juan & A. Martínez-Flor (Eds.), *Current Trends in the Development and Teaching of the Four Language Skills* (pp. 47-73). Mouton de Gruyter; Walter de Gruyter, Inc. <https://doi.org/10.1515/9783110197778.2.75>

Rost, M. (2011). *Teaching and Researching Listening* (2nd ed.). New York: Pearson Education.

Rost, M., & Wilson, J. J. (2013). *Active listening*. Pearson.

Rubin, J. (1994). A Review of Second Language Listening Comprehension Research. *The Modern Language Journal (Boulder, Colo.)*, 78(2), 199–221. <https://doi.org/10.1111/j.1540-4781.1994.tb02034.x>

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied linguistics*, 11(2), 129-158. <https://doi.org/10.1093/applin/11.2.129>

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language teaching research: LTR*, 12(3), 329-363. <https://doi.org/10.1177/1362168808089921>

Schmitt, N., Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 47(4), 484-503. <https://doi.org/10.1017/S0261444812000018>

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching*, 50(2), 212-226. <https://doi.org/10.1017/S0261444815000075>

- Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1), 26-43. <https://doi.org/10.1111/j.1540-4781.2011.01146.x>
- Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, 53(1), 109-120. <https://doi.org/10.1017/s0261444819000326>
- Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language testing*, 18(1), 55-88. <https://doi.org/10.1177/026553220101800103>
- Schmitt, N. & Zimmerman C.B. (2002). Derivative Word Forms: What Do Learners Know? *Tesol Quarterly*, 36, 145-171. <https://doi.org/10.2307/3588328>
- Schmuckler, M. A. (2001). What is ecological validity? A dimensional analysis. *Infancy*, 2 (4), 419-436. https://doi.org/10.1207/s15327078in0204_02
- Scrivener, J., & Jones, C. (2012). *Straightforward. Student's book* (2nd ed.). Macmillan Education.
- Side, R., (1990). Phrasal verbs: sorting them out, *ELT Journal* 44(2), 144–152. <https://doi.org/10.1093/elt/44.2.144>
- Siegel, J. (2013). Exploring L2 listening instruction: examinations of practice. *ELT Journal* 68(1), 22-30. <https://doi.org/10.1093/elt/cct058>
- Siegel, J. (2015). *Exploring listening strategy instruction through action research*. Basingstoke: Palmgrave Macmillan.
- Silva, B. B., & Otwinowska, A. (2019). VST as a reliable academic placement tool despite cognate inflation effects. *English for Specific Purposes*, 54, 35-49. <https://doi.org/10.1016/j.esp.2018.12.001>
- Skehan, P. (2002). Theorising and updating aptitude. In Robinson, P. (Ed.), *Individual differences and instructed language learning*, 2, 69-94. John Benjamins Publishing Co.
- Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2), 139-152. <https://doi.org/10.1080/09571730802389975>

- Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in second language acquisition*, 31(4), 577-607. <https://doi.org/10.1017/S0272263109990039>
- Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading research quarterly* 16(1), 32-71. <https://doi.org/10.2307/747348>
- Street, J. & Ingham, K. (2007). Publishing vocabulary lists for BEC Preliminary, PET and KET examinations. *Research Notes*, 27, 4-7.
- Summers, D., & Gadsby, A. (Eds.). (1987). *Longman dictionary of contemporary English*. Essex: Longman
- Suzuki, Y., Nakata, T., & Dekeyser, R. (2019). Optimizing second language practice in the classroom: Perspectives from cognitive psychology. *The Modern Language Journal*, 103(3), 551-561. <https://doi.org/10.1111/modl.12582>
- Tomlinson, B. (2013). *Developing materials for language teaching* (Second edition.). London: Bloomsbury.
- Tsui, A. M. B. and Fullilove, J. (1998). Bottom-up or Top-down Processing as a Discriminator of L2 Listening Performance. *Applied Linguistics*, 19(4), 432-451. <https://doi.org/10.1093/applin/19.4.432>
- UCLES (2012). The Cambridge English: Preliminary and Preliminary for Schools Vocabulary List. Retrieved from <https://www.cambridgeenglish.org/Images/84669-pet-vocabulary-list.pdf>.
- Accessed on 19th January 2019
- UCLES (2015). Cambridge English: The Cambridge English Scale. Retrieved from: <https://www.cambridgeenglish.org/Images/167506-cambridge-english-scale-factsheet.pdf>.
- UCLES (2019). Cambridge English: The Cambridge English Scale Explained. A guide to converting practice test scores to Cambridge English Scale scores. Retrieved from <https://www.cambridgeenglish.org/Images/210434-converting-practice-test-scores-to-cambridge-english-scale-scores.pdf>.
- van Zeeland, H. (2014a). *Second language vocabulary knowledge in and from listening* (Doctoral dissertation, University of Nottingham).

- van Zeeland, H. (2014b). Lexical inferencing in first and second language listening. *The Modern Language Journal*, 98(4), 1006-1021.
<https://doi.org/10.1111/modl.1215>
- van Zeeland, H. (2017). Christopher Brumfit Thesis Award Winner 2014 – Hilde van Zeeland: Four studies on vocabulary knowledge in and from listening: Findings and implications for future research. 50(1), 143–150.
<https://doi.org/10.1017/S0261444816000318>
- van Zeeland, H. (2018). Vocabulary in Listening. In Liontas, J.I., T. International Association and DelliCarpini, M. (Eds.), *The TESOL Encyclopedia of English Language Teaching*, (pp. 1-6). <https://doi.org/10.1002/9781118784235.eelt0614>
- van Zeeland, H., & Schmitt, N. (2013a). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, 41(3), 609-624.
<https://doi.org/10.1016/j.system.2013.07.012>
- van Zeeland, H., & Schmitt, N. (2013b). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, 34(4), 457-479. <https://doi.org/10.1093/applin/ams074>
- Vandergrift, L. (2003). Orchestrating strategy use: Toward a model of the skilled second language listener. *Language learning*, 53(3), 463-496.
<https://doi.org/10.1111/1467-9922.00232>
- Vandergrift, L. (2004). Listening to Learn or Learning to Listen? *Annual Review of Applied Linguistics*, 24, 3–25. <https://doi.org/10.1017/S0267190504000017>
- Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, 40(3), 191-210.
<https://doi.org/10.1017/S0261444807004338>
- Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, 65(2), 390-416. <https://doi.org/10.1111/lang.12105>
- Vandergrift, L. & Goh, C. C. (2012). *Teaching and Learning Second Language Listening: Metacognition in Action*. Taylor & Francis Ltd - M.U.A.
<https://doi.org/10.4324/9780203843376>

- Vandergrift, L., & Tafaghodtari, M. H. (2010). Teaching L2 learners how to listen does make a difference: An empirical study. *Language learning*, 60(2), 470-497. <https://doi.org/10.1111/j.1467-9922.2009.00559.x>
- Vandergrift, L., Goh, C. C., Mareschal, C. J., & Tafaghodtari, M. H. (2006). The metacognitive awareness listening questionnaire: Development and validation. *Language learning*, 56(3), 431-462. <https://doi.org/10.1111/j.1467-9922.2006.00373.x>
- Wang, Y., & Treffers-Daller, J. (2017). Explaining listening comprehension among L2 learners of English: The contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System (Linköping)*, 65, 139-150. <https://doi.org/10.1016/j.system.2016.12.013>
- Webb, S. (2002) Investigating the effects of learning tasks on vocabulary knowledge. Unpublished PhD thesis, Victoria University of Wellington, New Zealand.
- Webb, S., Sasao, Y., & Ballance, O. (2017). The updated vocabulary levels test. *ITL-International Journal of Applied Linguistics*, 168(1), 33-69. <https://doi.org/10.1075/itl.168.1.02web>
- White, B. J. (2012). A conceptual approach to the instruction of phrasal verbs. *The Modern Language Journal*, 96(3), 419-438. <https://doi.org/10.1111/j.1540-4781.2012.01365.x>
- Wiley, J., Sanchez, C., & Jaeger, A. (2014). The Individual Differences in Working Memory Capacity Principle in Multimedia Learning. In Mayer, R. E. (Ed.), *The Cambridge Handbook of Multimedia Learning*, 2nd edition, (pp. 598-620). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139547369.029>
- Wright, B. D. (1991). Scores, reliabilities and assumptions. *Rasch Measurement Transactions*, 5(3), 157-158
- Wright, B. D. (1997). A history of social science measurement. *Educational measurement: issues and practice*, 16(4), 33-45. <https://doi.org/10.1111/j.1745-3992.1997.tb00606.x>
- Wright, B. D. & Linacre J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.

- Xu, F. (2011). Anxiety in EFL Listening Comprehension. *Theory and Practice in Language Studies*, 1(12), 1709-1717. <https://doi.org/10.4304/tpls.1.12.1709-1717>
- Yi, F. (2011). Plateau of EFL learning: A psycholinguistic and pedagogical study. Retrieved from <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.565.6275&rep=rep1&type=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. <https://doi.org/10.1177/026553229801500102>
- Zhang, P. (2018). *Comparing different types of EFL vocabulary instruction for Chinese senior secondary school learners of English* (Doctoral dissertation, University of Reading). http://centaur.reading.ac.uk/77933/12/22841392_Zhang_thesis_redacted.pdf
- Zhang, P., & Graham, S. (2020). Learning vocabulary through listening: The role of vocabulary knowledge and listening proficiency. *Language Learning*, 70(4), 1017-1053. <https://doi.org/10.1111/lang.12411>
- Zhao, P., & Ji, X. (2018). Validation of the Mandarin version of the Vocabulary Size Test. *RELC Journal*, 49(3), 308-321. <https://doi.org/10.1177/0033688216639761>

APPENDICES

APPENDIX 1 – Listening Vocabulary Size Test – LVST (McLean et al., 2015)

This is a vocabulary test. Each English word will be read together with an example sentence. Select the Japanese word from the choices (a–d) that has the closest meaning to the English word being read.

Each question will be read only once.

Example Problem

1.

- a. 食べた (ate)
- b. 待った (waited)
- c. 見た (saw)
- d. 寝た (slept)

The correct answer is b.

If you do not know a word at all, please leave it blank.

However, if you think there is a chance that you may know the word, please try to answer.

Let's practice some problems.

Practice Problem 1

- a. 強い (strong)
- b. 幸せな (happy)
- c. 食べすぎる (eats too much)
- d. 親切的な (kind)

Practice Problem 2

- a. ~について話します (talk about)
- b. ~を運ぶ (carry)
- c. ~に名前を書く (write your name on)
- d. ~を振る (shake)

Because this is a listening test, please do not speak until it is finished.

Let's begin.

APPENDIX 2 – Phrasal Expressions List (Martinez & Schmitt, 2012)

The PHRASE List

Phrasal expressions divided to match 1K frequency bands of the most common word families in the BNC. The 'Integrated List Rank' represents where each item falls when both lists (individual and phrase lists) are merged together.

There are three categories of genre with frequency information to help the list user discern the appropriateness/usefulness of each phrase. The frequency information breaks down as follows:

- *** = phrase most common in this genre (or as common)
- ** = phrase less common in this genre
- * = phrase infrequent in this genre
- X = phrase rare or non-existent in this genre

Integrated List - Rank	Phrase	Frequency (per 100 million)	Spoken general	Written general	Written academic	Example
107	HAVE TO	83092	***	**	*	I exercise because I have to .
165	THERE IS/ARE	59833	***	***	**	There are some problems.
415	SUCH AS	30857	*	***	***	We have questions, such as how it happened.
463	GOING TO (FUTURE)	28259	***	**	x	I'm going to think about it.
483	OF COURSE	26966	***	**	*	He said he'd come of course .
489	A FEW	26451	***	**	*	After a few drinks, she started to dance.
518	AT LEAST	25034	***	**	**	Well, you could email me at least .
551	SUCH A(N)	23894	***	**	*	She had such a strange sense of humor.
556	I MEAN	23616	***	X	x	It's fine, but, I mean , is it worth the price?
598	A LOT	22332	***	*	x	They go camping a lot in the summer.
631	RATHER THAN	21085	***	***	***	Children, rather than adults, tend to learn quickly.
635	SO THAT	20966	***	**	*	Park it so that the wheels are curbed.
655	A LITTLE	20296	***	**	*	I like to work out a little before dinner.
674	A BIT (OF)	19618	***	**	x	There was a bit of drama today at the office.
717	AS WELL AS	18041	***	***	***	She jogs as well as swims.
803	IN FACT	15983	***	***	**	The researchers tried several approaches, in fact .
807	(BE) LIKELY TO	15854	***	***	***	To be honest, I'm likely to forget.

APPENDIX 3 – Listening Vocabulary Test – 150 Items (May 2019)

VOCABULARY SIZE TEST

This test has TWO PARTS. Each part will take you about 25 minutes to finish. It is very important that you do the two parts of the test and that you try to answer **ALL THE QUESTIONS** in the test. There are no negative marks for incorrect answers.

Listening Vocabulary Size Test – Listen to the recording and select the answer (a, b, c, OR d) with the closest Spanish translation to the key word in the question.

Example 1 – You will hear:

SCHOOL – This **school** is new.

- A. cama
- B. escuela
- C. parque
- D. supermercado

The closest translation for the target word that you have heard is '*escuela*', so the answer you have to mark is **B**.

Example 2 - You will hear:

PLAY – They **play** it very often.

- A. beber
- B. cocinar
- C. comer
- D. jugar

The closest translation for the target word that you have heard is '*jugar*', so the answer you have to mark is **D**.

Example 3 - You will hear:

STRONG – They are really **strong**.

- A. alto
- B. feliz
- C. fuerte
- D. rico

The closest translation for the target word that you have heard is '*fuerte*', so the answer you have to mark is **C**.

Example 4 - You will hear:

TODAY – They need it **today**.

- A. hoy
- B. siempre
- C. también
- D. todavía

The closest translation for the target word that you have heard is '*hoy*', so the answer you have to mark is **A**.

PRUEBA DE VOCABULARIO

Esta prueba tiene DOS PARTES. Terminar cada parte te llevará unos 20 minutos. Es muy importante que hagas las dos partes de la prueba y que intentes contestar TODAS LAS PREGUNTAS en la prueba. No hay puntos negativos por respuestas incorrectas.

Prueba de Comprensión Oral de Vocabulario – Escucha la grabación y selecciona la respuesta (a-d) con la traducción en español más próxima a la palabra clave de la pregunta

Ejemplo 1 – Escucharás:

SCHOOL – This **school** is new.

- A. cama
- B. escuela
- C. parque
- D. supermercado

La traducción más próxima a la palabra que has escuchado es '*escuela*', así que la respuesta que tienes que marcar es **B**.

Ejemplo 2 – Escucharás:

PLAY – They **play** it very often.

- A. beber
- B. cocinar
- C. comer
- D. jugar

La traducción más próxima a la palabra que has escuchado es '*jugar*', así que la respuesta que tienes que marcar es **D**.

Ejemplo 3 – Escucharás:

STRONG – They are really **strong**.

- A. alto
- B. feliz
- C. fuerte
- D. rico

La traducción más próxima a la palabra que has escuchado '*fuerte*', así que la respuesta que tienes que marcar es **C**.

Ejemplo 4 – Escucharás:

TODAY – They need it **today**.

- A. hoy
- B. siempre
- C. también
- D. todavía

La traducción más próxima a la palabra que has escuchado '*hoy*', así que la respuesta que tienes que marcar es **A**.

