The Impact of Input Task Characteristics on Performance on an Integrated Listening-Into-Writing EAP Assessment

Dr. Carolyn Westbrook

This is a digitised version of a dissertation submitted to the University of Bedfordshire.

It is available to view only.

This item is subject to copyright.
THE IMPACT OF INPUT TASK CHARACTERISTICS ON PERFORMANCE ON AN INTEGRATED LISTENING-INTO-WRITING EAP ASSESSMENT

CAROLYN WESTBROOK

A thesis submitted to the University of Bedfordshire, in fulfilment of the requirements for the degree of PhD

UNIVERSITY OF BEDFORDSHIRE

CRELLA

OCTOBER 2019
Author’s declaration
I, Carolyn Westbrook, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University;
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- Where I have drawn on or cited the published work of others, this is always clearly attributed;
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- Where the thesis or any part of it is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- Parts of this work have been presented at conferences as indicated on page vi.
Abstract

Over the last fifty years, as the number of students studying in English-speaking countries or studying through the medium of English has grown exponentially, so has the market for testing the language skills of these students to ensure that they have the language and skills necessary for tertiary study in English.

While traditional tests of academic English have been skills-based, more and more integrated tests are being designed to measure English for Academic Purposes (EAP) both for university entrance purposes and for in-sessional English courses due to the increased authenticity (Bachman and Palmer, 1996, p. 23) and interactiveness (p. 25) that such tests can offer.

Furthermore, as more and more teaching moves to a ‘flipped’ or blended model and computer-based testing increases, there is a need to ascertain how to best provide input for both testing and teaching. Traditionally, listening comprehension has been audio-only and tests have consequently used audio only input. However, the broader range of possibilities offered by technological developments means that offering video-based input as an alternative to audio only is now perfectly feasible.

This raises the question of how to ‘test for best’ (Weir, 2005, p. 54). A number of studies have investigated audio versus video or multimedia listening comprehension tests. Similarly, much research has been done into reading-into-writing or listening and reading-into-writing, yet very little has been done on integrated listening-into-writing. This study aimed to address that gap in the research by investigating the impact of audio versus video input on performance on an integrated EAP listening-into-writing test.

In the study, participants were exposed to a lecture which was divided in half and presented in both audio and video formats in a counterbalanced measures design.

The quantitative findings of this study revealed that there was a significant difference in scores between the audio first group, which was exposed to the audio input in the first half of the
lecture, and the video first group, which was exposed to the audio input in the second half of
the lecture, while there was only a small, non-significant difference between the two groups
when exposed to the video input. A follow-up textual analysis broadly supported these
findings.

In line with findings from Cumming et al. (2005a), the quantitative analyses suggest that
higher level learners tended to paraphrase more of the input while the lower-intermediate and
intermediate learners generated both paraphrased and verbatim reproductions of the input.
The very low levels learners appeared unable to make very much use of the input yet students
from both groups reproduced large numbers of word-level matches from the PowerPoint slides
when they had access to the video input.

While there was no clear preference for one or other of the input formats, around 40 per cent
of students expressed a preference for video while around 20 per cent said that they preferred
audio only as the video was distracting. This supports the findings of Chen, Wang and Xu.
(2014, p. 57).

The research has highlighted several areas for future research but also has important
implications for the construct of academic listening-into-writing.
Acknowledgements

Firstly, thank you to Solent University for part-funding this PhD and to the British Council ARAGs for funding my data collection and analysis.

Thank you to the staff and students of the universities of Bath and Reading who kindly took part in the exploratory study. Thank you to Dr. Viktoriya Levchenko, colleagues and students at Samara National Research University, Dr. Olga Kvasova and students at Taras Shevchenko National University of Kiev for your help and participation!

My heartfelt thanks go to my rater, Alex Brown and my coder, Leander Johnson.

I would also like to express my gratitude to my first supervisor, Professor Tony Green, and to my second supervisor, Dr John Field, and the CRELLA team, particularly Dr Fumiyo Nakatsuhara, who have always been incredibly supportive. Thank you for everything!

A huge thank you to Professor Barry O’Sullivan, OBE, Dr Karen Dunn and the British Council ARG team for believing in me and supporting me. Special thanks also to Dr Jamie Dunlea, whose unending motivation got me here! I could not wish for a better group of colleagues!

I also owe my mother, the late Vera Ogden so much. Thank you for your financial support and undying love. I am so sorry you did not manage to see me get to this point. Thank you to my mother-in-law, Kath, who has always been there for me, too.

Finally, to my family: Kai and Christian, thank you for everything. To my husband, Kevin: thank you for being one of my raters, my proofreader and my technical specialist but, most of all, thank you for your unending support. I could not have done it without you either! I love you all!
List of publications


# Table of contents

Author’s declaration ........................................................................................................... ii
Abstract .............................................................................................................................. iii
Acknowledgements ................................................................................................................ v
List of publications ................................................................................................................ vi
Table of contents .................................................................................................................... vii
List of Tables ........................................................................................................................... xii
List of Figures ........................................................................................................................... xiv
1. Introduction ......................................................................................................................... 1
   1.1 Background: The development of international student mobility and English-language instruction in higher education ................................................................................. 1
   1.2 Language testing in EAP ................................................................................................. 3
   1.3 My personal and professional interest in teaching and testing EAP.............................. 5
   1.4 Aims of the study ........................................................................................................... 6
   1.5 Structure of the thesis .................................................................................................... 7
2. Literature Review ................................................................................................................... 9
   2.1 Introduction ....................................................................................................................... 9
   2.2 Listening ............................................................................................................................. 9
      2.2.1 What is listening? ........................................................................................................ 9
      2.2.2 Idea units rather than sentences ............................................................................. 12
   2.3 Factors affecting listening ............................................................................................. 13
      2.3.1 Listener characteristics ........................................................................................... 13
      2.3.2 Text characteristics ................................................................................................ 16
      2.3.3 Testing conditions ................................................................................................... 20
      2.3.4 Input format ............................................................................................................. 22
   2.4 Academic listening ......................................................................................................... 31
      2.4.1 Note-taking .............................................................................................................. 31
   2.5 Writing ............................................................................................................................. 35
      2.5.1 Models of writing ...................................................................................................... 35
      2.5.2 Conclusion ................................................................................................................. 48
2.6 Independent tasks and real-world academic writing tasks ............................................... 49
   2.6.1 Background ................................................................................................................ 49
   2.6.2 ‘Real-world’ academic tasks.................................................................................... 50
   2.6.3 Independent writing tasks and ‘real-world’ performance ....................................... 53
   2.6.4 Independent writing tasks and authenticity ............................................................. 57
   2.6.5 Independent writing tasks and background knowledge ........................................... 58
2.7 Integrated writing tasks .................................................................................................... 61
2.7.1 Defining integrated test tasks ................................................................. 61
2.7.2 Integrated writing task types ............................................................... 63
2.7.3 Criticisms of integrated tasks ............................................................. 69
2.8 Independent vs integrated writing tasks ..................................................... 72
2.8.1 Independent vs integrated writing tasks ................................................ 72
2.8.2 Factors affecting performance on integrated writing tasks .................... 75
2.8.3 Source text use in integrated writing tasks ........................................... 79
2.8.4 Listening-into-writing research ............................................................. 84
2.8.5 Rating integrated tasks ........................................................................ 91
2.8.6 Integrated testing and plagiarism .......................................................... 92
2.9 Text difficulty ......................................................................................... 92
2.9.1 Factors contributing to text difficulty .................................................. 92
2.10 Conclusion ......................................................................................... 107
3. Exploratory study .................................................................................. 113
3.1 Introduction .......................................................................................... 113
3.2 Research approach ............................................................................. 113
3.3 Exploratory study ................................................................................ 114
  3.3.1 Participants ....................................................................................... 116
  3.3.2 Procedure ......................................................................................... 117
3.4 Exploratory study results and discussion .................................................. 127
  3.4.1 RQ1: Is there a relationship between listening ability and performance on
            an independent writing test and an integrated EAP listening-into-writing test? .....
  3.4.2 RQ2: To what extent does output performance on an independent EAP
            writing task predict output performance on an integrated EAP listening-into-writing test? ....
  3.4.3 RQ3: How does the input format of an academic lecture affect performance on
            an integrated EAP listening-into-writing task? ...........................................
  3.4.4 RQ4: How do learners’ perceptions of their performance compare with observed
            performance on an integrated EAP listening-into-writing task? .....................
  3.4.5 Questionnaire results ........................................................................
  3.4.6 Conclusions from the exploratory study findings ...................................
3.5 Changes for the main study .................................................................. 172
  3.5.1 Implications for study design ............................................................ 172
  3.5.2 Implications for task design ............................................................. 178
3.6 Conclusion .......................................................................................... 188
4. Main study: Methodology ...................................................................... 189
  4.1 Participants ......................................................................................... 189
  4.2 Ethical consent ................................................................................... 194
  4.3 Research Questions and hypotheses ..................................................... 194
Appendix 5: Exploratory Study - Lecture Transcript: Culture Shock ........................................ 369
Appendix 6: Exploratory study - IELTS cohesive devices ...................................................... 375
Appendix 7: Exploratory study - EAP cohesive devices .......................................................... 377
Appendix 8: Exploratory study - EAP lecture propositions ..................................................... 380
Appendix 9: Exploratory study - Percentage of propositions reproduced ............................... 388
Appendix 10: Exploratory study - Number of reproductions from the EAP slides ................. 389
Appendix 11: Exploratory study - Screenshot of input – part three ........................................ 390
Appendix 12: Exploratory study - Feedback questionnaire ....................................................... 391
Appendix 13: Main study - Information sheet and consent form ............................................ 396
Appendix 14: Ethics form ........................................................................................................ 400
Appendix 15: Main study - Quick Placement Test (OUP, 2001) .............................................. 408
Appendix 16: Main study - INTEGRATED EAP TEST .......................................................... 418
Appendix 17: Main study - PowerPoint Lecture Slides ............................................................ 425
Appendix 18: Main study - Lecture Transcript: Culture Shock ............................................. 427
Appendix 19: Main study - Text Inspector analysis of lecture language and slides .......... 433
Appendix 20: Main study - Text Inspector CEFR analysis ...................................................... 434
Appendix 21: Main study - Text Inspector AWL analyses ....................................................... 439
Appendix 22: Main study - Feedback questionnaire ................................................................. 441
Appendix 23: Main study - Example of a coded paper ............................................................. 446
Appendix 24: Main study - Culture shock lecture: expert panel propositions ....................... 447
Appendix 25: Main study - Propositions reproduced from part one: both groups and totals ..................................................................................................................... 449
Appendix 26: Main study - Propositions reproduced from part two: both groups and totals ..................................................................................................................... 451
Reference list ............................................................................................................................. 452
List of Tables