1. _____
a) cubo
b) entrada
c) factura
d) rama
2. _____
a) cuerpo
b) estantería
c) operación
d) pizarra
3. _____
a) garganta
b) piel
c) pierna
d) uña
4. _____
a) aburrido
b) ansioso
c) avergonzado
d) decepcionado
5. _____
a) ayudante
b) comerciante
c) representante
d) suplente
6. _____
a) boda
b) nacimiento
c) negocio
d) reunión
7. _____
a) al revés
b) al mismo tiempo
c) de nuevo
d) después
8. _____
a) pañuelo
b) payaso
c) peine
d) ternero
9. _____
a) cerradura
b) hielo
c) isla
d) tecla
10. _____
a) reconocer
b) recuperar
c) reducir
d) rehusar
11. _____
a) cerdo
b) enchufe
c) pipa
d) tarta
12. _____
a) frase
b) literatura
c) líquido
d) palacio
13. _____
a) al otro lado
b) antes
c) en el extranjero
d) en realidad
14. _____
a) adelante
b) además
c) del mismo modo
d) solo
15. _____
a) acera
b) cuero
c) etiqueta
d) pico
16. _____
a) equipo
b) prueba
c) sistema
d) trimestre
17. _____
a) carne
b) compañero
c) espejo
d) mono
18. _____
a) producto
b) profesión
c) progreso
d) proyecto
19. _____
a) alfabeto
b) barbacoa
c) camello
d) coco
20. _____
a) jersey
b) tableta
c) trompeta
d) vocabulario
21. _____
a) cine
b) destino
c) diccionario
d) documental
22. _____
a) comercio
b) hojalata
c) ladrón
d) maletero
23. _____
a) amplio
b) digno
c) húmedo
d) salvaje
24. _____
a) anunciar
b) arrestar
c) atraer
d) avanzar
25. _____
a) acampar
b) atrapar
c) comprar
d) cuidar
26. _____
a) cielo
b) olor
c) sonrisa
d) tamaño
27. _____
a) cualquiera
b) eso
c) otro
d) todo
28. _____
a) aunque
b) a menos que
c) por lo tanto
d) sin embargo
29. _____
a) criatura
b) cultura
c) defensa
d) diseño
30. _____
a) comprobar
b) contestar
c) creer
d) decir
31. _____
a) besar
b) dar patadas
c) mentir
d) reír
32. _____
a) descubrir
b) decidir
c) hacer
d) matar
33. _____
a) bastón
b) huelga
c) paso
d) tema

34. _____
a) dolorido
b) inteligente
c) libre
d) suave
35. _____
a) camión
b) éxito
c) interruptor
d) sueldo
36. _____
a) le
b) lo
c) nos
d) te
37. _____
a) dientes
b) noticias
c) personas
d) ropas
38. _____
a) escena
b) esquí
c) sección
d) señal
39. _____
a) despierto
b) listo
c) rubio
d) valiente
40. _____
a) especial
b) histórico
c) local
d) necesario
41. _____
a) campeón
b) conexión
c) corrección
d) elección
42. _____
a) aumentar
b) combinar
c) herir
d) mejorar
43. _____
a) batería
b) bicicleta
c) murciélago
d) puente
44. _____
a) delicioso
b) diferente
c) difícil
d) peligroso
45. _____
a) asiento
b) carretera
c) regla
d) roca
46. _____
a) asunto
b) escenario
c) gusto
d) impuesto
47. _____
a) grabación
b) informe
c) pensamiento
d) razón
48. _____
a) enfermedad
b) muñeca
c) sobre
d) tambor
49. _____
a) multitud
b) pato
c) polvo
d) techo
50. _____
a) confiado
b) crudo
c) frecuente
d) preciso
51. _____
a) cerrar
b) gastar
c) gritar
d) rasgar
52. _____
a) cabaña
b) contable
c) lavabo
d) taxi
53. _____
a) ascensor
b) combustible
c) experimento
d) puerto
54. _____
a) calefacción
b) grupo
c) historia
d) idea
55. _____
a) ¡Guay!
b) ¡Oye!
c) ¡Vale!
d) ¡Vaya!
56. _____
a) botón
b) capítulo
c) cesta
d) esfuerzo
57. _____
a) afilado
b) agradable
c) educado
d) satisfecho
58. _____
a) este
b) norte
c) oeste
d) sur
59. _____
a) alimentar
b) colgar
c) congelar
d) pegar
60. _____
a) a veces
b) absolutamente
c) con cuidado
d) diariamente
61. _____
a) biblioteca
b) laboratorio
c) liga
d) revista
62. _____
a) conseguir
b) haber
c) poder
d) tener
63. _____
a) derrota
b) ensayo
c) muestra
d) retraso
64. _____
a) corte de pelo
b) pelo
c) peluquero
d) secador de pelo
65. _____
a) alquiler
b) ausencia
c) hierro
d) mejora
66. _____
a) recibir
b) relajar
c) repetir
d) reservar

67. _____
a) cartón
b) cojín
c) desafío
d) goma
68. _____
a) competir
b) consistir
c) discrepar
d) persuadir
69. _____
a) artista
b) cazuela
c) patrón
d) ratón
70. _____
a) abrochar
b) arreglar
c) ordenar
d) subrayar
71. _____
a) colección
b) control
c) conversación
d) coste
72. _____
a) apoyo
b) asignatura
c) mantel
d) traje
73. _____
a) apropiado
b) bajo
c) grande
d) pobre
74. _____
a) bigote
b) jarrón
c) melocotón
d) tazón
75. _____
a) estudios
b) gafas
c) medias
d) peniques
76. _____
a) aprender
b) esconder
c) golpear
d) oír
77. _____
a) clase
b) color
c) ordenador
d) país
78. _____
a) empleado
b) excusa
c) familia
d) padre
79. _____
a) respuesta
b) uso
c) visita
d) voz
80. _____
a) gris
b) marrón
c) naranja
d) verde
81. _____
a) celebrar
b) derramar
c) disculparse
d) saludar
82. _____
a) imagen
b) paga
c) plan
d) precio
83. _____
a) concurso
b) devolución
c) planchado
d) vendaje
84. _____
a) abrir
b) aceptar
c) colorear
d) llegar
85. _____
a) botella
b) pájaro
c) paseo
d) reloj
86. _____
a) bailarín
b) esquina
c) montaña
d) peligro
87. _____
a) copa
b) mascota
c) papá
d) sentimiento
88. _____
a) culpa
b) elemento
c) mobiliario
d) rana
89. _____
a) aterrizar
b) quemar
c) reservar
d) unirse
90. _____
a) batir
b) manejar
c) permitir
d) sostener
91. _____
a) cambio
b) explicación
c) información
d) miembro
92. _____
a) aduanas
b) refrescos
c) saludos
d) servicios
93. _____
a) idioma
b) impresora
c) piscina
d) situación
94. _____
a) algún
b) cada
c) este
d) ningún
95. _____
a) aceite
b) fresa
c) mantequilla
d) pimienta
96. _____
a) añadir
b) cazar
c) lanzar
d) molestar
97. _____
a) ayuda
b) imaginación
c) vacación
d) vecino
98. _____
a) ducha
b) plata
c) silencio
d) sociedad
99. _____
a) decisión
b) defensa
c) detective
d) escritorio

100.	_____	111.	_____	122.	_____
	a) forma b) lado c) punto d) vista		a) alubia b) cebolla c) guisante d) lechuga		a) ahora b) aparte c) nunca d) en total
101.	_____	112.	_____	123.	_____
	a) horno b) magia c) mapa d) partido		a) chico b) jugador c) persona d) policía		a) fábrica b) pañuelo de papel c) traducción d) variedad
102.	_____	113.	_____	124.	_____
	a) consonante b) gimnasia c) limonada d) pingüino		a) cerca b) finalmente c) primero d) temprano		a) deseoso b) desordenado c) disponible d) tímido
103.	_____	114.	_____	125.	_____
	a) con b) para c) que d) sin		a) asustado b) cansado c) envejecido d) inusual		a) año b) azul c) fiesta d) mujer
104.	_____	115.	_____	126.	_____
	a) capaz b) enfadado c) satisfecho d) sorprendido		a) broma b) cocina c) líder d) trabajo		a) estúpido b) mejor c) perdido d) soleado
105.	_____	116.	_____	127.	_____
	a) camarero b) especia c) espía d) estatua		a) hambre b) juez c) sala d) sombrero		a) mediodía b) niebla c) pastilla d) seta
106.	_____	117.	_____	128.	_____
	a) acantilado b) cajón c) jaula d) mejilla		a) calle b) comienzo c) humo d) sonido		a) árbol b) ciencia c) pared d) vídeo
107.	_____	118.	_____	129.	_____
	a) bloguero b) cheque c) jirafa d) yogur		a) increíble b) pacífico c) pequeño d) típico		a) luna de miel b) maleta c) página de inicio d) poste
108.	_____	119.	_____	130.	_____
	a) abajo b) deprisa c) lejos d) tarde		a) millón b) naturaleza c) página d) radio		a) agujero b) colina c) equipaje d) esperanza
109.	_____	120.	_____	131.	_____
	a) decepción b) desarrollo c) intercambio d) meta		a) construir b) romper c) saltar d) soplar		a) aconsejar b) desear c) lamentar d) lograr
110.	_____	121.	_____	132.	_____
	a) edificio b) experiencia c) final d) película		a) altura b) beca c) ejemplo d) rodilla		a) armario b) moneda c) tasa d) tripulación

- 133.** _____
a) paquete
b) patinaje
c) peatón
d) postre
- 134.** _____
a) abrazo
b) cabra
c) guante
d) hoja
- 135.** _____
a) animar
b) castigar
c) reemplazar
d) situar
- 136.** _____
a) descuidado
b) incapaz
c) inconsciente
d) poco amable
- 137.** _____
a) llegada
b) sangre
c) trozo
d) zona
- 138.** _____
a) bolsa de mano
b) caligrafía
c) tablón de anuncios
d) titular
- 139.** _____
a) arquitectura
b) frase
c) profesor
d) recepcionista
- 140.** _____
a) actitud
b) fondo
c) grupo
d) promedio
- 141.** _____
a) antiguo
b) completado
c) desconocido
d) subterráneo
- 142.** _____
a) advertir
b) adivinar
c) amenazar
d) recomendar
- 143.** _____
a) casi nunca
b) de alguna forma
c) en algún lugar
d) por error
- 144.** _____
a) enlace
b) premio
c) rango
d) red
- 145.** _____
a) barbilla
b) codo
c) pulgar
d) tobillo
- 146.** _____
a) bufanda
b) folleto
c) investigación
d) monedero
- 147.** _____
a) canal
b) céntimo
c) círculo
d) costa
- 148.** _____
a) asqueroso
b) bochornoso
c) encantador
d) precioso
- 149.** _____
a) al otro lado de
b) dentro de
c) encima de
d) frente a
- 150.** _____
a) agua
b) guerra
c) tipo
d) verdad

APPENDIX 4 – Written Vocabulary Test – 150 Items (May 2019)

WRITTEN VOCABULARY SIZE TEST

This is the second part of the vocabulary test. Please, FINISH the listening part of the test BEFORE you do this written part. It is also very important that you try to answer **ALL THE QUESTIONS** in the test. There are no negative marks for incorrect answers.

Read the questions and select the answer (a, b, c, OR d) with the closest Spanish translation to the key word in each question.

Example 1 – SCHOOL – This **school** is new.

- A. cama
- B. escuela
- C. parque
- D. supermercado

The closest translation for this word is '*escuela*', so the answer you have to mark is **B**.

Example 2 - PLAY – They **play** it very often.

- A. beber
- B. cocinar
- C. comer
- D. jugar

The closest translation for this word is '*jugar*', so the answer you have to mark is **D**.

Example 3 - STRONG – They are really **strong**.

- A. alto
- B. feliz
- C. fuerte
- D. rico

The closest translation for this word that is '*fuerte*', so the answer you have to mark is **C**.

Example 4 - TODAY – They need it **today**.

- A. hoy
- B. siempre
- C. también
- D. todavía

The closest translation for this word is '*hoy*', so the answer you have to mark is **A**.

PRUEBA DE VOCABULARIO

Esta es la segunda parte de la prueba de vocabulario. Por favor, termina la parte oral del test ANTES de hacer esta parte escrita. Es también muy importante que intentes contestar TODAS LAS PREGUNTAS en la prueba. No hay puntos negativos por respuestas incorrectas.

Lee las preguntas y selecciona la respuesta (a, b, c, d) con la traducción en español más próxima a la palabra clave en cada pregunta.

Ejemplo 1 – SCHOOL – This **school** is new.

- A. cama
- B. escuela
- C. parque
- D. supermercado

La traducción más próxima a esta palabra es '*escuela*', así que la respuesta que tienes que marcar es **B**.

Ejemplo 2 - PLAY – They **play** it very often.

- A. beber
- B. cocinar
- C. comer
- D. jugar

La traducción más próxima a esta palabra es '*jugar*', así que la respuesta que tienes que marcar es **D**.

Ejemplo 3 - STRONG – They are really **strong**.

- A. alto
- B. feliz
- C. fuerte
- D. rico

La traducción más próxima a esta palabra es '*fuerte*', así que la respuesta que tienes que marcar es **C**.

Ejemplo 4 - TODAY – They need it **today**.

- A. hoy
- B. siempre
- C. también
- D. todavía

La traducción más próxima a esta palabra es '*hoy*', así que la respuesta que tienes que marcar es **A**.

1. TICKET: This **ticket** is perfect.
a) cubo
b) entrada
c) factura
d) rama
2. OPERATION: This type of **operation** is perfect.
a) cuerpo
b) estantería
c) operación
d) pizarra
3. SKIN: This **skin** is perfect.
a) garganta
b) piel
c) pierna
d) uña
4. BORED: They are really **bored**.
a) aburrido
b) ansioso
c) avergonzado
d) decepcionado
5. ASSISTANT: The **assistant** is here.
a) ayudante
b) comerciante
c) representante
d) suplente
6. WEDDING: This type of **wedding** is perfect.
a) boda
b) nacimiento
c) negocio
d) reunión
7. AGAIN: They need it **again**.
a) al revés
b) al mismo tiempo
c) de nuevo
d) después
8. CLOWN: The **clown** is here.
a) pañuelo
b) payaso
c) peine
d) ternero
9. ICE: This **ice** is perfect.
a) cerradura
b) hielo
c) isla
d) tecla
10. REFUSE: They want to **refuse** it today.
a) reconocer
b) recuperar
c) reducir
d) rehusar
11. PIG: This **pig** is new.
a) cerdo
b) enchufe
c) pipa
d) tarta
12. LITERATURE: This type of **literature** is new.
a) frase
b) literatura
c) líquido
d) palacio
13. BEFORE: They were teachers three years before.
a) al otro lado
b) antes
c) en el extranjero
d) en realidad
14. FORWARD: They want to go **forward**.
a) adelante
b) además
c) del mismo modo
d) solo
15. PAVEMENT: This type of **pavement** is new.
a) acera
b) cuero
c) etiqueta
d) pico
16. TERM: This **term** is perfect.
a) equipo
b) prueba
c) sistema
d) trimestre
17. MATE: This **mate** is new here.
a) carne
b) compañero
c) espejo
d) mono
18. PROJECT: This **project** is perfect.
a) producto
b) profesión
c) progreso
d) proyecto
19. CAMEL: This **camel** is new.
a) alfabeto
b) barbacoa
c) camello
d) coco
20. TABLET: The **tablet** is here.
a) jersey
b) tableta
c) trompeta
d) vocabulario
21. DESTINATION: This type of **destination** is new.
a) cine
b) destino
c) diccionario
d) documental
22. TIN: This type of **tin** is new.
a) comercio
b) hojalata
c) ladrón
d) maletero
23. WIDE: This is really **wide**.
a) amplio
b) digno
c) húmedo
d) salvaje
24. ANNOUNCE: They want to **announce** it today.
a) anunciar
b) arrestar
c) atraer
d) avanzar
25. CAMP: They **camp** very often.
a) acampar
b) atrapar
c) comprar
d) cuidar
26. SIZE: This **size** is new to me.
a) cielo
b) olor
c) sonrisa
d) tamaño
27. THAT: They need **that**.
a) cualquiera
b) eso
c) otro
d) todo
28. ALTHOUGH: I am happy **although** this is new to me.
a) aunque
b) a menos que
c) por lo tanto
d) sin embargo
29. CREATURE: This type of **creature** is new to me.
a) criatura
b) cultura
c) defensa
d) diseño
30. THINK: They **think** this is new.
a) comprobar
b) contestar
c) creer
d) decir
31. LAUGH: They **laugh** very often.
a) besar
b) dar patadas
c) mentir
d) reír
32. DISCOVER: They want to **discover** it today.
a) descubrir
b) decidir
c) hacer
d) matar
33. STEP: This **step** is new.
a) bastón
b) huelga
c) paso
d) tema

34. SMOOTH: This is really **smooth**.
 a) dolorido
 b) inteligente
 c) libre
 d) suave
35. SWITCH: This **switch** is new.
 a) camión
 b) éxito
 c) interruptor
 d) sueldo
36. YOU: They need **you** today.
 a) le
 b) lo
 c) nos
 d) te
37. PEOPLE: This type of **people** is new here.
 a) dientes
 b) noticias
 c) personas
 d) ropas
38. SECTION: This **section** is new.
 a) escena
 b) esquí
 c) sección
 d) señal
39. BRAVE: They are very **brave**.
 a) despierto
 b) listo
 c) rubio
 d) valiente
40. LOCAL: They are **local** schools.
 a) especial
 b) histórico
 c) local
 d) necesario
41. CONNECTION: The **connection** is here.
 a) campeón
 b) conexión
 c) corrección
 d) elección
42. IMPROVE: They **improve** very often.
 a) aumentar
 b) combinar
 c) herir
 d) mejorar
43. BIKE: This **bike** is new.
 a) batería
 b) bicicleta
 c) murciélago
 d) puente
44. DIFFICULT: This is really **difficult**.
 a) delicioso
 b) diferente
 c) difícil
 d) peligroso
45. ROAD: This **road** is perfect.
 a) asiento
 b) carretera
 c) regla
 d) roca
46. STAGE: This **stage** is new.
 a) asunto
 b) escenario
 c) gusto
 d) impuesto
47. RECORDING: This **recording** is new.
 a) grabación
 b) informe
 c) pensamiento
 d) razón
48. DISEASE: This **disease** is new.
 a) enfermedad
 b) muñeca
 c) sobre
 d) tambor
49. DUST: The **dust** is here.
 a) multitud
 b) pato
 c) polvo
 d) techo
50. CONFIDENT: They are really **confident**.
 a) confiado
 b) crudo
 c) frecuente
 d) preciso
51. SHUT: They **shut** it very often.
 a) cerrar
 b) gastar
 c) gritar
 d) rasgar
52. CABIN: The **cabin** is here.
 a) cabaña
 b) contable
 c) lavabo
 d) taxi
53. ELEVATOR: This **elevator** is new.
 a) ascensor
 b) combustible
 c) experimento
 d) puerto
54. IDEA: This **idea** is perfect.
 a) calefacción
 b) grupo
 c) historia
 d) idea
55. HEY: **Hey**, Peter! How are you?
 a) ¡Guay!
 b) ¡Oye!
 c) ¡Vale!
 d) ¡Vaya!
56. EFFORT: This **effort** is new.
 a) botón
 b) capítulo
 c) cesta
 d) esfuerzo
57. PLEASANT: They are really **pleasant**.
 a) afilado
 b) agradable
 c) educado
 d) satisfecho
58. WEST: This is the **west** coast of the country.
 a) este
 b) norte
 c) oeste
 d) sur
59. HANG: They want to **hang** them today.
 a) alimentar
 b) colgar
 c) congelar
 d) pegar
60. DAILY: They need it **daily**.
 a) a veces
 b) absolutamente
 c) con cuidado
 d) diariamente
61. LABORATORY: This **laboratory** is new.
 a) biblioteca
 b) laboratorio
 c) liga
 d) revista
62. HAVE: They **have** done it.
 a) conseguir
 b) haber
 c) poder
 d) tener
63. DELAY: This **delay** is new.
 a) derrota
 b) ensayo
 c) muestra
 d) retraso
64. HAIRCUT: This type of **haircut** is perfect for me.
 a) corte de pelo
 b) pelo
 c) peluquero
 d) secador de pelo
65. IMPROVEMENT: This type of **improvement** is new.
 a) alquiler
 b) ausencia
 c) hierro
 d) mejora