Table 3.1: Breakdown of participants in the exploratory study ........................................... 114
Table 3.2: Word length and time for input sections .............................................................. 118
Table 3.3: Descriptive statistics for test tasks ........................................................................ 123
Table 3.4: Spearman Rank Order results for test tasks .......................................................... 124
Table 3.5: Kendall’s tau results for test tasks ........................................................................ 125
Table 3.6: Non-parametric test results for cohesive devices .................................................. 129
Table 3.7: Descriptive statistics for Coh-Metrix analyses ........................................................ 130
Table 3.8: Percentage increase in word length across the two essay types ......................... 131
Table 3.9: Coh-Metrix content analyses ................................................................................. 135
Table 3.10: Comparison of cohesive devices .......................................................................... 142
Table 3.11: Word counts for each of the four sections of the essay ....................................... 146
Table 3.12: Analysis of performance ...................................................................................... 154
Table 3.13: Strongest and weakest performance comparison ................................................ 155
Table 3.14: Features of the study design to be changed ......................................................... 168
Table 3.15: Features of the task design to be retained .............................................................. 174
Table 3.16: Refinements to the task design ............................................................................. 174
Table 3.17: Speech rates by section ......................................................................................... 178
Table 3.18: Linguistic features of input ................................................................................... 180
Table 4.1: Background information of the test takers ............................................................. 188
Table 4.2: Data collection methods and analyses .................................................................... 192
Table 4.3: Descriptive Statistics and Lexical Diversity measures: transcript parts one and two ......................................................................................................................... 199
Table 4.4: CEFR levels – Type percentages for lecture transcript parts one and two ............. 200
Table 4.5: CEFR levels - Token percentages for lecture transcript parts one and two ............ 200
Table 4.6: AWL analyses – Types / token percentages for lecture transcript parts one and two ................................................................................................................................. 201
Table 4.7: EAP lecture: order of input ..................................................................................... 202
Table 4.8: Text Inspector analyses ......................................................................................... 209
Table 4.9: Coding system for Content Analysis ..................................................................... 211
Table 5.1: Descriptive statistics and lexical diversity measures from Text Inspector .......... 228
Table 5.2: Percentage of CEFR types reproduced ................................................................. 231
Table 5.3: Percentage of CEFR tokens reproduced ................................................................. 234
Table 5.4: Percentage of AWL types and tokens analysis ...................................................... 235
Table 5.5: Coding system for coding analysis ....................................................................... 237
Table 5.6: Summary of coding results – number of matches .................................................. 239
Table 5.7: Group results for part one matches ....................................................................... 242
Table 5.8: Group results for part two matches ...................................................................... 243
Table 5.9: Most frequently reproduced propositions in part one across both groups .......... 246
Table 5.10: Least frequently reproduced propositions in part one across both groups ....... 247
Table 5.11: Part one PowerPoint propositions reproduced by the video first group ............. 248
Table 5.12: Most frequently reproduced propositions from part two across both groups ... 251
Table 5.13: Least frequently reproduced propositions from part two across both groups ... 253
Table 5.14: Part two PowerPoint propositions reproductions by the audio first group ......... 254
Table 5.15: Audio first group - Perceived performance compared to observed test performance ................................................................................................................................. 260
Table 5.16: Audio first group - Person by person count of matches between perceived and observed best performances ................................................................................................. 261
Table 5.17: Audio first group - Person by person count of matches between perceived and observed worst performances ................................................................. 261
Table 5.18: Video first group - Perceived performance compared to observed performance ........................................................................................................ 262
Table 5.19: Video first group - Person by person count of matches between perceived and observed best performances ................................................................. 263
Table 5.20: Video first group - Person by person count of matches between perceived and observed worst performances ................................................................. 263
Table 5.21: Q12. Any other factors that affected your performance? ............................. 275
List of Figures

Figure 5.1: Performance by group ................................................................. 224
Figure 5.2: Performance by input mode .......................................................... 225
Figure 5.3: Ratio of reproductions to total word count for audio first group: audio input (part 1) .................................................................................................................. 255
Figure 5.4: Ratio of reproductions to total word count for video first group: video input (part 1) .................................................................................................................. 256
Figure 5.5: Ratio of reproductions to total word count for audio first group: video input (part 2) .................................................................................................................. 257
Figure 5.6: Ratio of reproductions to total word count for video first group: audio input (part 2) .................................................................................................................. 257
Figure 5.7: Reasons for best performance ......................................................... 264
Figure 5.8: Reasons for worst performance ....................................................... 266
Figure 5.9: Q1. How well did you do on the listening test? .............................. 268
Figure 5.10: Q2. How difficult was the writing test? ........................................ 269
Figure 5.11: Q3. How much did you know about the topic? ............................ 269
Figure 5.12: Q4. What effect did the topic have on your performance? .......... 270
Figure 5.13: Q9. Did anything else have a positive impact on your understanding of the lecture? ................................................................................ 271
Figure 5.14: Q10. Did anything else have a negative impact on your understanding of the lecture? ................................................................................ 272
Figure 5.15: Q11. Overall, how difficult was the essay? ................................. 275
1. Introduction

This thesis will investigate how performance on an integrated English for Academic Purposes (EAP) listening-into-writing task for students in higher education differs according to input format.

This chapter will first set out the background to the study, including both the bigger picture behind the study and my personal impetus for the choice of study. The former will briefly outline the development of international student mobility in the UK and further afield, and how this has led to the need to ensure that students wanting to study through the medium of English have a suitable language level, irrespective of whether they are studying in an English-language environment or through English-medium instruction in a non-English-language environment, which, in turn, has led to a number of developments in English language testing for university entrance purposes over the last 60 or so years. These changes in testing were the basis for my personal interest in the study.

1.1 Background: The development of international student mobility and English-language instruction in higher education

International students have been coming to the UK since before World War II. As far back as 1935, it was mainly European students but by 1943, students from the United States, Turkey, Latin America and China were studying in the UK (Eastment, 1982, p. 60-62, in Weir and O’Sullivan, 2017, p. 107-108). The number of international students coming to the UK continued to grow over the following 50 years such that in 1996, well over 200,000 students came to study here (p. 109). The exponential growth continued into this century and in 2017-18, there were 458,490 international students studying in the UK, a good proportion of whom (n=106,530) were Chinese (UKCISA, 2019). Although growth has slowed down in the last 10 years, the UK still is and has long been the second most popular destination for international students (Marginson, 2018, p. 1).
The growing importance of internationalisation and the desire to improve the reputation of higher education institutions as well as to attract international students has led to a huge increase in demand for both English language education and English Medium Instruction (EMI)\(^1\) (Galloway, Kriukow and Numajiri, 2017, p. 4). The “‘galloping’ phenomenon’ of ‘pandemic ... proportion’ (Chapple, 2015, p. 1) that is EMI has experienced exponential growth in Europe, increasing from 725 courses\(^2\) identified in a 2001 survey to over 8089 in 2014 (Wächter and Maiworm, 2014, p. 15-16). Thus, while internationalisation involves much more than merely offering courses in English, EMI courses have been described as the ‘most significant trend in educational internationalisation’ (Chapple, 2015, p. 1).

In order to gain access to these courses, students need good language skills because they are required to meet admissions criteria, the most important of which are their ‘academic and intellectual potential’ and their English proficiency, which are often measured by the applicants’ certificates, an English language test, an entrance test or a combination of the above (Wächter and Maiworm, 2014, p. 71-72).

Internationalisation ‘has become a priority for higher education institutes’ (Galloway, Kriukow and Numajiri, 2017, p. 4) as universities try to establish a ‘global presence’ and become more competitive on a global scale (p. 4). For the students, however, there is also the ‘moral and financial need to ensure that fee-paying international students who undertake university education are capable of succeeding’ (Feast, 2002, p. 70). Since language proficiency has been shown to be a major challenge to students studying in an English-speaking environment or through the medium of English (Belhiah and Elhami, 2015; Murphy, 2011), it is in everyone’s best interests to ensure that the students have a sufficient level of English. As such, there is

---

\(^1\) Dearden (2014, p. 4) defines EMI as ‘The use of the English language to teach academic subjects in countries or jurisdictions where the first language (L1) of the majority of the population is not English’.

\(^2\) The authors recognise that this figure may have been higher in 2001 but they were unable to identify any more while in 2014, the availability of information online was much more widespread.
a huge demand for tests which can be used as evidence of language proficiency in EAP for university entrance purposes.

**1.2 Language testing in EAP**

As far back as 1953, concerns were raised about the language level of students coming to the UK to study and the need to ensure that students coming to the UK had the necessary language competence (Phillips, 1953, in Weir and O'Sullivan, 2017, p. 111). The use of a pre-departure language questionnaire to ensure proficiency in all four language skills was introduced as the first type of ‘language test’ for students coming to the UK. As dissatisfaction with the questionnaire grew over the next 10 years as student numbers increased, the need for a more appropriate language test became apparent as the costs of selecting students with inadequate levels of proficiency, who were subsequently ill-prepared for English-medium higher education, were no longer acceptable (p. 113). This led, initially, to the Davies Test / English Proficiency Test Battery (p. 114) and other tests, which tested linguistic aspects of English. While this was useful to measure knowledge of the grammatical system and lower-level processing skills, it became clear over time that tertiary level students needed higher-order processing skills and these were not being measured by the tests in use at the time (pp. 132-133).

Conversations in second language pedagogy at around the same time were focussing on communicative competence and communicative language ability (Hymes, 1972; Canale and Swain, 1980). This led to a paradigm shift in language teaching from earlier approaches and methods such as grammar-translation or the Direct Method to a communicative approach to language teaching, which focussed on the use of more authentic tasks in the classroom. With the development of frameworks of communicative competence in the classroom came the

---

3 Canale and Swain (1980, p. 27) describe communicative competence as comprising competence in grammatical, sociolinguistic and pragmatic competence. As such, they posit that a communicative approach must be based around the learner’s ‘communication needs’ (p. 27).
need to reflect this authentic use of language in tests. It was against this backdrop that first the ELTS test (the predecessor of the IELTS test) and later the IELTS test emerged.

The IELTS test is a four skills test for university entrance. Unlike earlier tests, which tested structural aspects of the language such as grammar and knowledge of phonology, the IELTS test is designed to reflect ‘some of the features of academic language’ (IELTS, 2019) with the intention of assessing a candidate’s readiness to study in an English language environment.

The IELTS test has become one of the most widely accepted tests for university entrance and, along with the TOEFL test, has dominated the market for university entrance tests in recent decades.

However, more recently, there has been a drive towards more integrated skills assessment in both large-scale assessments and in classroom use (Plakans and Gebril, 2012, p. 217), because this reflects the demands of the academic environment better (Cumming, 2014, p. 224). Bachman and Palmer (1996, p. 18) define the environment where the language is used as the ‘target language use’ (TLU) domain while the language use task used in that domain is a ‘TLU task’ (p. 18). It is these TLU tasks that a test should reflect. In an EAP domain, the TLU tasks involve understanding both written and spoken texts as well as integrating knowledge from these input formats into extended discourse, which itself could be written or spoken.

Shaw and Weir (2007) point out that the increased interest in integrated EAP testing is due to the fact that “real-life performance ... [is] seen as the criterion of choice ...” (p. 17).

Currently, then, the academic environment in both the UK and around the world is experiencing a push towards integrated skills assessment as a way of reflecting the needs and, indeed, the cognitive processing demands of students in higher education.

1.3 My personal and professional interest in teaching and testing EAP

Within this wider context, I found myself being promoted to Course Manager for the International Foundation Programme at a UK university with responsibility, among other
things, for the EAP provision on the course. The intention of the Foundation Programme was to take students from an entry level of IELTS 4.5 to the university entrance level of IELTS 6.0.