66. RELAX: They want to **relax** them today.
- recibir
 - relajar
 - repetir
 - reservar
67. CUSHION: This **cushion** is new.
- cartón
 - cojín
 - desafío
 - goma
68. CONSIST: They **consist** of parts.
- competir
 - consistir
 - discrepar
 - persuadir
69. MOUSE: This **mouse** is new.
- artista
 - cazuela
 - patrón
 - ratón
70. MEND: They want to **mend** them today.
- abrochar
 - arreglar
 - ordenar
 - subrayar
71. CONVERSATION: This type of **conversation** is new to me.
- colección
 - control
 - conversación
 - coste
72. SUBJECT: This **subject** is new.
- apoyo
 - asignatura
 - mantel
 - traje
73. LOW: They are very **low**.
- apropiado
 - bajo
 - grande
 - pobre
74. MUG: The mug is here.
- bigote
 - jarrón
 - melocotón
 - tazón
75. TIGHTS: The **tights** are here.
- estudios
 - gafas
 - medias
 - peniques
76. HEAR: They want to **hear** it today.
- aprender
 - esconder
 - golpear
 - oír
77. COLOUR: This **colour** is new.
- clase
 - color
 - ordenador
 - país
78. EXCUSE: This type of **excuse** is perfect for me.
- empleado
 - excusa
 - familia
 - padre
79. USE: This type of **use** is new.
- respuesta
 - uso
 - visita
 - voz
80. BROWN: This **brown** is perfect.
- gris
 - marrón
 - naranja
 - verde
81. APOLOGISE: They **apologise** very often.
- celebrar
 - derramar
 - disculpase
 - saludar
82. PAY: This **pay** is new.
- imagen
 - paga
 - plan
 - precio
83. IRONING: This **ironing** is new.
- concurso
 - devolución
 - planchado
 - vendaje
84. COLOUR: They **colour** it very often.
- abrir
 - aceptar
 - colorear
 - llegar
85. WALK: This type of **walk** is perfect for me.
- botella
 - pájaro
 - paseo
 - reloj
86. MOUNTAIN: The **mountain** is here.
- bailarín
 - esquina
 - montaña
 - peligro
87. DAD: This **dad** is perfect.
- copa
 - masкота
 - papá
 - sentimiento
88. ITEM: This **item** is new.
- culpa
 - elemento
 - mobiliario
 - rana
89. LAND: They want to **land** today.
- aterrizar
 - quemar
 - reservar
 - unirse
90. HANDLE: They **handle** it very often.
- batir
 - manejar
 - permitir
 - sostener
91. EXPLANATION: This type of **explanation** is perfect.
- cambio
 - explicación
 - información
 - miembro
92. REGARDS: The **regards** are here.
- aduanas
 - refrescos
 - saludos
 - servicios
93. LANGUAGE: This type of **language** is new to me.
- idioma
 - impresora
 - piscina
 - situación
94. EVERY: **Every** object here is perfect.
- algún
 - cada
 - este
 - ningún
95. PEPPER: The **pepper** is here.
- aceite
 - fresa
 - mantequilla
 - pimienta
96. THROW: They want to **throw** them today.
- añadir
 - cazar
 - lanzar
 - molestar

97. IMAGINATION: This type of **imagination** is perfect for me.
- ayuda
 - imaginación
 - vacación
 - vecino
98. SILENCE: This **silence** is perfect.
- ducha
 - plata
 - silencio
 - sociedad
99. DECISION: This **decision** is new.
- decisión
 - defensa
 - detective
 - escritorio
100. SIDE: This **side** is new.
- forma
 - lado
 - punto
 - vista
101. MATCH: This **match** is new.
- horno
 - magia
 - mapa
 - partido
102. LEMONADE: This type of **lemonade** is new to me.
- consonante
 - gimnasia
 - limonada
 - pingüino
103. THAN: They are better **than** my brother.
- con
 - para
 - que
 - sin
104. PLEASED: They are very **pleased**.
- capaz
 - enfadado
 - satisfecho
 - sorprendido
105. WAITER: The **waiter** is here.
- camarero
 - especia
 - espía
 - estatua
106. CHEEK: The **cheek** is here.
- acantilado
 - cajón
 - jaula
 - mejilla
107. BLOGGER: This type of **blogger** is new to me.
- bloguero
 - cheque
 - jirafa
 - yogur
108. FAST: They need it **fast**.
- abajo
 - deprisa
 - lejos
 - tarde
109. DEVELOPMENT: This **development** is new.
- decepción
 - desarrollo
 - intercambio
 - meta
110. FINAL: The **final** is here.
- edificio
 - experiencia
 - final
 - película
111. PEA: This type of **pea** is new.
- alubia
 - cebolla
 - guisante
 - lechuga
112. PERSON: This **person** is perfect.
- chico
 - jugador
 - persona
 - policía
113. FIRST: They need it **first**.
- cerca
 - finalmente
 - primero
 - temprano
114. TIRED: They are really **tired**.
- asustado
 - cansado
 - envejecido
 - inusual
115. JOB: This **job** is perfect for me.
- broma
 - cocina
 - líder
 - trabajo
116. HALL: This **hall** is new.
- hambre
 - juez
 - sala
 - sombrero
117. START: This **start** is new.
- calle
 - comienzo
 - humo
 - sonido
118. INCREDIBLE: They are really **incredible**.
- increíble
 - pacífico
 - pequeño
 - típico
119. MILLION: This **million** is perfect for me.
- millón
 - naturaleza
 - página
 - radio
120. BLOW: They want to **blow** them today.
- construir
 - romper
 - saltar
 - soplar
121. HEIGHT: This **height** is perfect.
- altura
 - beca
 - ejemplo
 - rodilla
122. NOW: They need it **now**.
- ahora
 - aparte
 - nunca
 - en total
123. TISSUE: This **tissue** is perfect.
- fábrica
 - pañuelo de papel
 - traducción
 - variedad
124. MESSY: They are really **messy**.
- deseoso
 - desordenado
 - disponible
 - tímido
125. PARTY: This **party** is perfect.
- año
 - azul
 - fiesta
 - mujer
126. SUNNY: They are really **sunny**.
- estúpido
 - mejor
 - perdido
 - soleado
127. PILL: This **pill** is perfect.
- mediodía
 - niebla
 - pastilla
 - seta

- 128.** VIDEO: The **video** is here.
- árbol
 - ciencia
 - pared
 - vídeo
- 129.** HOMEPAGE: This type of **homepage** is new to me.
- luna de miel
 - maleta
 - página de inicio
 - poste
- 130.** HILL: The **hill** is here.
- agujero
 - colina
 - equipaje
 - esperanza
- 131.** REGRET: They **regret** them very often.
- aconsejar
 - desear
 - lamentar
 - lograr
- 132.** CABINET: This type of **cabinet** is perfect.
- armario
 - moneda
 - tasa
 - tripulación
- 133.** PEDRESTRIAN: The **pedestrian** is here.
- paquete
 - patinaje
 - peatón
 - postre
- 134.** GLOVE: The **glove** is here.
- abrazo
 - cabra
 - guante
 - hoja
- 135.** PUNISH: They want to **punish** them today.
- animar
 - castigar
 - reemplazar
 - situuar
- 136.** UNKIND: They are really **unkind**.
- descuidado
 - incapaz
 - inconsciente
 - poco amable
- 137.** BLOOD: The **blood** is here.
- llegada
 - sangre
 - trozo
 - zona
- 138.** HANDWRITING: This type of **handwriting** is perfect.
- bolsa de mano
 - caligrafía
 - tablón de anuncios
 - títular
- 139.** PROFESSOR: This **professor** is new.
- arquitectura
 - frase
 - profesor
 - repcionista
- 140.** BACKGROUND: This **background** is perfect.
- actitud
 - fondo
 - grupo
 - promedio
- 141.** UNDERGROUND: This is the **underground** part.
- antiguo
 - completado
 - desconocido
 - subterráneo
- 142.** WARN: They want to **warn** them today.
- advertir
 - adivinar
 - amenazar
 - recomendar
- 143.** SOMEHOW: They need them **somehow**.
- casi nunca
 - de alguna forma
 - en algún lugar
 - por error
- 144.** NET: This **net** is perfect.
- enlace
 - premio
 - rango
 - red
- 145.** THUMB: The **thumb** is here.
- barbilla
 - codo
 - pulgar
 - tobillo
- 146.** BROCHURE: This type of **brochure** is perfect.
- bufanda
 - folleto
 - investigación
 - monedero
- 147.** CENT: The **cent** is here.
- canal
 - céntimo
 - círculo
 - costa
- 148.** CHARMING: They are really **charming**.
- asqueroso
 - bochornoso
 - encantador
 - precioso
- 149.** IN: They are **in** it.
- al otro lado de
 - dentro de
 - encima de
 - frente a
- 150.** WATER: This **water** is perfect.
- agua
 - guerra
 - tipo
 - verdad

APPENDIX 5 – Listening Vocabulary Test – 81 Items (October 2019)

This test has TWO PARTS. Each part will take you about 20 minutes to finish. It is very important that you do the two parts of the test and that you try to answer **ALL THE QUESTIONS** in the test. There are no negative marks for incorrect answers.

Listening Vocabulary Size Test – Listen to the recording and select the answer (a, b, c, OR d) with the closest Spanish translation to the key word in the question.

Example 1 – You will hear:

SCHOOL – This **school** is new.

- A. cama
- B. escuela
- C. parque
- D. supermercado

The closest translation for the target word that you have heard is '*escuela*', so the answer you have to mark is **B**.

Example 2 - You will hear:

PLAY – They **play** it very often.

- A. beber
- B. cocinar
- C. comer
- D. jugar

The closest translation for the target word that you have heard is '*jugar*', so the answer you have to mark is **D**.

Example 3 - You will hear:

STRONG – They are really **strong**.

- A. alto
- B. feliz
- C. fuerte
- D. rico

The closest translation for the target word that you have heard is '*fuerte*', so the answer you have to mark is **C**.

Example 4 - You will hear:

TODAY – They need it **today**.

- A. hoy
- B. siempre
- C. también
- D. todavía

The closest translation for the target word that you have heard is '*hoy*', so the answer you have to mark is **A**.

PRUEBA DE VOCABULARIO

Esta prueba tiene DOS PARTES. Terminar cada parte te llevará unos 20 minutos. Es muy importante que hagas las dos partes de la prueba y que intentes contestar TODAS LAS PREGUNTAS en la prueba. No hay puntos negativos por respuestas incorrectas.

Prueba de Comprensión Oral de Vocabulario – Escucha la grabación y selecciona la respuesta (a-d) con la traducción en español más próxima a la palabra clave de la pregunta

Ejemplo 1 – Escucharás:

SCHOOL – This **school** is new.

- A. cama
- B. escuela
- C. parque
- D. supermercado

La traducción más próxima a la palabra que has escuchado es '*escuela*', así que la respuesta que tienes que marcar es **B**.

Ejemplo 2 – Escucharás:

PLAY – They **play** it very often.

- A. beber
- B. cocinar
- C. comer
- D. jugar

La traducción más próxima a la palabra que has escuchado es '*jugar*', así que la respuesta que tienes que marcar es **D**.

Ejemplo 3 – Escucharás:

STRONG – They are really **strong**.

- A. alto
- B. feliz
- C. fuerte
- D. rico

La traducción más próxima a la palabra que has escuchado es '*fuerte*', así que la respuesta que tienes que marcar es **C**.

Ejemplo 4 – Escucharás:

TODAY – They need it **today**.

- A. hoy
- B. siempre
- C. también
- D. todavía

La traducción más próxima a la palabra que has escuchado es '*hoy*', así que la respuesta que tienes que marcar es **A**.

1. _____
a) cubo
b) entrada
c) factura
d) rama
2. _____
a) ayudante
b) comerciante
c) representante
d) suplente
3. _____
a) pañuelo
b) payaso
c) peine
d) ternero
4. _____
a) reconocer
b) recuperar
c) reducir
d) rehusar
5. _____
a) cerdo
b) enchufe
c) pipa
d) tarta
6. _____
a) adelante
b) además
c) del mismo modo
d) solo
7. _____
a) acera
b) cuero
c) etiqueta
d) pico
8. _____
a) equipo
b) prueba
c) sistema
d) trimestre
9. _____
a) carne
b) compañero
c) espejo
d) documental
10. _____
a) comercio
b) hojalata
c) ladrón
d) maletero
11. _____
a) amplio
b) digno
c) húmedo
d) salvaje
12. _____
a) aunque
b) a menos que
c) por lo tanto
d) sin embargo
13. _____
a) criatura
b) cultura
c) defensa
d) diseño
14. _____
a) besar
b) dar patadas
c) mentir
d) reír
15. _____
a) dolorido
b) inteligente
c) libre
d) suave
16. _____
a) camión
b) éxito
c) interruptor
d) sueldo
17. _____
a) lo
b) me
c) nos
d) te
18. _____
a) especial
b) histórico
c) local
d) necesario
19. _____
a) aumentar
b) combinar
c) herir
d) mejorar
20. _____
a) asunto
b) escenario
c) gusto
d) impuesto
21. _____
a) grabación
b) informe
c) pensamiento
d) razón
22. _____
a) enfermedad
b) muñeca
c) sobre
d) tambor
23. _____
a) multitud
b) pato
c) polvo
d) techo
24. _____
a) confiado
b) crudo
c) frecuente
d) preciso
25. _____
a) cerrar
b) gastar
c) gritar
d) rasgar
26. _____
a) cabaña
b) contable
c) lavabo
d) taxi
27. _____
a) ¡Guay!
b) ¡Oye!
c) ¡Vale!
d) ¡Vaya!
28. _____
a) botón
b) capítulo
c) cesta
d) esfuerzo
29. _____
a) afilado
b) agradable
c) disponible
d) educado
30. _____
a) este
b) norte
c) oeste
d) sur
31. _____
a) alimentar
b) colgar
c) congelar
d) pegar
32. _____
a) a veces
b) absolutamente
c) con cuidado
d) diariamente
33. _____
a) conseguir
b) haber
c) poder
d) ver

34. _____
a) derrota
b) ensayo
c) muestra
d) retraso
35. _____
a) alquiler
b) ausencia
c) hierro
d) mejora
36. _____
a) cartón
b) cojín
c) desafío
d) goma
37. _____
a) competir
b) consistir
c) discrepar
d) persuadir
38. _____
a) abrochar
b) arreglar
c) ordenar
d) subrayar
39. _____
a) apoyo
b) asignatura
c) mantel
d) traje
40. _____
a) apropiado
b) bajo
c) grande
d) pobre
41. _____
a) bigote
b) jarrón
c) melocotón
d) tazón
42. _____
a) estudios
b) gafas
c) medias
d) peniques
43. _____
a) imagen
b) paga
c) plan
d) precio
44. _____
a) concurso
b) devolución
c) planchado
d) vendaje
45. _____
a) botella
b) pájaro
c) paseo
d) reloj
46. _____
a) culpa
b) elemento
c) mobiliario
d) rana
47. _____
a) aterrizar
b) quemar
c) reservar
d) unirse
48. _____
a) batir
b) manejar
c) permitir
d) sostener
49. _____
a) aduanas
b) refrescos
c) saludos
d) servicios
50. _____
a) algún
b) cada
c) este
d) ningún
51. _____
a) añadir
b) cazar
c) lanzar
d) molestar
52. _____
a) forma
b) lado
c) punto
d) vista
53. _____
a) con
b) para
c) que
d) sin
54. _____
a) capaz
b) enfadado
c) satisfecho
d) sorprendido
55. _____
a) acantilado
b) cajón
c) jaula
d) mejilla
56. _____
a) abajo
b) deprimido
c) lejos
d) tarde
57. _____
a) decepción
b) desarrollo
c) intercambio
d) meta
58. _____
a) alubia
b) cebolla
c) guisante
d) lechuga
59. _____
a) hambre
b) juez
c) sala
d) sombrero
60. _____
a) construir
b) romper
c) saltar
d) soplar
61. _____
a) altura
b) beca
c) ejemplo
d) rodilla
62. _____
a) fábrica
b) pañuelo de papel
c) traducción
d) variedad
63. _____
a) deseoso
b) desordenado
c) roto
d) tímido
64. _____
a) mediodía
b) niebla
c) pastilla
d) seta
65. _____
a) luna de miel
b) maleta
c) página de inicio
d) poste
66. _____
a) agujero
b) colina
c) equipaje
d) esperanza

67. _____
a) aconsejar
b) desear
c) lamentar
d) lograr
68. _____
a) armario
b) moneda
c) tasa
d) tripulación
69. _____
a) paquete
b) patinaje
c) peatón
d) postre
70. _____
a) brazo
b) cabra
c) guante
d) hoja
71. _____
a) animar
b) castigar
c) reemplazar
d) situar
72. _____
a) descuidado
b) incapaz
c) inconsciente
d) poco amable
73. _____
a) bolsa de mano
b) caligrafía
c) tablón de anuncios
d) titular
74. _____
a) actitud
b) fondo
c) grupo
d) promedio
75. _____
a) advertir
b) adivinar
c) amenazar
d) recomendar
76. _____
a) casi nunca
b) de alguna forma
c) en algún lugar
d) por error
77. _____
a) enlace
b) premio
c) rango
d) red
78. _____
a) barbilla
b) codo
c) pulgar
d) tobillo
79. _____
a) bufanda
b) folleto
c) investigación
d) monedero
80. _____
a) asqueroso
b) bochornoso
c) encantador
d) precioso
81. _____
a) al otro lado de
b) dentro de
c) detrás de
d) frente a

APPENDIX 6 – Written Vocabulary Test – 81 Items (October 2019)

WRITTEN VOCABULARY SIZE TEST

This is the second part of the vocabulary test. Please, DO NOT CORRECT any answers in the previous test. It is also very important that you try to answer **ALL THE QUESTIONS** in the test. There are no negative marks for incorrect answers.

Read the questions and select the answer (a, b, c, OR d) with the closest Spanish translation to the key word in each question.

Example 1 – You will hear:

SCHOOL – This **school** is new.

- A. cama
- B. escuela
- C. parque
- D. supermercado

Example 2 - You will hear:

PLAY – They **play** it very often.

- A. beber
- B. cocinar
- C. comer
- D. jugar

The closest translation for the target word that you have heard is *'escuela'*, so the answer you have to mark is **B**.

The closest translation for the target word that you have heard is *'jugar'*, so the answer you have to mark is **D**.

1. TICKET: This **ticket** is perfect.

- a) cubo
- b) entrada
- c) factura
- d) rama

2. ASSISTANT: The **assistant** is here.

- a) ayudante
- b) comerciante
- c) representante
- d) suplente

3. CLOWN: The **clown** is here.

- a) pañuelo
- b) payaso
- c) peine
- d) ternero

4. REFUSE: They want to **refuse** it today.

- a) reconocer
- b) recuperar
- c) reducir
- d) rehusar

5. PIG: This **pig** is new.

- a) cerdo
- b) enchufe
- c) pipa
- d) tarta

6. FORWARD: They want to go **forward**.

- a) adelante
- b) además
- c) del mismo modo
- d) solo

7. PAVEMENT: This type of **pavement** is new.

- a) acera
- b) cuero
- c) etiqueta
- d) pico

8. TERM: This **term** is perfect.

- a) equipo
- b) prueba
- c) sistema
- d) trimestre

9. MATE: This **mate** is new here.

- a) carne
- b) compañero
- c) espejo
- d) modo

10. TIN: This type of **tin** is new.

- a) comercio
- b) hojalata
- c) ladrón
- d) maletero

11. WIDE: This is really **wide**.

- a) amplio
- b) digno
- c) húmedo
- d) salvaje

12. ALTHOUGH: I am happy **although** this is new to me.

- a) aunque
- b) a menos que
- c) por lo tanto
- d) sin embargo

13. CREATURE: This type of **creature** is new to me.

- a) criatura
- b) cultura
- c) defensa
- d) diseño

14. LAUGH: They **laugh** very often.

- a) besar
- b) dar patadas
- c) mentir
- d) reír

15. SMOOTH: This is really **smooth**.

- a) dolorido
- b) inteligente
- c) libre
- d) suave

16. SWITCH: This **switch** is new.

- a) camión
- b) éxito
- c) interruptor
- d) sueldo

17. YOU: They need **you** today.

- a) lo
- b) me
- c) nos
- d) te

18. LOCAL: They are **local** schools.

- a) especial
- b) histórico
- c) local
- d) necesario

19. IMPROVE: They **improve** very often.
 a) aumentar
 b) combinar
 c) herir
 d) mejorar
20. STAGE: This **stage** is new.
 a) asunto
 b) escenario
 c) gusto
 d) impuesto
21. RECORDING: This **recording** is new.
 a) grabación
 b) informe
 c) pensamiento
 d) razón
22. DISEASE: This **disease** is new.
 a) enfermedad
 b) muñeca
 c) sobre
 d) tambor
23. DUST: The **dust** is here.
 a) multitud
 b) pato
 c) polvo
 d) techo
24. CONFIDENT: They are really **confident**.
 a) confiado
 b) crudo
 c) frecuente
 d) preciso
25. SHUT: They **shut** it very often.
 a) cerrar
 b) gastar
 c) gritar
 d) rasgar
26. CABIN: The **cabin** is here.
 a) cabaña
 b) contable
 c) lavabo
 d) taxi
27. HEY: **Hey**, Peter! How are you?
 a) ¡Guay!
 b) ¡Oye!
 c) ¡Vale!
 d) ¡Vaya!
28. EFFORT: This **effort** is new.
 a) botón
 b) capítulo
 c) cesta
 d) esfuerzo
29. PLEASANT: They are really **pleasant**.
 a) afilado
 b) agradable
 c) disponible
 d) educado
30. WEST: This is the **west** coast of the country.
 a) este
 b) norte
 c) oeste
 d) sur
31. HANG: They want to **hang** them today.
 a) alimentar
 b) colgar
 c) congelar
 d) pegar
32. DAILY: They need it **daily**.
 a) a veces
 b) absolutamente
 c) con cuidado
 d) diariamente
33. HAVE: They **have** done it.
 a) conseguir
 b) haber
 c) poder
 d) ver
34. DELAY: This **delay** is new.
 a) derrota
 b) ensayo
 c) muestra
 d) retraso
35. IMPROVEMENT: This type of **improvement** is new.
 a) alquiler
 b) ausencia
 c) hierro
 d) mejora
36. CUSHION: This **cushion** is new.
 a) cartón
 b) cojín
 c) desafío
 d) goma
37. CONSIST: They **consist** of parts.
 a) competir
 b) consistir
 c) discrepar
 d) persuadir
38. MEND: They want to **mend** them today.
 a) abrochar
 b) arreglar
 c) ordenar
 d) subrayar
39. SUBJECT: This **subject** is new.
 a) apoyo
 b) asignatura
 c) mantel
 d) traje
40. LOW: They are very **low**.
 a) apropiado
 b) bajo
 c) grande
 d) pobre
41. MUG: The **mug** is here.
 a) bigote
 b) jarrón
 c) melocotón
 d) tazón
42. TIGHTS: The **tights** are here.
 a) estudios
 b) gafas
 c) medias
 d) peniques
43. PAY: This pay is **new**.
 a) imagen
 b) paga
 c) plan
 d) precio
44. IRONING: This **ironing** is new.
 a) concurso
 b) devolución
 c) planchado
 d) vendaje
45. WALK: This type of **walk** is perfect for me.
 a) botella
 b) pájaro
 c) paseo
 d) reloj
46. ITEM: This **item** is new.
 a) culpa
 b) elemento
 c) mobiliario
 d) rana
47. LAND: They want to **land** today.
 a) aterrizar
 b) quemar
 c) reservar
 d) unirse
48. HANDLE: They **handle** it very often.
 a) batir
 b) manejar
 c) permitir
 d) sostener

49. REGARDS: The **regards** are here.
 a) aduanas
 b) refrescos
 c) saludos
 d) servicios
50. EVERY: **Every** object here is perfect.
 a) algún
 b) cada
 c) este
 d) ningún
51. THROW: They want to **throw** them today.
 a) añadir
 b) cazar
 c) lanzar
 d) molestar
52. SIDE: This **side** is new.
 a) forma
 b) lado
 c) punto
 d) vista
53. THAN: They are better **than** my brother.
 a) con
 b) para
 c) que
 d) sin
54. PLEASED: They are very **pleased**.
 a) capaz
 b) enfadado
 c) satisfecho
 d) sorprendido
55. CHEEK: The **cheek** is here.
 a) acantilado
 b) cajón
 c) jaula
 d) mejilla
56. FAST: They need it **fast**.
 a) abajo
 b) deprisa
 c) lejos
 d) tarde
57. DEVELOPMENT: This **development** is new.
 a) decepción
 b) desarrollo
 c) intercambio
 d) meta
58. PEA: This type of **pea** is new.
 a) alubia
 b) cebolla
 c) guisante
 d) lechuga
59. HALL: This **hall** is new.
 a) hambre
 b) juez
 c) sala
 d) sombrero
60. BLOW: They want to **blow** them today.
 a) construir
 b) romper
 c) saltar
 d) soplar
61. HEIGHT: This **height** is perfect.
 a) altura
 b) beca
 c) ejemplo
 d) rodilla
62. TISSUE: This **tissue** is perfect.
 a) fábrica
 b) pañuelo de papel
 c) traducción
 d) variedad
63. MESSY: They are really **messy**.
 a) deseoso
 b) desordenado
 c) roto
 d) tímido
64. PILL: This **pill** is perfect.
 a) mediodía
 b) niebla
 c) pastilla
 d) seta
65. HOMEPAGE: This type of **homepage** is new to me.
 a) luna de miel
 b) maleta
 c) página de inicio
 d) poste
66. HILL: The **hill** is here.
 a) agujero
 b) colina
 c) equipaje
 d) esperanza
67. REGRET: They **regret** them very often.
 a) aconsejar
 b) desear
 c) lamentar
 d) lograr
68. CABINET: This type of **cabinet** is perfect.
 a) armario
 b) moneda
 c) tasa
 d) tripulación
69. PEDRESTRIAN: The **pedestrian** is here.
 a) paquete
 b) patinaje
 c) peatón
 d) postre
70. GLOVE: The **glove** is here.
 a) abrazo
 b) cabra
 c) guante
 d) hoja
71. PUNISH: They want to **punish** them today.
 a) animar
 b) castigar
 c) reemplazar
 d) situar
72. UNKIND: They are really **unkind**.
 a) descuidado
 b) incapaz
 c) inconsciente
 d) poco amable
73. HANDWRITING: This type of **handwriting** is perfect.
 a) bolsa de mano
 b) caligrafía
 c) tablón de anuncios
 d) titular
74. BACKGROUND: This **background** is perfect.
 a) actitud
 b) fondo
 c) grupo
 d) promedio
75. WARN: They want to **warn** them today.
 a) advertir
 b) adivinar
 c) amenazar
 d) recomendar
76. SOMEHOW: They need them **somehow**.
 a) casi nunca
 b) de alguna forma
 c) en algún lugar
 d) por error
77. NET: This **net** is perfect.
 a) enlace
 b) premio
 c) rango
 d) red
78. THUMB: The **thumb** is here.
 a) barbilla
 b) codo
 c) pulgar
 d) tobillo

79. BROCHURE: This **type** of **brochure** is perfect.

- a) bufanda
- b) folleto
- c) investigación
- d) monedero

80. CHARMING: They are really **charming**.

- a) asqueroso
- b) bochornoso
- c) encantador
- d) precioso

81. IN: They are **in** it.

- a) al otro lado de
- b) dentro de
- c) detrás de
- d) frente a

APPENDIX 7 – PET Vocabulary List (retrieved from Cambridge Assessment English in PDF format)

A

a/an (det)	activity (n)
ability (n)	actor (n)
able (adj)	actress (n)
• be able to	actually (adv)
about (adv & prep)	• She seems a bit strict at first, but she's actually very nice.
• about 500 students (adv)	• Are you actually going to take
• The film is about a small boy. (prep)	the job?
above (adj, adv & prep)	ad (advertisement) (n)
abroad (adv)	add (v)
absent (adj)	addition (n)
absolutely (adv)	• in addition
• The movie was absolutely awful.	address (n)
accent(n)	admire (v)
• She has a beautiful French accent.	admission (n)
accept (v)	• charges/cost/price
access (n)	admit (v)
• disabled access	adult (adj & n)
• internet access	advance (n)
accident (n)	• book in advance
accommodation (n)	advanced (adj)
accompany (v)	advantage (n)
according to (prep phr)	adventure (n)
account (n)	advert (n)
accountant (n)	advertise (v)
accurate (adj)	advertisement (n)
ache (n)	advice (n)
	advise (v)
	aeroplane (n)

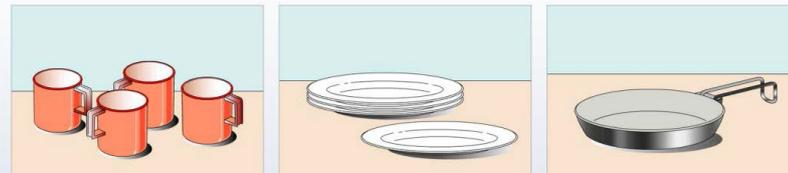
APPENDIX 8 – Listening Comprehension Test – Part 1 (original test retrieved from Cambridge Assessment English in PDF format)

LISTENING TEST – PART 1 - You will hear seven short recordings. For questions 1-7, choose the correct picture and mark it on your answer sheet. You will hear each recording TWICE.

1. What has the girl bought today?



2. What have they forgotten?



3. How will the girl get home?



4. Which room are the flowers in?



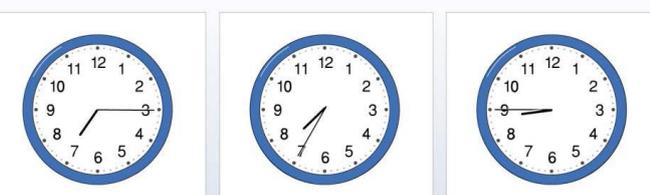
5. What is at the art gallery this week?



6. Which is the woman's suitcase?



7. What time does the woman's flight leave?



APPENDIX 9 – Listening Comprehension Test – Part 2 (original test retrieved from Cambridge Assessment English in PDF format)

LISTENING TEST – PART 2 - You will hear a radio interview with Darren Hubberd, a runner who takes part in athletics competitions. For questions 8-13 choose the correct answer and mark it on your answer sheet. You will hear each recording TWICE.

<p>8. At the February competition, Darren</p>	<p>a) ran in a new event. b) hurt himself. c) came last.</p>
<p>9. Darren's situation began to improve when he</p>	<p>a) started a job with fewer hours. b) was offered a place on the British team. c) signed contract with a sportswear company.</p>
<p>10. Darren got fit again quickly because he</p>	<p>a) changed the way he trained. b) started to work with a new trainer. c) increased the time he spends training.</p>
<p>11. Darren wants to win his next athletics competition so that he</p>	<p>a) can retire early. b) can pay for his wedding. c) can show people that he is fit.</p>
<p>12. In the next competition, Darren will run the 400-metre race on</p>	<p>a) the first day. b) the second day. c) the third day.</p>
<p>13. In the future, Darren</p>	<p>a) hopes to write about his career. b) wants to change the distance he runs. c) would like more people to recognise him.</p>

APPENDIX 10 – Listening Comprehension Test – Part 3 (original test retrieved from Cambridge Assessment English in PDF format)

LISTENING TEST – PART 3 - You will hear a man giving details about a photography competition. For questions 14-19, write the correct answer in the gap on your answer sheet. You will hear the recording TWICE.

PHOTOGRAPY OF THE YEAR COMPETITION

First prize: £2,000 and a painting of ⁽¹⁴⁾ _____ by John Stevens.

Second prize: £1,000 and camera equipment worth £200

Competition closing date: ⁽¹⁵⁾ _____.

Subjects:

1. British Nature
2. Wild Places
3. Animals at ⁽¹⁶⁾ _____.

Exhibition: Victoria Museum

Countries which the exhibition will tour: UK, USA, ⁽¹⁷⁾ _____ and Japan.

To enter, write to

Radio TYL

63 ⁽¹⁸⁾ _____ Road.

London

6TY 9JN

Tel: ⁽¹⁹⁾ _____.

APPENDIX 11 – Listening Comprehension Test – Part 4 (original test retrieved from Cambridge Assessment English in PDF format)

LISTENING TEST – PART 4 - You will hear a boy called Jack and a girl called Helen, talking about a rock festival. Decide if each sentence is correct or not. If it is correct, select YES. If the sentence is not correct, select NO.

20. The festival was better than Jack expected it to be.	YES / NO
21. Helen bought her ticket for the festival in advance.	YES / NO
22. Jack was disappointed that he had to change his plans.	YES / NO
23. Helen complains about having to wait a long time for food.	YES / NO
24. They both say that it was the sunshine that made the afternoon enjoyable.	YES / NO
25. Jack prefers listening to loud bands.	YES / NO

APPENDIX 12 – Listening Comprehension Test – Edited Transcript (original transcript retrieved from Cambridge Assessment English in PDF format)

1: *What has the girl bought today?*

- Oh ... you've been to the duty-free shop, what did you get? Perfume?
- You must be joking. It costs much less at the supermarket at home. There was some nice jewellery, but what was really good value was this T-shirt ... look.
- Oh ... £4.50, well that's cheaper than the box of chocolates you bought last year anyway.

2: *What have they forgotten?*

- Now we've put the tent up, let's make something to drink. I'll get the cups. They're in the plastic bag in the back of the car, aren't they?
- No, that's got the new frying pan in it. You packed the cups in the box with the plates.
- Ah yes, that's right. Here they are. But I can't see the plastic bag anywhere.
- Oh dear, we've left it behind, so we can't cook anything. Well, we can still have a cup of tea.