One of the key issues I faced in this role was that, in the days before the UKVI regulations on students’ language levels and the requirement to have a Secure English Language Test (SELT), large numbers of students were being admitted to the course with, in reality, a much lower proficiency level than would be expected of someone with an IELTS level of 4.5. This was due, in no small part, to a lack of understanding on the part of admissions officers about which tests were acceptable for university entrance purposes and what levels in these tests equated to the required IELTS levels. As such, students with test results from various sources and at various levels were being accepted as ‘equivalent’ to the required IELTS level.

This naturally led to students having problems with the course as they were unable to meet the demands of the preparation course, much less the demands of academic study. I became interested in assessing language proficiency for university entrance purposes and in order to design what I felt was an authentic test which reflected the TLU domain, I needed to know what the TLU domain was. Therefore, I undertook a survey of lecturers (n=100), along with a colleague from another university, to investigate the ‘real-world’ needs of students studying for degrees in UK universities so that I could better prepare my students on the Foundation Course but also design an entrance test which would reflect students’ real-world needs, and which could replace the multiple choice lexico-grammatical and phoneme differentiation test that was being used as the main university entrance test at the time.

As one might expect, the results of the study showed that a frequently cited listening task was to listen to lectures and take notes (90.4 per cent). In terms of written output, 50 per cent of participants required their students to write a summary or a précis. While other written tasks such as writing reports / projects (72.3 per cent), written responses based on lecture input (61.7 per cent) and essays (60.6 per cent) were more frequently cited (Westbrook and Howell,
the ability to summarise and paraphrase information is a skill that is also needed for these other tasks.

Further to the study outlined above, I designed an integrated EAP test, which included, among other sections, a listening-into-writing test, which I used for a number of years alongside a lexico-grammatical placement test, the Oxford Quick Placement Test (QPT), to assess students when they arrived and to place them in groups on the basis of these results. The two tests together gave me a good insight into their underlying lexico-grammatical knowledge and provided evidence of the students' level and ability in relation to their EAP skills. In an effort to gather validity evidence for this test, it was important to investigate the extent to which my test demonstrated concurrent validity with non-integrated tests such as IELTS. I also wanted to develop a better understanding of how different input formats in a lecture could impact on output performance in an integrated listening-into-writing task. These two areas became the focus of the exploratory study in this thesis while the second area became the focus for the more in-depth, main study of the PhD.

1.4 Aims of the study
As integrated assessment is designed to reflect the demands of real-life academic practice, it is important to consider what academic practice looks like in today's technology-influenced world. More and more universities are 'flipping their classrooms' and placing content online for students to access outside of the class so that contact time can be used for more interactive purposes. Indeed, the current Covid-19 pandemic is evidence of how quickly we can be forced to change from a face-to-face environment to a completely online environment. As a result, there is a question around how best to present the input: should this be in the form of a podcast (i.e. audio-only input) or a vodcast (i.e. video-based input)? While this is a question for the teaching of the input, so, too, is it a question for testing the competence of students wishing to study in higher education.
This study therefore aims to investigate the differences in performance on an integrated EAP test of listening-into-writing when test takers are presented with lecture input in the form of a podcast and a vodcast. It aims to answer the question of which of the two input media results in a better score for the test taker as well as investigating the extent to which test takers focus on the written input in the vodcast compared to the content from the speaker. Finally, it also aims to examine which of the two input media students prefer. This research will therefore have implications for both language teaching and testing.

1.5 Structure of the thesis

The thesis is divided into seven chapters. Chapter 1, the Introduction, has outlined the demand for EAP testing both in the UK and around the world. It has also presented a very brief overview of the developments in EAP testing that have led to the need for integrated assessments.

Chapter 2 discusses the literature on listening and writing processes, factors affecting listening, input media, real-world academic writing, independent and integrated writing tasks and text difficulty.

Chapter 3 discusses the research questions for the exploratory study as well as the methodology, results and discussion for the exploratory study. It also discussed the changes made to the main study on the basis of the exploratory study results.

Chapter 4 presents the research questions (RQs) and discussed the methodology for the main study in detail, including how the data will be collected and analysed.

Chapter 5 presents the findings from the data collection with regard to the RQs outlined in Chapter 4 while Chapter 6 discusses the results with reference to findings from other studies.

Chapter 7 outlines the limitations of the study and suggests some areas for future research before concluding the thesis with some closing remarks.

The reference list is presented after the Appendices, which can be found after Chapter 7.
2. Literature Review

2.1 Introduction
This literature review begins by giving a brief overview of the processes involved in listening and then discusses factors that affect listening. Studies in listening comprehension which investigate the impact of different input media are critiqued and the research relating to academic listening is also outlined. After that, models of writing and the processes involved in writing are discussed. I then consider the research on independent and integrated writing tasks in the EAP domain, starting with an outline of studies that have investigated what real-world academic tasks involve and how these compare with independent writing tasks, looking particularly at the cognitive demands, authenticity and background knowledge. I go on to define integrated test tasks and task types, focussing specifically on summary writing, as well as highlighting some criticisms of integrated tasks before examining the literature comparing independent and integrated test tasks. After that, a review of the literature on integrated tasks is presented, focussing on factors affecting performance on integrated tasks. Finally, the literature around text difficulty is considered, since this relates to both the comprehensibility of the input and the learner’s performance on the output.