3: *How will the girl get home?*

- ... Hi Mum, it's me ... it's all right, I'm not phoning for a lift ... I am going to be late though ... Mmm ... when I got to the railway station I found the seven o'clock was cancelled, so I'll just wait for the next one – there aren't any buses at this time of night. See you soon, I hope ... Next time I'll go by bike!

4: *Which room are the flowers in?*

Hi! I'm home. Oh, where have you put the flowers that Robin bought me? I left them on the table here in the hall with some letters I need to post.

Well, they were in the way there, so I've put them in a jug in the bedroom.

Okay thanks, but I think I'll put them in the kitchen. They'll look nicer there. Would you like a cup of coffee?

Umm. That sounds good!

5: *What is at the art gallery this week?*

- Thank you for calling the Central Art Gallery. This week, and next, there is a special exhibition of paintings by a local artist, John Temple, on the subject of 'Growing Old'. He is now quite well known and we hope this exhibition will be even more popular than his last one on 'Animals in the Wild'. Next week we will also have a small exhibition of children's paintings of the seaside.

6: *Which is the woman's suitcase?*

- Good afternoon Madam, I understand you've lost a piece of luggage. Could you describe it to me please?

- Yes, it's a small black suitcase, with a set of wheels at one end and a metal handle which pulls out of the other end, so you can pull it along.

7: *What time does the woman's flight leave?*

- Excuse me, I've come to the airport rather early. I'm booked on flight number 645 to London which leaves at 8.45. I've got these two heavy bags, and the check-in time isn't until 7.35. Would it be possible to check them in a little earlier?

- I'm sorry Madam, but there's nobody here from that company yet. They usually come in at about 7.15. Perhaps you can come back then?

You will hear a radio interview with Darren Hubbard, a runner who takes part in athletics competitions.

- Our next guest is the runner Darren Hubbard. Darren, the year started badly for you.
- It did. In the February competition I was running in my normal events, the 200, 400 and 800-metre races. I'd done quite badly in the first race – though I wasn't last – but the problems really began with the 800 metres. During the race I was injured, and it took me quite a while to recover.
- When did things start to get better?

- In the summer, really. I was disappointed because I hadn't got into the British team but then I was offered a contract with a Japanese company that makes running shoes. The money meant I could stop work. I'd only been working part-time in a shop but, as you know, this can make things quite difficult for athletes. I accepted the contract immediately.

- Has it taken long to get fit again?

- No – not long because I now do some different exercises as part of my training. For example, we've introduced swimming and weight-training into my programme. I've had the same trainer since I started running, and I still train for 5 hours a day as before but, of course, I don't have to fit that in around work any more.

- So you're confident about the next competition, then?

- Yes. I don't have any plans to retire! I've been in other races since February and I've already proved that I'm fit. But the next competition is important to me. I'm hoping to get married soon and the prize money would be very useful to pay for the celebrations. In fact, it will be very difficult without it.

- Which races are you in?

- On day one, I start with the 800 metres and the following day there's the 400 metres. That's the race I'm most confident about. I'll finish with the 200 metres on day three.

- And what are you hoping the future will bring?

- I'm aiming to get faster at the distances I run. That's one thing. And, although I don't want to be really famous, I mean, I don't want the newspapers writing about me all the time, I would like to get to the point where I walk down the street and everybody says 'There's Darren!' Yes, I'd quite like that.

- Well, good luck with that Darren, and thank you for joining us...

You will hear a radio announcer giving details about a photography competition

- Now, this morning I'd like to tell you about this year's competition for the best photograph of animals, birds or plants. We have some great prizes for you – first prize for the most original photo is a cheque for £2,000 and a picture of elephants painted by the artist John Stevens. The second prize is £1,000 and camera equipment worth £200. The lucky winner will receive his or her prize in London on 16th October this year. So, all you photographers, get your cameras and start taking some great photographs, as you must send them to us by 14th May. Now for the details. You can enter up to three colour photographs in each of the following areas. First of all, British Nature. For this your photos must only include plants or animals which are found living in Britain. Secondly, Wild Places. Your photos should be of lonely places. And finally, our third subject is Animals at Night. Pictures must be taken between sunset and sunrise and must include animals. All the winning photographs can be seen in a special exhibition at the Victoria Museum in London, from the end of November until January next year. The exhibition will tour the UK and the USA in the spring, followed by France and Japan during the summer. Remember, the judges want to see some original ideas – they don't want photos of pets or animals in zoos. Now, to enter, the first thing you should do is contact us to get an application form. Our address is Radio TYL, 63 Beechwood Road, that's spelled B E E C H W O O D, Road, London 6 9. Of course, if you have any questions about the competition we'll be glad to hear from you. You can either telephone us on 0163-55934 or fax us on 0163-33298.

You will hear a boy called Jack and a girl called Helen, talking about a rock festival.

- Hi Jack, how are you?

- Fine, Helen. Did you go to the rock festival last Saturday? I didn't see you there.

- Well, there *were* lots of people! It was great, wasn't it?

- Well, one or two bands were brilliant, yes, but I have to say it *wasn't* as good as I thought it would be.

- Oh, why's that? - Well, perhaps I expected too much ... It did cost a lot of money to get in – £20.

- Didn't you book early? My ticket was much less.

- But you had to buy that so long ago!

- So?
- Well, I mean until last Wednesday I thought I wasn't even going to the festival.
- Oh that's right. You were supposed to go to Canada, weren't you? I'm sorry that didn't happen.
- Don't remind me about it! ... I doubt if I'll ever get the same chance again.
- I'm sure you will, Jack. Anyway ... talking about the festival, what did you think of the food there?
- It wasn't bad.
- So much choice, especially for vegetarians like me ... and there never seemed to be many queues.
- Mmm. You know, I *did* enjoy the afternoon ...
- Yes, that was the best thing, wasn't it, when it got really sunny?
- Did it? I didn't notice! That's when my favourite band were playing.
- Flashbang? They had a problem with their sound system, didn't they? I had to cover my ears at one point.
- Helen, it's supposed to be like that! That's what so good about them ... the drums were like thunder. It's my favourite kind of music.
- Well, that wouldn't be my choice, Jack.
- So what did you like best then?
- Oh, Maria Crevel – definitely – she sang so beautifully ...

APPENDIX 13 – Preliminary Study – Tables with fit statistics – May 2019

Table 3.12 – Preliminary Study (May 2019) – Persons in the listening vocabulary test with highest misfit values

PERSON	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
P37	72	81	2.65	.39	1.15	.65	2.30	1.83
P20	76	81	3.39	.49	.94	-.04	1.77	1.04
P70	65	81	1.82	.31	1.19	1.05	1.63	1.53
P31	79	81	4.43	.74	1.12	.40	1.57	.81
P6	58	81	1.20	.28	1.13	.91	1.52	1.78
P39	71	81	2.51	.37	1.03	.20	1.49	.97
P30	53	81	.83	.27	1.27	2.04	1.47	1.95
P13	58	81	1.20	.28	1.24	1.66	1.41	1.46
P17	56	81	1.05	.27	1.06	.46	1.40	1.54
P14	69	81	2.25	.35	1.12	.63	1.38	.87
P62	45	81	.28	.26	1.11	1.02	1.37	1.95
P43	63	81	1.63	.30	1.07	.47	1.32	.99
P32	52	81	.76	.27	1.15	1.23	1.30	1.39
P71	44	81	.22	.26	1.13	1.20	1.29	1.61
P33	61	81	1.45	.29	1.10	.66	1.26	.90
P73	61	81	1.45	.29	1.17	1.12	1.25	.85
P49	60	81	1.37	.29	1.13	.86	1.21	.79
BETTER FITTING PARTICIPANTS NOT SHOWN								
P40	63	81	1.63	.30	.75	-1.64	.59	-1.33
P51	69	81	2.25	.35	.75	-1.24	.41	-1.45
P28	77	81	3.66	.54	.73	-.56	.20	-1.06
P3	66	81	1.92	.32	.70	-1.76	.50	-1.42
P2	67	81	2.02	.33	.87	-.62	.68	-.72
P21	72	81	2.65	.39	.86	-.50	.54	-.75
P26	75	81	3.17	.45	.84	-.40	.71	-.19
P45	77	81	3.66	.54	.84	-.27	.31	-.75
P55	62	81	1.54	.30	.84	-1.00	.78	-.64
P59	67	81	2.02	.33	.83	-.90	.58	-1.03
P1	60	81	1.37	.29	.80	-1.35	.71	-1.01
P22	73	81	2.80	.40	.80	-.70	.55	-.63

Table 3.13 – Preliminary Study (May 2019) – Persons in the written vocabulary test with highest misfit values

PERSON	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
P20	77	81	3.85	.55	.97	.07	4.70	2.28
P6	73	81	2.96	.42	1.41	1.49	3.14	2.19
P66	71	81	2.65	.38	1.26	1.10	2.55	2.02
P5	79	81	4.64	.74	1.17	.47	2.37	1.23
P34	67	81	2.13	.34	1.29	1.39	2.00	1.86
P70	69	81	2.37	.36	1.41	1.77	1.76	1.36
P30	60	81	1.41	.30	1.23	1.36	1.72	2.03
P27	78	81	4.18	.62	1.12	.40	1.69	.89
P41	75	81	3.35	.46	1.19	.67	1.61	.88
P17	64	81	1.80	.32	1.17	.93	1.51	1.29
P53	77	81	3.85	.55	1.13	.44	1.49	.75
P13	64	81	1.80	.32	1.12	.67	1.45	1.16
P7	60	81	1.41	.30	1.29	1.65	1.27	.91
P62	40	81	-.13	.26	1.15	1.25	1.27	1.33
P50	69	81	2.37	.36	1.13	.64	1.25	.61
P67	75	81	3.35	.46	1.25	.85	1.25	.55
BETTER FITTING PARTICIPANTS NOT SHOWN								
P22	78	81	4.18	.62	.84	-.19	.20	-.76
P29	78	81	4.18	.62	.84	-.18	.20	-.77
P52	78	81	4.18	.62	.84	-.18	.20	-.77
P63	73	81	2.96	.42	.84	-.54	.31	-1.11
P72	56	81	1.06	.29	.82	-1.23	.71	-1.18
P10	69	81	2.37	.36	.81	-.87	.48	-1.08
P59	70	81	2.51	.37	.81	-.85	.48	-.96
P21	72	81	2.80	.40	.80	-.81	.40	-1.00
P45	75	81	3.35	.46	.80	-.60	.57	-.29
P12	71	81	2.65	.38	.79	-.90	.45	-.95
P51	69	81	2.37	.36	.69	-1.58	.34	-1.58
P3	66	81	2.01	.33	.68	-1.83	.68	-.71
P73	69	81	2.37	.36	.63	-1.95	.31	-1.67

APPENDIX 14 – MAIN STUDY – FIRST DATA GATHERING – TABLES WITH FIT STATISTICS – OCTOBER 2019

Table 3.20 – Main Study (October 2019) – Items in the listening vocabulary test with highest misfit values (shaded cells represent flagged items)

ITEM	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
L51	33	284	2.90	.19	1.10	.77	1.51	2.22
L52	177	284	.01	.13	1.28	5.47	1.45	5.57
L108	238	284	-1.26	.17	1.08	.82	1.39	2.15
L5	263	284	-2.20	.23	1.02	.17	1.34	1.19
L1	185	284	-.13	.13	1.16	3.10	1.30	3.48
L55	264	284	-2.26	.24	1.02	.19	1.25	.88
L70	64	284	2.01	.15	1.13	1.54	1.24	1.93
L132	117	284	1.00	.13	1.17	3.44	1.24	3.62
L121	240	284	-1.32	.17	1.05	.53	1.17	.98
BETTER FITTING ITEMS NOT SHOWN								
L136	112	284	1.08	.13	.90	-2.11	.88	-1.96
L31	205	284	-.50	.14	.89	-1.82	.83	-1.76
L88	180	284	-.05	.13	.89	-2.41	.86	-1.93
L89	195	284	-.31	.14	.88	-2.28	.83	-1.94
L135	194	284	-.30	.14	.88	-2.18	.80	-2.37
L63	142	284	.59	.13	.87	-3.12	.84	-3.06
L130	204	284	-.48	.14	.87	-2.21	.81	-1.99
L49	149	284	.47	.13	.86	-3.44	.83	-3.19
L65	193	284	-.28	.14	.86	-2.71	.77	-2.81

Table 3.21 – Main Study (October 2019) – Items in the written vocabulary test with highest misfit values (shaded cells represent flagged items)

ITEM	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
W52	142	283	1.30	.13	1.47	9.44	1.72	9.51
W1	217	283	-.07	.15	1.23	2.83	1.58	3.77
W108	262	283	-1.51	.23	1.05	.34	1.41	1.30
W51	46	283	3.23	.17	1.17	1.49	1.39	2.14
W16	106	283	1.91	.13	1.23	4.08	1.33	4.13
W5	261	283	-1.46	.23	1.03	.25	1.31	1.05
W58	212	283	.04	.15	1.11	1.51	1.28	2.11
W121	221	281	-.19	.15	1.11	1.34	1.28	1.81
W62	205	283	.19	.14	1.11	1.69	1.27	2.27
W29	242	283	-.71	.18	1.07	.64	1.24	1.21
W70	72	283	2.57	.15	1.14	1.79	1.21	1.82
W132	137	280	1.37	.13	1.15	3.21	1.19	2.93
BETTER FITTING ITEMS NOT SHOWN								
W19	208	283	.13	.14	.91	-1.39	.86	-1.19
W53	266	283	-1.75	.26	.91	-.36	.72	-.80
W57	219	281	-.14	.15	.91	-1.18	.81	-1.44
W65	265	281	-1.81	.26	.91	-.36	.58	-1.33
W14	240	283	-.65	.17	.90	-.91	.70	-1.72
W34	168	283	.86	.13	.90	-2.33	.86	-1.97
W10	191	283	.46	.14	.89	-1.98	.85	-1.69
W35	228	283	-.32	.16	.89	-1.28	.72	-1.98
W47	218	283	-.09	.15	.86	-1.85	.78	-1.74
W66	222	281	-.21	.15	.85	-1.86	.75	-1.83
W71	210	280	.05	.15	.82	-2.60	.69	-2.72
W19	208	283	.13	.14	.91	-1.39	.86	-1.19

Table 3.22 – Main Study (October 2019) – Items in the listening comprehension test with highest misfit values (shaded cells represent flagged items)

ITEM	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
LISTEN	245	284	-2.08	.19	1.07	.62	1.64	2.41
LISTEN	171	284	-.38	.13	1.20	3.67	1.31	3.05
LISTEN	126	284	.41	.13	.99	-.14	1.27	2.87
LISTEN	120	284	.51	.13	1.15	2.80	1.20	2.10
LISTEN	42	284	2.25	.18	1.02	.24	1.18	.86
LISTEN	178	284	-.51	.14	1.14	2.48	1.16	1.59
LISTEN	144	284	.09	.13	1.00	.05	1.14	1.69
LISTEN	123	284	.46	.13	1.03	.52	1.10	1.14
LISTEN	199	284	-.91	.14	1.04	.67	1.08	.69
LISTEN	78	284	1.32	.15	1.03	.47	1.07	.56
LISTEN	16	284	3.46	.27	1.03	.23	.84	-.31
LISTEN	165	284	-.27	.13	1.02	.32	1.02	.25
LISTEN	204	284	-1.01	.14	.98	-.26	1.01	.10
LISTEN	109	284	.71	.14	1.00	-.04	.99	-.05
LISTEN	168	284	-.33	.13	.99	-.10	.95	-.50
LISTEN	213	284	-1.21	.15	.94	-.75	.99	-.04
LISTEN	240	284	-1.92	.18	.97	-.20	.82	-.84
LISTEN	146	284	.06	.13	.95	-1.13	.92	-1.01
LISTEN	178	284	-.51	.14	.93	-1.30	.85	-1.60
LISTEN	39	284	2.35	.19	.93	-.55	.87	-.50
LISTEN	145	284	.08	.13	.91	-1.95	.85	-1.91
LISTEN	87	284	1.13	.14	.90	-1.60	.85	-1.24
LISTEN	223	284	-1.44	.16	.88	-1.37	.80	-1.23
LISTEN	165	284	-.27	.13	.87	-2.58	.79	-2.45
LISTEN	243	284	-2.01	.18	.87	-1.13	.62	-1.91

Table 3.23 – Main Study (October 2019) – Persons in the listening vocabulary test with highest misfit values (shaded cells represent flagged persons)