2.2 Listening
2.2.1 What is listening?
In order to teach or test listening, it is imperative to understand what listening is. Contrary to Buck’s (2001, p. 148) assertion that there is no evidence to suggest that L2 listening differs from L1 listening, Field (2013) posits that ‘the L2 listening experience differs markedly from the L1 listening one for most listeners, with a much greater reliance on remedial strategy use’ (p. 104). He argues that performance in L2 listening can be affected by both linguistic knowledge (including cultural knowledge) and cognitive processing.

Linguistic knowledge can affect comprehension for lower-level learners since they may be lacking the phonological knowledge necessary to decode the incoming input, and may
therefore rely on word-level matching, which may distort the overall message. Similarly, if they do not possess sufficient lexical knowledge, this may also hamper their attempts to understand the input. Learners need to be able to recognise words presented in both the written and spoken format but also need to be able to recognise these as parts of larger chunks rather than individual words, for example, when presented as connected speech, contracted forms or stressed versus unstressed forms, all of which can affect the speed at which listeners can process the incoming signals and thus parse the information to make sense of it (Field, 2013, p. 104-105). Field (2008, p. 113-116) outlines a model of listening which builds up from ‘phoneme-level’ decoding (p. 115), in which listeners identify syllables then go on to word level decoding. After this, the listener starts to identify chunks of language then adds in grammar and intonation to lead to their being able to piece together the meaning of an utterance. To get to this stage, the learner draws on lower-level processes comprising phonological, lexical and syntactic knowledge (Field, 2013, p. 97). Building on his earlier model (Field, 2008), Field (2013, p. 101) suggests that once the speaker has pieced together the meaning of a proposition, higher-level processes are brought into play. Pragmatic knowledge, external knowledge (the speaker’s background knowledge) and knowledge of the current topic as well as discourse representation all feed into the mix, to enable the learner to recognise the speaker’s intentions and work out the context (p. 101). It is at this point that cultural knowledge can also affect comprehension as the cultural background knowledge that learners bring to a situation may affect their interpretation of the meaning. Similarly, a lack of language proficiency can affect their cultural knowledge if learners do not understand pragmatic meaning and are therefore unable to understand the speaker’s intentions (pp. 105-106). Understanding the speaker’s intentions and the context, in turn, enables the learner to build up, via inferencing and referencing, to a representation of the meaning. From here, learners bring in their external knowledge and, through a process of monitoring their understanding of the input, this leads to a representation of the discourse (p. 104).
The second factor that can help or hinder comprehension is cognitive processing. One aspect of this relates to the speed at which the various steps outlined above happen. Field (2008, p. 115) highlights the fact that, for the novice listener, this may be quite challenging – and quite a slow process – while for the more experienced listener, this happens both automatically and effortlessly (p. 115). This is an important aspect of listening comprehension as decoding, unlike reading, where the reader can go back over the input at their own speed, needs to happen in real-time and, for the novice listener, this can lead to problems keeping pace with the incoming auditory signals. If the listener is unable to decode and process the incoming information in real-time, then they fall behind and miss important details, which leads to a lack of comprehension. Conversely, the more automatic the listener’s ability to decode, the easier it is to follow what is being said (Buck, 2001, p. 7).

Thus, in order to decode the incoming information, it is clear that the listener must possess a certain amount of knowledge including grammar, vocabulary and an understanding of different types of discourse. Processing this kind of knowledge to make sense of the incoming input is known as ‘bottom-up’ processing (Buck, 2001, pp. 1-2). We can also use our knowledge of the world and topical knowledge to ‘piece together’ the information even if we do not understand every word. This is called ‘top-down’ processing (p. 2). In reality, however, we often employ both of these types of processing to help us decode input.

In a testing situation, the test taker may engage in cognitive processes to help with comprehension but may also fall back on communication strategies which can assist with comprehension when a lack of language knowledge causes a problem. Another option is to employ test-wiseness to cope with a lack of listening ability. It is hoped that the processes that the test taker engages in reflect those required in the TLU domain to provide cognitive validity evidence for the test. However, if test takers engage in test-wiseness strategies, this may lead to an over-inflated estimate of their listening ability, which represents a challenge to the cognitive validity of the test (Field, 2013, p. 107). Regarding the use of communication
strategies, these can ‘extend the comprehension of a listener well beyond what his/her knowledge and expertise might otherwise permit’ (p. 108). As such, Field (2013) recommends that strategic competence be considered a key aspect of L2 listening proficiency.

Rost (2016, p. 1) proposes a similar but slightly different model of listening based on neurological, linguistic, semantic and pragmatic processing arguing that neurological processing involves ‘hearing, awareness, consciousness, and attention, … activating a kind of experiential field in which all other processes operate’. Neurological processing and linguistic processing relate very closely to the lower-level processes, that is the linguistic knowledge discussed in Field (2013), and to Field’s model of listening (Field, 2008, p. 117). Linguistic processing relates to the linguistic input that the listener has to decode and the steps that the listener goes through to decode the input and parse the information into ideational content. Semantic processing consists of meaning-making by activating and integrating ‘memory and prior experience into understanding’ (Rost, 2016, p. 1). Finally, pragmatic processing centres on the ‘social and cultural context’ of listening (p. 1). While pragmatic processing is in line with Field’s (2013) thoughts on the impact of cultural knowledge (Field, 2013, pp. 105-106), Rost’s (2016) descriptions of semantic processing and pragmatic processing relate to Field’s higher-level processes which focus on meaning and discourse construction (Field, 2013, pp. 100-103).

2.2.2 Idea units rather than sentences
We have already seen that, although reading and listening are both receptive skills, there is a big difference between them in terms of processing time. Another point to consider about listening that is different to reading is that people do not usually talk in full sentences. Instead, we talk in ‘idea units’ (Chafe, 1979, cited in Field, 2013, p. 121). These are short phrases or clauses which may appear ungrammatical yet make sense due to their semantic contribution to the discourse. This can be challenging for learners as they need to make links between the different idea units in order to make sense of the discourse (p. 122).
2.3 Factors affecting listening
Several factors have been found to affect listening comprehension. Bloomfield et al. (2010) divides these into characteristics relating to the listener, the text and the testing conditions. Listener characteristics include working memory, anxiety and L2 proficiency while text characteristics relate to text length, text complexity, and auditory features such as accent and speech rate. The testing conditions consider number of playbacks and note-taking (p. 1). L2 proficiency has been touched up on Section 2.2.1 above and will also be dealt with in Section 2.9. Linguistic complexity will also be dealt with separately in Section 2.9. Note-taking will be dealt with separately in Section 2.4.1. The other factors mentioned above will be outlined in the following sections. In addition, other factors such as background knowledge and input format have been found to impact on performance so these will be discussed below, too.

2.3.1 Listener characteristics
2.3.1.1 Working memory
Working memory is linked to discourse building as discussed in Sections 2.2.1 and 2.2.2 above because the longer the listener has to keep the idea units in their head before they can construct the discourse, the better their working memory needs to be (Elliott, 2013, p. 38). Similarly, working memory links back to automaticity discussed in 2.2.1 above because the more automatic the lower-level decoding skills, the more working memory is available for higher-level processes such as inferring meaning, understanding the speaker's intentions and organising the input (Field, 2013, p. 107).

Automaticity of processing, bottom up and top down processing, the ability to link idea units and working memory are all cognitive characteristics of listening which can affect the listener’s ability to understand the input.

2.3.1.2 Anxiety
Anxiety may also have an impact on listening comprehension, particularly in combination with the factors mentioned above. Indeed, several studies have investigated L2 listening anxiety
from different angles: Elkhafaifi (2005) looked at listening anxiety in combination with foreign language learning anxiety while Chen (2007, cited in Graham, 2011) looked at listening anxiety as a predictor of L2 listening scores. Capan and Karaca (2013) correlated listening anxiety and reading anxiety while, in a novel study, Brunfaut and Révész (2015) investigated ‘the influence of [working memory] and listening anxiety on listening difficulty’ (p. 142) and found a significant negative correlation between anxiety and listening performance.

As mentioned above, anxiety has been found to correlate negatively with listening performance (Brunfaut and Révész, 2015; Elkhafaifi, 2005). In a study carried out by Field (2015), he intentionally omitted any reference to stress or anxiety regarding the number of playbacks when interviewing his subjects. Nonetheless, as he reports, over 50 per cent of his respondents ‘freely volunteered the information that they felt less nervous, more relaxed or more confident when permitted a second play’ (original emphasis) (p. 36) compared to their experience doing an earlier test with a single playback.

The studies above all found that anxiety played a role in listening performance. Therefore, in order to minimise anxiety, it is important that test designers consider how they can get the best out of the test takers. This could be achieved by ensuring that test takers are familiar with the types of tasks they will be asked to complete or, as Field (2015) suggests by allowing a second playback.

2.3.1.3 Topic or background knowledge in receptive skills tests
As discussed above, listening involves using both top-down and bottom-up processing to decode the input. Thus, one would expect that topic knowledge or background knowledge would play a role in comprehension. This has been investigated with respect to both listening comprehension and reading comprehension but the findings have been contradictory.

With regard to the former, Long (1990) investigated the effect of background knowledge on L2 Spanish listening comprehension of two topics: the gold rush and the band, U2. Her results indicate that, for the gold rush text, students could rely on their linguistic knowledge when
they lacked topic knowledge but for the U2 text, where participants had more background knowledge, both topic schemata and linguistic knowledge worked together. On the basis of these results, she suggests that the combination of linguistic knowledge and background knowledge led to students being more confident about their levels of comprehension of the passages, which was illustrated by the longer recall protocols that they wrote after listening. On the other hand, she highlights the fact that a subgroup of 13 of the participants - who 'considered themselves average or very good listeners’ (p. 72) - wrote recall protocols which demonstrated the dysfunctional effect of background knowledge. These participants wrote comments which were incongruous with the content of the listening texts, introducing inconsistencies related to temporal details regarding the gold rush, and both phonological and recasting inconsistencies for the U2 text, demonstrating that their background knowledge was leading to a negative effect on their recall protocols (pp. 72-73).

On the other hand, investigating whether performance in an EAP listening comprehension test was improved as a result of prior study of a lecture topic, Jensen and Hansen (1995, p. 99) found that background knowledge had a 'significant main effect in five of the 11 lectures' that their participants listened to. They stress, however, that the 'prior knowledge variable ... does not appear to create anything more than a trivial differential effect’ (p. 113) between participants with or without prior lecture topic knowledge. They also highlight the fact that this difference is more likely to be present in technical rather than non-technical lectures.

In a test of reading comprehension, Alderson and Urquhart (1985, cited in Clapham, 2000, p. 514) found that, although some groups of students performed better on texts that were from their subject specialisms, this was not always the case and it was dependent on the discipline. Similarly, they found that, in some cases, students from some disciplines, e.g. Science and Engineering, performed better than students from other disciplines, e.g. Business and Economics, yet this also did not happen consistently.
In a famous study, Clapham (1996) investigated the issue of topic knowledge on performance on the IELTS test and found that, out of the three topic areas in the reading module, students generally performed better on the module in their subject area. This raises the issue of bias towards and against students from different subject areas.

Similarly, Lee and Anderson (2007) found that using certain topics may bias against some examinees while being particularly advantageous to others who are studying a similar subject.

Clapham (2000) posits that language proficiency may play as important a role in comprehension as background knowledge and suggests that a general topic may remove any bias.