PERSON	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
P132	73	81	2.59	.39	1.07	.34	2.73	2.60
P106	23	81	-1.11	.27	1.23	1.76	2.14	4.11
P88	71	81	2.30	.36	.99	.04	1.85	1.77
P180	58	81	1.10	.27	1.33	2.40	1.78	3.18
P271	74	81	2.75	.42	.97	.01	1.72	1.29
P204	73	81	2.59	.39	1.05	.28	1.59	1.19
P163	26	81	-.91	.26	1.10	.92	1.55	2.57
P54	60	81	1.25	.28	1.06	.47	1.46	1.87
P125	48	81	.45	.25	1.16	1.61	1.43	2.74
P209	55	81	.89	.26	1.25	2.08	1.42	2.13
P40	30	81	-.65	.25	1.09	.93	1.40	2.27
P46	25	81	-.97	.26	1.16	1.31	1.40	1.87
P30	57	81	1.03	.27	1.10	.80	1.37	1.75
P45	39	81	-.10	.24	1.24	2.57	1.37	2.61
P114	40	81	-.04	.24	1.10	1.18	1.37	2.67
P185	64	81	1.58	.29	1.07	.50	1.36	1.26
P67	31	81	-.58	.25	1.10	.97	1.35	2.09
P84	38	81	-.16	.24	1.17	1.80	1.35	2.49
P250	36	81	-.28	.25	1.12	1.36	1.35	2.37
P76	45	81	.26	.25	1.18	1.87	1.34	2.42
P255	38	81	-.16	.24	1.27	2.86	1.34	2.41
P147	48	81	.45	.25	1.13	1.35	1.33	2.16
P123	50	81	.57	.25	1.09	.89	1.32	2.02
P63	57	81	1.03	.27	1.00	.05	1.30	1.47
P216	64	81	1.58	.29	1.15	.91	1.30	1.09
P36	52	81	.70	.25	1.14	1.31	1.27	1.60
P57	51	81	.63	.25	1.17	1.64	1.27	1.66
P14	70	81	2.18	.35	1.07	.39	1.26	.75
P58	44	81	.20	.25	1.23	2.41	1.26	1.90
P153	57	81	1.03	.27	1.03	.26	1.26	1.31
P248	34	81	-.40	.25	1.20	2.05	1.26	1.74
P6	52	81	.70	.25	1.20	1.80	1.25	1.52
P83	54	81	.83	.26	1.07	.66	1.25	1.42
P258	32	81	-.52	.25	1.13	1.37	1.25	1.57
P81	56	81	.96	.26	1.14	1.20	1.23	1.20
P222	44	81	.20	.25	1.11	1.22	1.23	1.70
P210	66	81	1.76	.31	1.04	.28	1.22	.76
P228	46	81	.32	.25	1.12	1.32	1.22	1.60
P113	64	81	1.58	.29	1.21	1.26	1.16	.65
BETTER FITTING PARTICIPANTS NOT SHOWN								
P157	53	81	.76	.26	.85	-1.39	.77	-1.42
P186	63	81	1.49	.29	.84	-1.02	.75	-.96
P276	63	81	1.49	.29	.84	-1.05	.71	-1.16
P279	44	81	.20	.25	.84	-1.91	.79	-1.75
P27	36	81	-.28	.25	.83	-1.92	.77	-1.79
P38	77	81	3.40	.53	.83	-.26	.45	-.69
P128	47	81	.38	.25	.83	-1.86	.82	-1.34
P75	73	81	2.59	.39	.82	-.60	.78	-.31
P140	35	81	-.34	.25	.82	-2.01	.76	-1.85
P145	36	81	-.28	.25	.78	-2.60	.81	-1.45
P198	70	81	2.18	.35	.81	-.83	.62	-.99
P25	54	81	.83	.26	.80	-1.89	.72	-1.77
P5	63	81	1.49	.29	.78	-1.46	.59	-1.77
P236	67	81	1.85	.32	.76	-1.35	.54	-1.63
P21	67	81	1.85	.32	.75	-1.38	.53	-1.65

Table 3.24 – Main Study (October 2019) – Persons in the written vocabulary test with highest misfit values (shaded cells represent flagged persons)

PERSON	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
P20	61	81	1.40	.29	1.48	2.85	2.95	4.69
P112	65	81	1.75	.31	1.11	.70	2.40	3.04
P132	72	81	2.56	.38	1.09	.41	1.84	1.44
P185	71	81	2.42	.37	1.24	1.02	1.76	1.41
P232	50	81	.60	.26	1.23	1.98	1.71	3.27
P1	56	81	1.02	.27	1.30	2.23	1.67	2.53
P248	37	81	-.22	.25	1.32	3.00	1.64	3.37
P183	53	81	.81	.26	1.35	2.71	1.51	2.25
P181	68	81	2.06	.33	1.02	.16	1.50	1.19
P12	53	81	.81	.26	1.09	.82	1.46	2.05
P225	30	81	-.68	.26	1.37	3.15	1.45	2.13
P65	48	81	.47	.25	1.31	2.76	1.44	2.28
P72	50	81	.60	.26	1.25	2.16	1.41	2.04
P209	68	81	2.06	.33	1.16	.82	1.39	.99
P68	41	81	.03	.25	1.18	1.76	1.37	2.18
P170	68	81	2.06	.33	1.07	.40	1.37	.94
P264	52	81	.74	.26	.96	-.33	1.37	1.77
P162	64	81	1.66	.30	1.10	.62	1.36	1.09
P180	61	81	1.40	.29	1.12	.80	1.34	1.18
P129	71	81	2.42	.37	.95	-.12	1.33	.76
P46	46	81	.34	.25	1.14	1.38	1.31	1.77
P85	52	81	.74	.26	1.09	.79	1.31	1.49
P120	57	81	1.09	.27	1.15	1.15	1.31	1.28
P257	66	81	1.85	.31	1.03	.21	1.31	.88
P114	42	81	.09	.25	1.21	2.04	1.30	1.79
P254	39	81	-.10	.25	1.04	.44	1.30	1.80
P15	59	81	1.24	.28	1.04	.32	1.29	1.11
P83	51	81	.67	.26	1.15	1.31	1.28	1.42
P126	39	81	-.10	.25	1.20	1.95	1.28	1.69
P255	40	81	-.04	.25	1.20	1.96	1.28	1.71
P50	69	81	2.17	.34	.93	-.26	1.27	.70
P146	32	56	.28	.31	1.26	2.01	1.27	1.29
P81	60	81	1.32	.28	1.19	1.32	1.26	.98
P14	68	81	2.06	.33	1.08	.43	1.25	.70
P163	49	81	.54	.26	1.12	1.11	1.25	1.37
P3	63	81	1.57	.30	1.07	.46	1.24	.82
P27	44	81	.22	.25	1.24	2.25	1.24	1.43
P149	51	81	.67	.26	1.21	1.83	1.24	1.26
P175	53	81	.81	.26	1.20	1.68	1.24	1.17
P261	39	81	-.10	.25	1.22	2.12	1.24	1.44
P270	41	81	.03	.25	1.14	1.41	1.23	1.41
P41	50	81	.60	.26	1.21	1.87	1.22	1.20
P54	70	81	2.29	.35	.90	-.39	1.22	.59
P242	50	81	.60	.26	1.16	1.41	1.22	1.19
P124	73	81	2.71	.40	1.07	.33	.76	-.24
BETTER FITTING PARTICIPANTS NOT SHOWN								
P91	62	81	1.49	.29	.80	-1.36	.67	-1.16
P111	59	81	1.24	.28	.80	-1.51	.67	-1.36
P227	62	81	1.49	.29	.79	-1.39	.65	-1.26
P275	72	81	2.56	.38	.79	-.77	.72	-.40
P115	58	81	1.17	.28	.78	-1.73	.67	-1.43
P62	56	81	1.02	.27	.77	-1.90	.66	-1.65
P104	78	81	3.87	.61	.77	-.33	.26	-.82
P214	78	81	3.87	.61	.77	-.32	.28	-.78
P204	72	81	2.56	.38	.76	-.94	.40	-1.29
P276	66	81	1.85	.31	.76	-1.42	.53	-1.43
P102	52	81	.74	.26	.75	-2.33	.66	-1.94
P125	64	81	1.66	.30	.74	-1.66	.69	-.94
P237	64	81	1.66	.30	.71	-1.88	.49	-1.82

Table 3.25 – Main Study (October 2019) – Persons in the listening comprehension test with highest misfit values

PERSON	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
P49	21	25	2.17	.63	1.30	.81	3.97	2.46
P7	12	25	-.17	.46	1.52	2.60	3.14	3.99
P39	9	25	-.82	.47	1.51	2.32	2.79	2.73
P106	1	25	-3.82	1.04	1.15	.46	2.78	1.33
P224	20	25	1.81	.58	1.59	1.53	2.77	2.08
P192	20	25	1.81	.58	1.19	.61	2.65	1.98
P176	1	25	-3.82	1.04	1.14	.45	2.04	1.04
P69	17	25	.94	.50	1.50	1.81	2.00	2.01
P46	15	25	.47	.47	1.45	1.99	1.99	2.32
P118	5	25	-1.82	.55	1.37	1.23	1.98	1.21
P77	13	25	.04	.46	1.62	2.95	1.97	2.33
P67	8	25	-1.04	.48	1.27	1.24	1.93	1.55
P4	8	25	-1.04	.48	1.06	.35	1.92	1.54
P98	13	25	.04	.46	1.02	.14	1.91	2.21
P45	7	25	-1.28	.50	.88	-.46	1.90	1.37
P251	19	25	1.49	.55	1.59	1.70	1.89	1.48
P34	20	25	1.81	.58	1.36	1.02	1.88	1.29
P63	18	25	1.20	.52	1.49	1.61	1.87	1.64
P231	18	25	1.20	.52	.97	-.01	1.87	1.64
P166	8	25	-1.04	.48	1.11	.59	1.86	1.47
P171	3	25	-2.53	.65	1.31	.81	1.84	.99
P153	14	25	.25	.47	1.57	2.62	1.82	2.06
P133	8	25	-1.04	.48	.97	-.08	1.80	1.39
P18	15	25	.47	.47	1.29	1.38	1.78	1.93
P177	14	25	.25	.47	1.14	.77	1.78	1.98
P220	16	25	.70	.48	1.39	1.62	1.77	1.80
P255	14	25	.25	.47	1.47	2.23	1.74	1.90
P122	8	25	-1.04	.48	1.49	2.07	1.72	1.29
P121	15	25	.47	.47	1.15	.76	1.66	1.70
P254	6	25	-1.54	.52	1.37	1.37	1.64	.99
P3	20	25	1.81	.58	1.62	1.60	1.17	.47
P114	10	25	-.60	.46	1.26	1.36	1.61	1.35
P40	8	25	-1.04	.48	1.54	2.27	1.32	.72
P196	21	25	2.17	.63	1.51	1.22	1.37	.68
P73	19	25	1.49	.55	1.14	.53	1.50	.98
P28	16	25	.70	.48	1.19	.85	1.49	1.26
P204	23	25	3.20	.83	1.48	.93	1.03	.48
P263	14	25	.25	.47	1.28	1.39	1.48	1.35
P51	16	25	.70	.48	1.10	.52	1.46	1.21
P102	10	25	-.60	.46	1.43	2.09	1.38	.94
P170	20	25	1.81	.58	1.43	1.19	1.36	.71
P152	17	25	.94	.50	.98	-.01	1.42	1.04
P208	10	25	-.60	.46	1.40	1.97	1.36	.91
P145	6	25	-1.54	.52	1.27	1.04	1.35	.68
P149	11	25	-.38	.46	1.34	1.79	1.33	.89
P162	15	25	.47	.47	1.24	1.14	1.34	1.00
BETTER FITTING PARTICIPANTS NOT SHOWN								
P27	12	25	-.17	.46	.66	-2.08	.56	-1.35
P150	4	25	-2.14	.59	.66	-1.03	.39	-.50
P256	15	25	.47	.47	.66	-1.85	.53	-1.51
P17	12	25	-.17	.46	.65	-2.18	.54	-1.42
P174	14	25	.25	.47	.65	-2.05	.53	-1.56
P26	20	25	1.81	.58	.63	-1.10	.46	-.83
P33	20	25	1.81	.58	.63	-1.10	.46	-.83
P92	5	25	-1.82	.55	.62	-1.40	.40	-.68
P104	21	25	2.17	.63	.61	-.99	.40	-.73
P131	20	25	1.81	.58	.61	-1.15	.40	-1.00
P280	21	25	2.17	.63	.61	-.97	.41	-.70
P31	16	25	.70	.48	.60	-1.98	.47	-1.64
P38	24	25	4.09	1.10	.57	-.34	.10	-.65
P32	19	25	1.49	.55	.53	-1.71	.36	-1.41

APPENDIX 15 – Main Study – Second Data Gathering – Tables with fit statistics – JUNE 2020

Table 3.33 – Main Study (June 2020) – Items in the listening vocabulary test with highest misfit values (shaded cells represent flagged items)

ITEM	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
L8	14	17	-.12	.69	1.55	1.31	6.96	3.47
L11	14	17	-.12	.69	1.41	1.05	2.71	1.66
L5	15	17	-.67	.80	1.39	.83	2.28	1.20
L74	10	17	1.33	.55	1.46	1.86	1.69	1.62
L100	15	17	-.67	.80	1.24	.60	1.69	.88
L131	12	17	.68	.59	1.48	1.61	1.53	1.03
L111	15	17	-.67	.80	.92	.03	1.51	.77
L136	13	17	.31	.63	1.31	.98	1.36	.70
L52	7	17	2.23	.56	1.34	1.42	1.32	.93
L16	12	17	.68	.59	1.16	.65	1.26	.62
L22	14	17	-.12	.69	1.25	.72	.90	.17
L108	15	17	-.67	.80	1.21	.55	.89	.29
L120	15	17	-.67	.80	1.21	.55	.89	.29
L88	15	17	-.67	.80	1.20	.52	.84	.24
L17	11	17	1.01	.57	1.18	.77	1.19	.54
L55	16	17	-1.51	1.07	1.18	.48	1.16	.58
L103	16	17	-1.51	1.07	1.18	.48	1.16	.58
L144	15	17	-.67	.80	1.15	.45	.82	.22
L127	14	17	-.12	.69	.96	.04	1.13	.43
L82	15	17	-.67	.80	1.11	.38	.78	.18
L104	15	17	-.67	.80	1.11	.38	.78	.18
L94	16	17	-1.51	1.07	1.10	.40	.71	.24
L124	16	17	-1.51	1.07	1.10	.40	.71	.24
L51	1	17	5.15	1.09	1.09	.38	.49	.02
L28	12	17	.68	.59	1.08	.39	1.05	.26
L46	13	17	.31	.63	1.07	.31	.89	.06
L143	13	17	.31	.63	1.06	.28	.79	-.11
L31	16	17	-1.51	1.07	1.05	.34	.55	.09
L75	16	17	-1.51	1.07	1.05	.34	.55	.09
L59	14	17	-.12	.69	1.02	.17	.66	-.16
L1	16	17	-1.51	1.07	1.01	.29	.47	.00
L73	16	17	-1.51	1.07	1.01	.29	.47	.00
L109	16	17	-1.51	1.07	1.01	.29	.47	.00
L58	15	17	-.67	.80	.89	-.04	.97	.36
L70	3	17	3.71	.70	.97	.07	.97	.23
L116	12	17	.68	.59	.94	-.15	.94	.07
L63	14	17	-.12	.69	.92	-.06	.82	.06
L130	14	17	-.12	.69	.92	-.06	.82	.06
BETTER FITTING NOT SHOWN								
L145	7	17	2.23	.56	.89	-.45	.77	-.58
L83	13	17	.31	.63	.89	-.24	.75	-.19
L138	14	17	-.12	.69	.84	-.28	.59	-.28
L67	15	17	-.67	.80	.83	-.16	.64	-.03
L56	15	17	-.67	.80	.82	-.19	.59	-.04
L62	14	17	-.12	.69	.81	-.39	.58	-.30
L14	14	17	-.12	.69	.81	-.39	.58	-.30
L57	13	17	.31	.63	.79	-.60	.61	-.46
L92	12	17	.68	.59	.77	-.82	.60	-.68
L140	15	17	-.67	.80	.75	-.34	.44	-.23
L90	15	17	-.67	.80	.75	-.34	.44	-.23
L148	6	17	2.55	.57	.73	-1.14	.60	-.99
L23	14	17	-.12	.69	.72	-.65	.45	-.52
L49	7	17	2.23	.56	.70	-1.41	.60	-1.15
L89	14	17	-.12	.69	.68	-.77	.43	-.57
L42	14	17	-.12	.69	.68	-.77	.43	-.57
L106	16	17	-1.51	1.07	.66	-.18	.20	-.41
L134	16	17	-1.51	1.07	.66	-.18	.20	-.41
L10	16	17	-1.51	1.07	.66	-.18	.20	-.41
L48	15	17	-.67	.80	.65	-.57	.33	-.41
EXTREME SCORES/MINIMUM MEASURES NOT SHOWN								

Table 3.34 – Main Study (June 2020) – Items in the written vocabulary test with highest misfit values

ITEM	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
W55	16	17	-.97	-.97	1.05	1.16	.46	1.99
W52	8	17	2.26	2.26	.53	1.50	2.48	1.62
W35	16	17	-.97	-.97	1.05	1.14	.44	1.56
W82	16	17	-.97	-.97	1.05	1.14	.44	1.56
W127	15	17	-.17	-.17	.78	1.23	.58	1.34
W146	11	17	1.42	1.42	.55	1.21	.97	1.29
W140	16	17	-.97	-.97	1.05	1.12	.42	1.27
W51	5	17	3.16	3.16	.58	1.21	.82	1.24
W16	13	17	.76	.76	.61	1.22	.76	1.22
W36	15	17	-.17	-.17	.78	1.13	.41	1.22
W74	9	17	1.99	1.99	.53	1.22	1.21	1.18
W62	12	17	1.11	1.11	.57	1.17	.72	1.09
W132	14	17	.35	.35	.67	1.01	.15	1.17
W111	15	17	-.17	-.17	.78	1.16	.46	1.11
W142	14	17	.35	.35	.67	1.15	.48	1.09
W131	14	17	.35	.35	.67	.93	-.05	1.06
W42	16	17	-.97	-.97	1.05	1.05	.35	.76
W46	16	17	-.97	-.97	1.05	1.05	.35	.76
W48	16	17	-.97	-.97	1.05	1.05	.35	.76
W106	16	17	-.97	-.97	1.05	1.05	.35	.76
W134	16	17	-.97	-.97	1.05	1.05	.35	.76
W50	16	17	-.97	-.97	1.05	1.03	.32	.66
W58	16	17	-.97	-.97	1.05	1.03	.32	.66
W67	16	17	-.97	-.97	1.05	1.03	.32	.66
W145	12	17	1.11	1.11	.57	1.02	.15	.96
BETTER FITTING NOT SHOWN								
W28	14	17	.35	.35	.67	.91	-.11	.89
W90	11	17	1.42	1.42	.55	.90	-.41	.82
W120	15	17	-.17	-.17	.78	.89	-.03	.87
W136	14	17	.35	.35	.67	.88	-.18	.86
W144	14	17	.35	.35	.67	.88	-.18	.86
W143	14	17	.35	.35	.67	.86	-.24	.69
W10	15	17	-.17	-.17	.78	.80	-.23	.51
W63	15	17	-.17	-.17	.78	.80	-.23	.51
W70	9	17	1.99	1.99	.53	.78	-1.24	.73
W14	16	17	-.97	-.97	1.05	.75	-.03	.27
W23	13	17	.76	.76	.61	.75	-.78	.58
W83	16	17	-.97	-.97	1.05	.75	-.03	.27
W124	16	17	-.97	-.97	1.05	.75	-.03	.27
W59	14	17	.35	.35	.67	.73	-.62	.50
W49	14	17	.35	.35	.67	.68	-.79	.45
EXTREME SCORES / MINIMUM MEASURES NOT SHOWN								

Table 3.35 – Main Study (June 2020) – Items in the listening comprehension test with highest misfit values (shaded cells represent flagged items)

ITEM	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
LISTEN	0	17	5.54	1.84	MAXIMUM MEASURE			
LISTEN	10	17	.49	.54	1.22	1.10	2.24	2.78
LISTEN	12	17	-.12	.57	1.56	1.99	1.79	1.46
LISTEN	14	17	-.88	.67	1.23	.68	1.52	.84
LISTEN	12	17	-.12	.57	1.40	1.50	1.46	.97
LISTEN	12	17	-.12	.57	1.33	1.29	1.38	.84
LISTEN	16	17	-2.21	1.06	1.12	.42	1.20	.60
LISTEN	10	17	.49	.54	1.15	.80	1.18	.60
LISTEN	9	17	.78	.53	1.11	.60	1.07	.31
LISTEN	15	17	-1.40	.78	1.09	.34	.98	.33
LISTEN	15	17	-1.40	.78	.95	.08	1.06	.40
LISTEN	5	17	1.99	.59	1.05	.28	.94	.00
LISTEN	12	17	-.12	.57	.95	-.13	.83	-.19
LISTEN	11	17	.20	.55	.91	-.38	.91	-.10
LISTEN	11	17	.20	.55	.90	-.41	.81	-.37
LISTEN	14	17	-.88	.67	.88	-2.20	.59	-.37
LISTEN	11	17	.20	.55	.87	-.53	.77	-.47
LISTEN	13	17	-.47	.61	.83	-.48	.69	-.36
LISTEN	10	17	.49	.54	.82	-.90	.73	-.73
LISTEN	14	17	-.88	.67	.78	-.48	.53	-.47
LISTEN	6	17	1.66	.56	.76	-1.00	.64	-1.07
LISTEN	11	17	.20	.55	.75	-1.21	.66	-.80
LISTEN	13	17	-.47	.61	.67	-1.15	.49	-.82
LISTEN	4	17	2.36	.63	.58	-1.31	.42	-1.25
LISTEN	10	17	.49	.54	1.22	1.10	2.24	2.78
LISTEN	17	17	-3.46	1.83	MINIMUM MEASURE			

Table 3.36 – Main Study (June 2020) – Persons with their fit values in the LVT (shaded cells represent flagged persons)

PERSON	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
PERSON1	66	81	1.41	.34	1.40	1.84	1.73	2.02
PERSON14	79	81	4.52	.83	.87	.01	1.68	.87
PERSON2	76	81	3.18	.55	1.32	.85	.89	.13
PERSON6	73	81	2.46	.45	1.21	.73	1.04	.26
PERSON16	61	81	.88	.31	1.08	.57	1.13	.58
PERSON8	56	81	.42	.29	1.07	.61	1.12	.54
PERSON4	66	81	1.41	.34	1.06	.38	1.00	.10
PERSON5	44	81	-.55	.28	1.02	.21	.98	.07
PERSON13	65	81	1.29	.34	.97	-.08	.84	-.48
PERSON17	70	81	1.94	.39	.97	-.05	.85	-.25
PERSON15	73	81	2.46	.45	.95	-.07	.86	-.08
PERSON7	58	81	.60	.30	.94	-.43	.93	-.19
PERSON9	74	81	2.67	.47	.89	-.22	.86	-.03
PERSON3	74	81	2.67	.47	.86	-.34	.46	-.94
PERSON10	68	81	1.66	.36	.83	-.71	.59	-1.23
PERSON12	61	81	.88	.31	.82	-1.14	.72	-1.11
PERSON11	74	81	2.67	.47	.71	-.85	.33	-1.32

Table 3.37 – Main Study (June 2020) – Persons with their fit values in the WVT

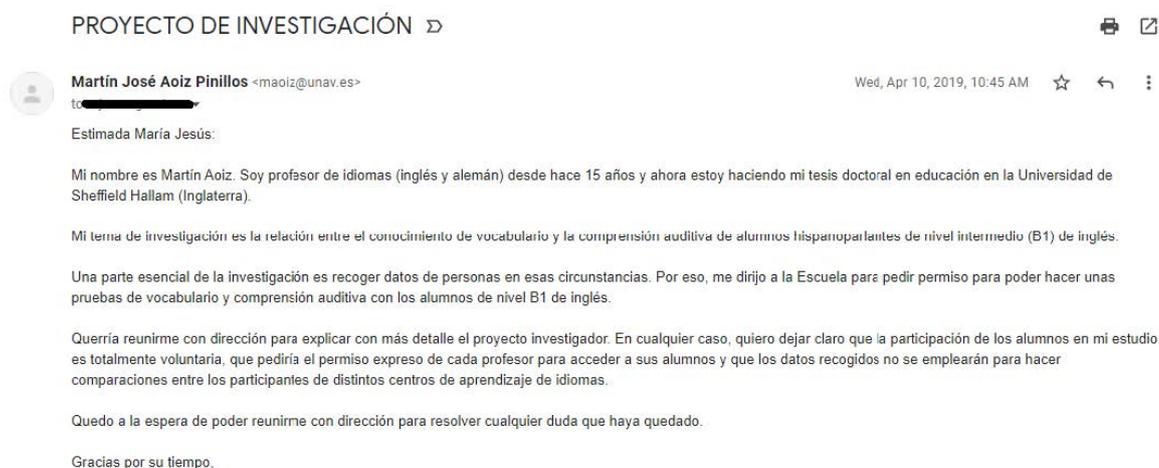
PERSON	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
PERSON4	75	81	2.26	.48	1.15	.56	1.65	1.21
PERSON5	71	81	1.50	.40	1.18	.82	1.53	1.46
PERSON1	76	81	2.51	.52	1.28	.85	1.52	.94
PERSON13	70	81	1.35	.39	1.14	.69	1.24	.82
PERSON16	62	81	.33	.34	1.14	.97	1.15	.84
PERSON6	77	81	2.81	.57	1.14	.47	.78	-1.10
PERSON12	74	81	2.04	.46	1.06	.29	1.13	.41
PERSON7	69	81	1.20	.38	1.08	.44	1.12	.52
PERSON9	76	81	2.51	.52	1.07	.31	.79	-1.17
PERSON17	71	81	1.50	.40	1.02	.18	.95	-.03
PERSON2	78	81	3.17	.64	.91	-.03	.32	-.80
PERSON15	75	81	2.26	.48	.87	-.31	.78	-.28
PERSON10	73	81	1.85	.43	.83	-.58	.58	-1.05
PERSON8	68	81	1.06	.37	.79	-1.16	.79	-.79
PERSON3	77	81	2.81	.57	.76	-.50	.39	-.91
PERSON14	80	81	4.43	1.04	.70	-.07	.10	-.67
PERSON11	77	81	2.81	.57	.57	-1.13	.22	-1.45

Table 3.38 – Main Study (June 2020) – Persons with their fit values in the LCT

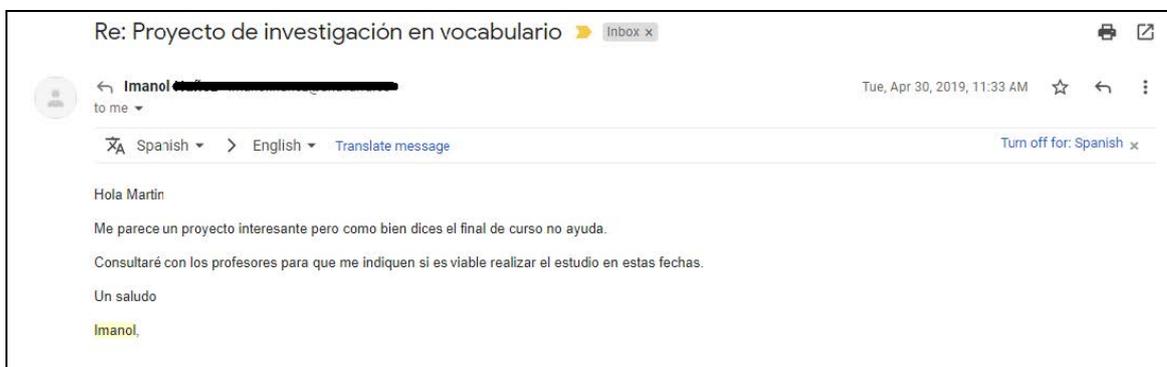
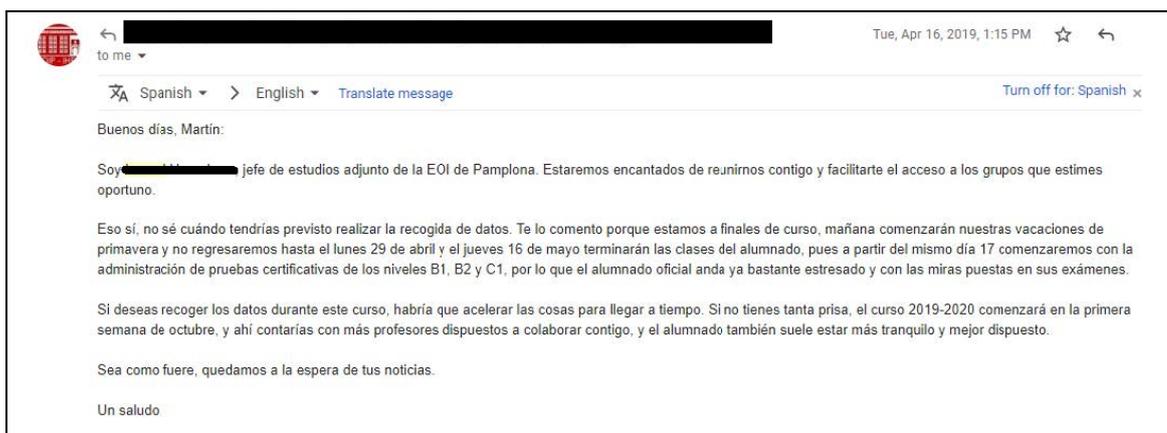
PERSON	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
PERSON13	16	25	.75	.48	1.14	.72	1.66	1.84
PERSON14	20	25	1.87	.60	1.25	.76	1.40	.78
PERSON9	17	25	.99	.50	1.14	.63	1.30	.86
PERSON16	18	25	1.25	.52	1.01	.13	1.26	.70
PERSON3	12	25	-.12	.46	1.21	1.31	1.14	.60
PERSON4	18	25	1.25	.52	1.16	.65	1.00	.14
PERSON15	20	25	1.87	.60	1.16	.54	.97	.16
PERSON2	23	25	3.61	1.06	1.15	.45	1.03	.49
PERSON11	21	25	2.27	.67	1.10	.36	.93	.18
PERSON17	15	25	.52	.47	1.05	.32	1.01	.14
PERSON6	14	25	.30	.46	.94	-.29	.89	-.36
PERSON1	9	25	-.77	.48	.92	-.36	.80	-.52
PERSON12	11	25	-.33	.46	.89	-.62	.81	-.65
PERSON8	14	25	.30	.46	.83	-.97	.73	-1.04
PERSON5	13	25	.09	.46	.79	-1.36	.71	-1.19
PERSON7	18	25	1.25	.52	.79	-.75	.67	-.71
PERSON10	18	25	1.25	.52	.72	-1.04	.60	-.94

APPENDIX 16 – Access to students at language school in Preliminary Study – Emails Exchange (April 2019)

1) First contact with language school to explain the project and ask for permission (10th April 2019)



2) Answers from dean of studies granting permission and including mails for teachers involved (16th April 2019)



  Tue, May 7, 2019, 12:15 PM ☆ ↶

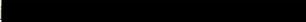
to me ▾

🌐 Spanish ▾ > English ▾ [Translate message](#) [Turn off for: Spanish](#) x

Hola, Martín:

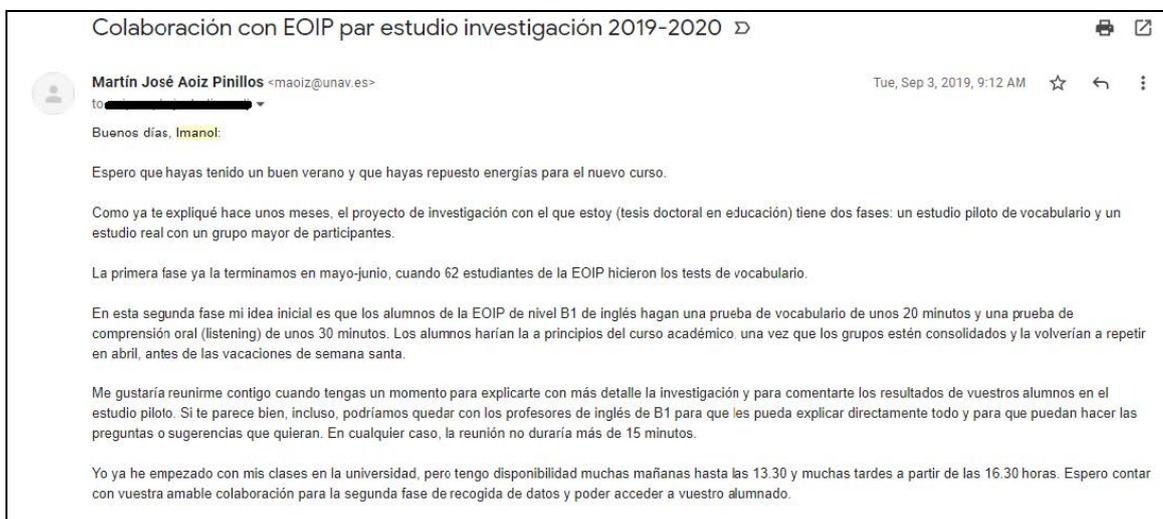
Cristina, Genoveva, Ramón, José Miguel y Noelia están avisados de tu visita a sus grupos, por lo que no debes tener problema.

Un saludo,

Hau idatzí du eoipamplo-jestudios-adj eoipamplo-jestudios-adj  erabiltzaileak (2019 mai. 7, ar. (12:03)):
De acuerdo.

APPENDIX 17 – Access to students at language school – Main Study – First Data Collection – Emails Exchange (September 2019)

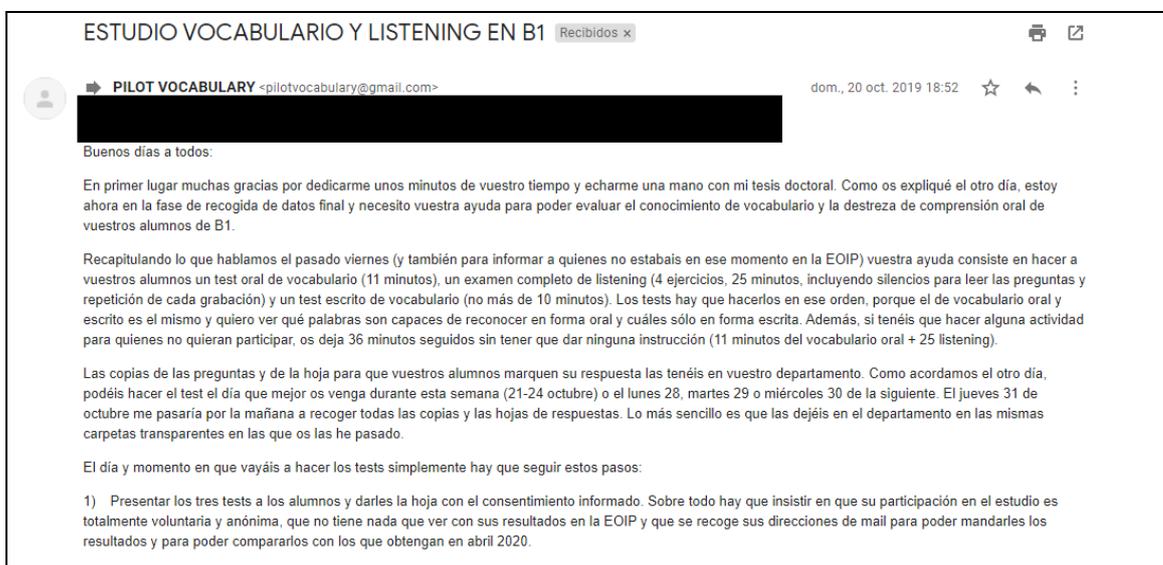
1) First contact with language school to ask for permission for data collection in the main study (3rd September 2019)



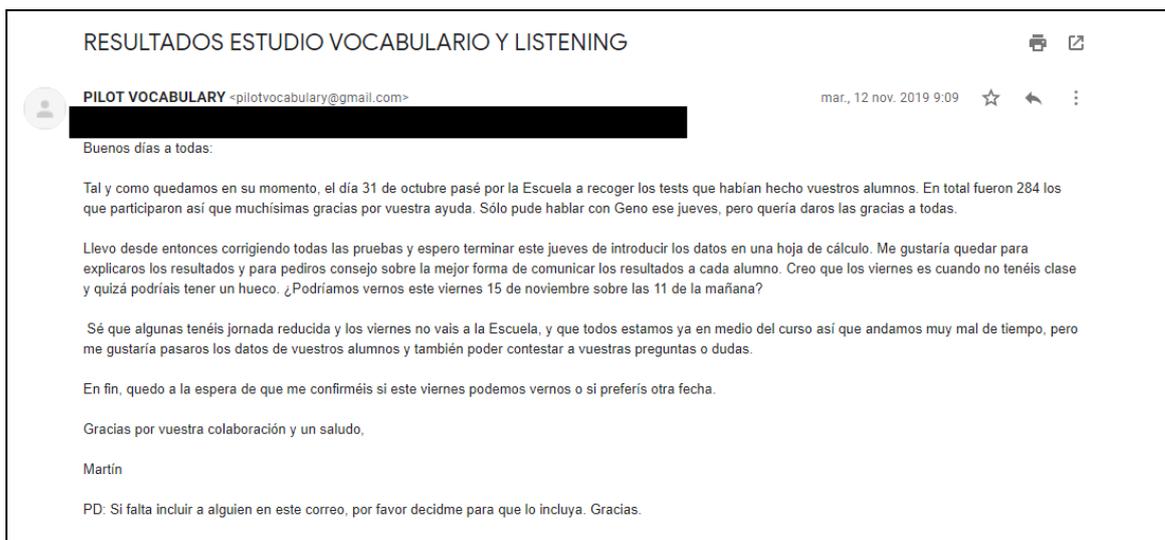
2) Answer from dean of studies granting permission and including mails for teachers involved (17th October 2019)



3) Mail to teachers with details about how to deliver the tests in their classes. Mail included the files. (20th October 2019)



4) Mail to thank the teachers for their help, and to arrange a meeting to discuss their students' overall results. (12th November 2019)

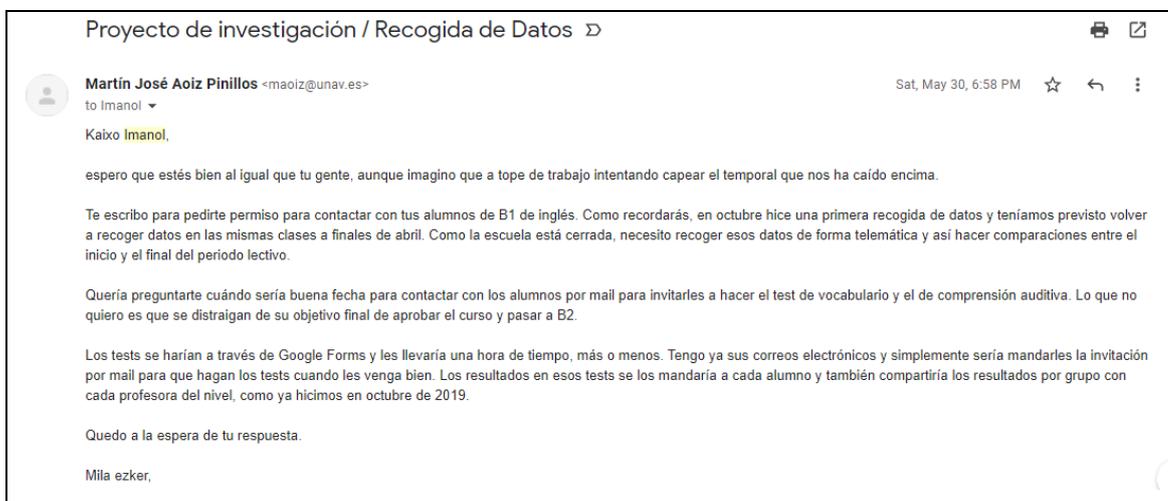


5) Mail from the teacher coordinator for the B1-level to confirm the meeting to discuss results. (14th November 2019)

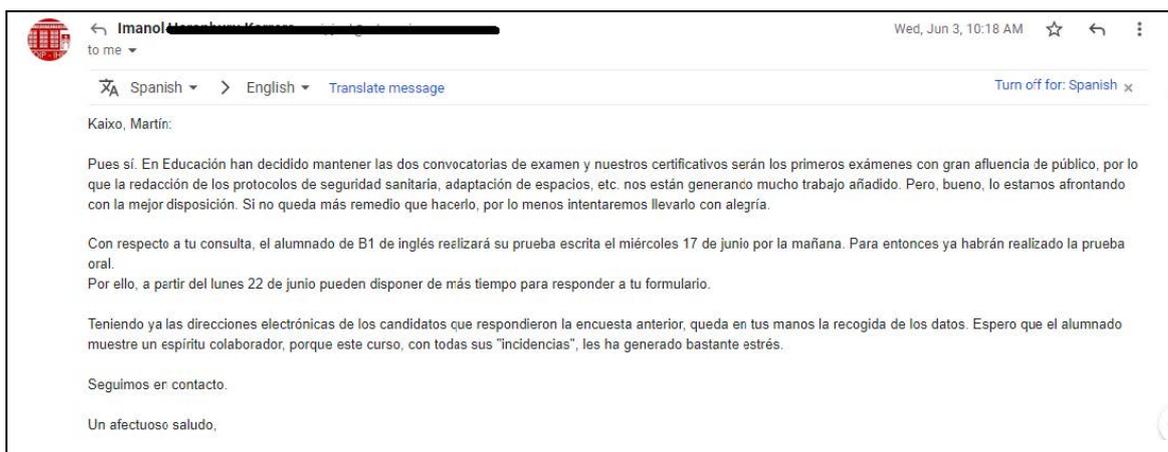


APPENDIX 18 – Access to students at language school – Emails Exchange (May 2020)

1) Mail to ask for permission to contact students for second data collection through Google forms (30th May 2020)



2) Answer from dean of studies granting permission for the second data collection (3rd June 2020)



APPENDIX 19 – Emails exchange with participants in Preliminary Study (May 2019)

1) Invitation to access the Google Form® with the test (6th May 2019)

 **Martin Aoiz Pinillos** <pilotvocabulary@gmail.com>
para Martín, maoiz ▾

lun., 6 may. 2019 8:38 ☆ ↶ ⋮

Dear English student,
This is the formulary with the vocabulary tests you are interested in doing. Follow the instructions on the screen to select your answers to the questions. At the end of the test, click on the ENVIAR button and I will receive your answers. In a few weeks, I will contact you again to send your results.
Thanks for participating in this vocabulary study,
Martín

VOCABULARY SIZE TEST

INFORMATION ABOUT THE RESEACH STUDY (text in Spanish below)

The University undertakes research as part of its function for the community under its legal status. Data protection allows us to use personal data for research with appropriate safeguards in place under the legal basis of public tasks that are in the public interest. A full statement of your rights can be found at <https://www.shu.ac.uk/about-this-website/privacy-policy/privacy-notices/privacy-notice-for-research>. However, all University research is reviewed to ensure that participants are treated appropriately and their rights respected. This study was approved by UREC with Converis number ER9510976. Further information at: <https://www.shu.ac.uk/research/ethics-integrity-and-practice>

1. The purpose of this pilot study is to test the validity of the items included in two vocabulary tests. You have been invited to participate because the main research study focuses on students like you, who are attending English classes in Spain and who have Spanish as their first language.
2. Your participation in this study is beneficial because it will help me in my research study and therefore, it might contribute to advance research in language

2) Reminder with the link to the Google Form® (10th May 2019)

VOCABULARY SIZE TEST 

 **PILOT VOCABULARY** <pilotvocabulary@gmail.com>
para maoiz ▾

vie., 10 may. 2019 13:19 ☆ ↶

Dear student,
here you have a link to the vocabulary tests you have agreed to do.

https://docs.google.com/forms/d/e/1FAIpQLSckRexkpN0qUgJDKD1HyRfNzqbsCELaOSFCNzkBTPCWtWitfg/viewform?usp=sf_link

Click on the link to access the study.
Read the information about this study. Then, enter your email address and click the button to access the formulary.
You have time to do the tests until 2nd June.
Thanks for your help,
Martín Aoiz

3) Email with one participant's tests results (13th May 2019)

VOCABULARY TEST - RESULTS

 **PILOT VOCABULARY** <pilotvocabulary@gmail.com>
para [REDACTED] ▾

lun., 13 may. 2019 8:58 ☆

Dear student,

here are your results in the two vocabulary tests you have done:

LISTENING VOCABULARY TEST: 128/150 (85% correct answers)
WRITTEN VOCABULARY TEST: 144/150 (96% correct answers)

APPENDIX 20 – Emails exchange with participants in Main Study – First Data Collection (October 2019)

1) Email with one participant's tests results in the first dataset (16th November 2019)

RESULTS



PILOT VOCABULARY <pilotvocabulary@gmail.com>

sáb., 16 nov. 2019 8:31



para [REDACTED]

Dear student,

Here are your results in the tests you did in October 2019:

LISTENING VOCABULARY TEST: 66.67%
LISTENING COMPREHENSION TEST: 44.00%
WRITTEN VOCABULARY TEST: 69.14%

The average results for the 284 participants at EOI were:

LISTENING VOCABULARY TEST: 61.04%
LISTENING COMPREHENSION TEST: 53.34%
WRITTEN VOCABULARY TEST: 72.21%

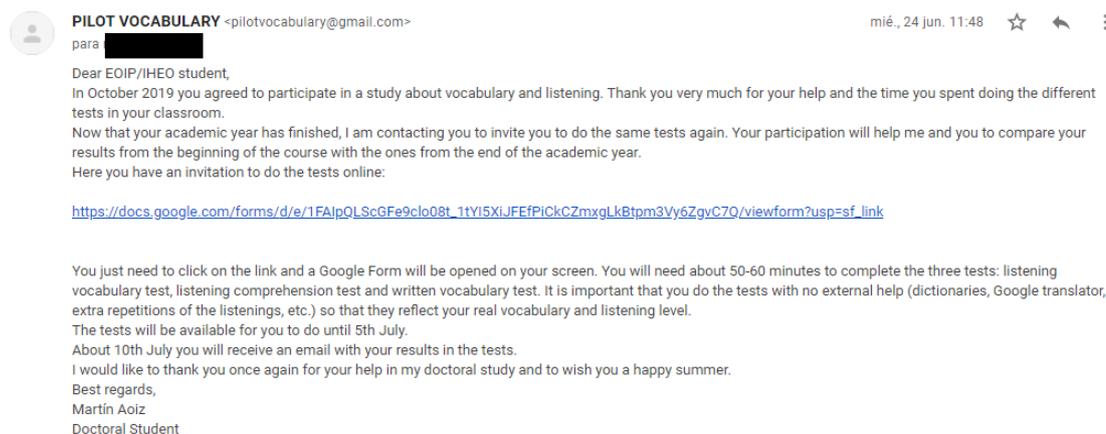
If you have any questions, please contact me.

Thank you very much for your help in my doctoral investigation.

Martín Aoiz
PhD Candidate
Sheffield Hallam University

APPENDIX 21 – Emails exchange with participants in Main Study – Second Data Collection (June 2020)

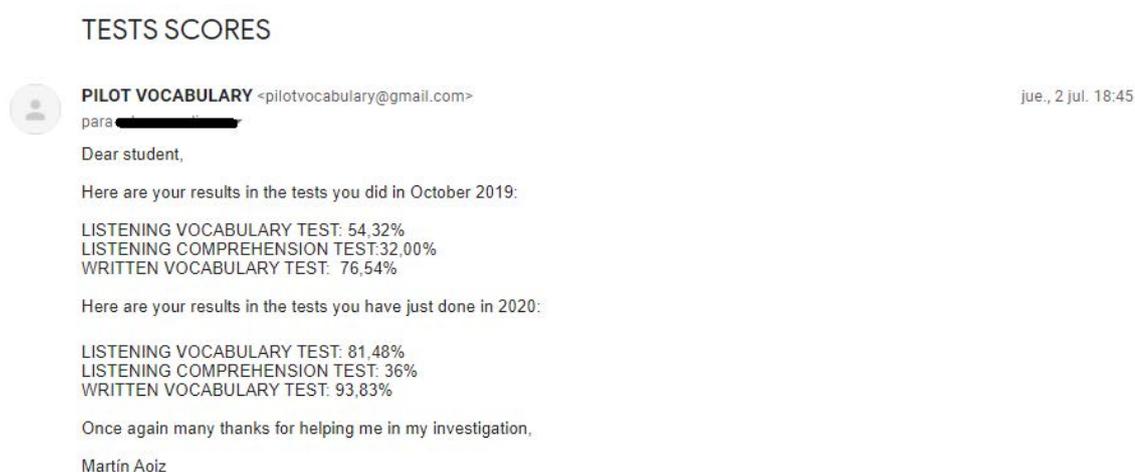
1) Invitation to access the Google Form® with the test (24th June 2020)



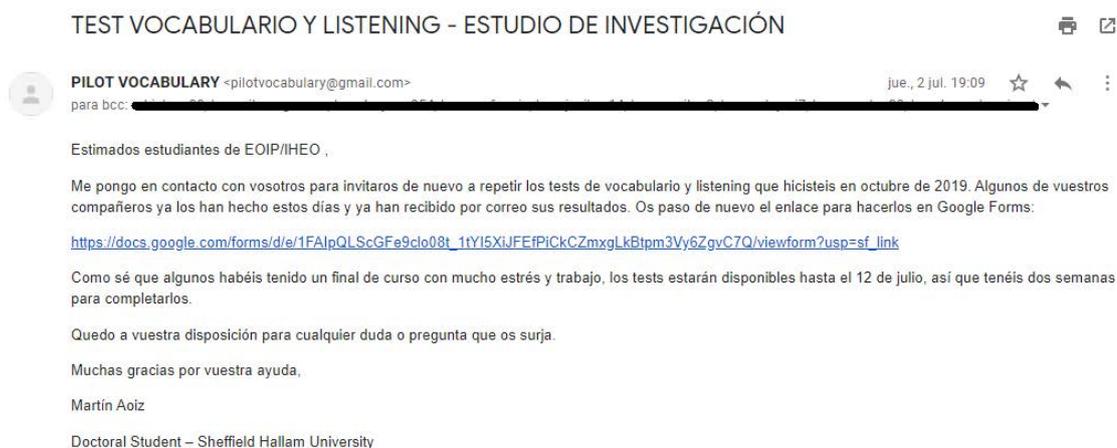
2) Email informing teachers about the second data collection and including the invitation sent to students (24th June 2020).



3) Email to a participant with their test results (2nd July 2020).



4) Reminder to students with the invitation to participate in the study (2nd July 2020).



5) Last reminder to students with the invitation to participate in the study (14th July 2020).



APPENDIX 22 – Screenshot of the first section of the vocabulary tests – (online version - May 2019)

VOCABULARY SIZE TEST

INFORMATION ABOUT THE RESEARCH STUDY (text in Spanish below)

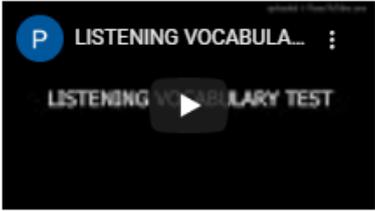
The University undertakes research as part of its function for the community under its legal status. Data protection allows us to use personal data for research with appropriate safeguards in place under the legal basis of public tasks that are in the public interest. A full statement of your rights can be found at <https://www.shu.ac.uk/about-this-website/privacy-policy/privacy-notices/privacy-notice-for-research>. However, all University research is reviewed to ensure that participants are treated appropriately and their rights respected. This study was approved by UREC with Converis number ER9510976.

Further information at: <https://www.shu.ac.uk/research/ethics-integrity-and-practice>

1. The purpose of this pilot study is to test the validity of the items included in two vocabulary tests. You have been invited to participate because the main research study focuses on students like you, who are attending English classes in Spain and who have Spanish as their first language.
2. Your participation in this study is beneficial because it will help me in my research study and therefore, it might contribute to advance research in language learning. Besides, you will have the opportunity of testing your own vocabulary size. Knowing this might help you in your study of the English language because it might provide you with useful information for the future.
3. You will have to do two vocabulary tests. It will take you about 35 minutes to finish the two tests.
4. Participation in this research is absolutely voluntary and will take place after you have signed an informed consent form. You will be able to withdraw from the study at any

APPENDIX 23 – Screenshot of the LVT – (online version - May 2019)

CLICK ON THE SCREEN TO LISTEN TO THE RECORDING



1 *

- a) cubo
- b) entrada
- c) basura
- d) rama

2 *

- a) cuerpo
- b) estantería
- c) operación
- d) pizarra