2.3.2 Text characteristics

2.3.2.1 Text length and fatigue

Another factor which can hinder performance in listening comprehension is the length of the input. McGarrigle et al. (2017, p. 95) state that ‘effortful listening’ can lead to stress and fatigue while mental fatigue is the slowdown in cognitive function brought about by an extended period of mental effort (DeLuca, 2005, cited in McGarrigle et al., 2014, p. 5).

Flowerdew (1994, p. 11) highlights the fact that listening in an academic environment often involves listeners needing to be able to ‘concentrate on and understand long stretches of talk’ without being able to interrupt or engage in any kind of interaction with the speaker.

Underwood (1989) argues that there are no specific rules about the length of texts that should be used for listening comprehension. On the other hand, James (1986, p. 41 quoted in Long, 1990, p. 68-69) states that an L2 listening text which is more than three minutes long will ‘induce sensory overload’ unless the participants consider the topic to be interesting and entertaining. He recommends that, for classroom purposes, texts should be no longer than one minute long (p. 69). Unfortunately, while this may be helpful for developing listening skills in a classroom environment, in an academic environment, students are unlikely to experience such short texts in lectures.
Hamouda (2013) found that his Saudi undergraduate students stated that text length was one of the reasons for their lack of understanding of English texts and that they experienced fatigue with the longer texts. However, unfortunately, there was no mention of the length of the texts being referred to in the study. On the other hand, in his study looking at note-taking in university lectures, Locke’s (1977) participants took notes on authentic lectures of between 50 and 70 minutes in length. Using ANOVA, he analysed the percentage of thought units produced during different time intervals during the lectures (0-20 minutes, 20-40 minutes and 50-70 minutes) and found a significant fatigue effect with the completeness of notes dropping steadily over the three time intervals from 57 per cent in the first 20 minutes to 51 per cent in the second 20 minutes and down to 46 per cent in the third interval. Similarly, O’Malley, Chamot and Küpper (1989, p. 428) found that the students in their study were affected by the length of their texts and thus stopped paying attention to the input but, unfortunately, they also do not give any information about how long the input was. Nonetheless, they do point out that, despite being affected by fatigue, the effective listeners noticed when their attention was lapsing and refocused. Conversely, Griffiths (1990, p. 331) did not find any effect on scores between text lengths, although he does concede that more specific research on text length and difficulty needs to be done to fully understand how comprehension and text difficulty are related.

2.3.2.2 Accent

One factor which can impact on learners’ ability to decode aural input is familiarity with the speaker’s accent. Traditionally, listening comprehension tests and materials were recorded by native speakers with standard accents, be they British English, American English or Australian English. However, in recent years, developments in English as an International Language (EIL) have led to calls for the inclusion of non-native speaker accents in an effort to move away from ‘native speaker norms’ (Harding, 2011, p. 163). However, there is debate about which accent is best for aiding comprehension in a listening test: a standard native English accent
or a local L2 accent (by which is meant the L2 accent which is predominant in the area where the test takers are) (Ockey and French, 2014, cited in Ockey, Papageorgiou and French, 2016, p. 85). Some research found that the standard accent was easier to understand (Ockey, Papageorgiou and French, 2016) while others have found that the local L2 accent was easier for some test takers but made no difference to others (Harding, 2011). Abeywickrama (2013) found no significant difference in scores among non-native speakers (Brazilian, Sri Lankan and Korean) listening to accents similar to their own (Sri Lankan and Korean) or different to their own (Chinese, United States). Berkowitz (1979, cited in Eisenstein and Berkowitz, 1981, p. 75-76) found that preferences varied between the proficiency levels of the test takers, with lower level learners claiming that non-native speakers were easier to understand while higher proficiency students found it easier to follow native speakers. Of course, how different the accent is may have an impact on comprehension but with the exception of Ockey, Papageorgiou and French (2016), these studies did not consider this. Nonetheless, since accent has been found to have an impact on comprehension, this may impact the test takers in the current study.

2.3.2.3 Speech rate
Speech rate has been widely researched with regard to its effect on listening comprehension. Most lectures have been found to be delivered at between 100-180 words per minute (wpm) (Carver, 1982, cited in Robinson et al., 1997). While 188 wpm might be considered an average speed (Griffiths, 1992), Ladas (1980, in Robinson et al., 1997, p. 260) found that 135 wpm was the maximum speed at which students could take notes. Unsurprisingly, then, fast speech (200 wpm) has been found to cause difficulty for L2 listeners (Brindley and Slayter, 2002; Griffiths, 1990) while moderate speech (150 wpm) does not appear to have a significantly different effect to slow speech (100 wpm) (Griffiths, 1990, p. 311).

Robinson et al. (1997) investigated whether lecture rate affects comprehension and perceived topic importance. Students listened to either an audiotaped or videotaped lecture, which had
been recorded at either a slow, medium or fast speed. The audiotaped lecture speeds were 102, 155 and 209 wpm respectively while the videotaped speeds were 106, 157 and 200 wpm respectively (p. 262). They found a ‘significant main effect for speaking rate’ with students in the ‘slow lecture group’ outperforming the medium and fast lecture groups (p. 263). They concluded that students can comprehend information better at a lecture rate of approximately 100 wpm than when the information is delivered at the faster rates of approximately 150 and 200 wpm. Although they distinguished between the input modes of audio vs audiovisual input, they point out that they ‘were not able to experimentally examine differences in paralinguistic nonverbal information’ such as pauses, tone and volume (p. 265). Similarly, Brindley and Slayter (2002) found that speech rate and item format influenced text difficulty.

Révész and Brunfaut (2013) looked at the impact of speed of delivery, linguistic complexity and explicitness of the input text to find out what had the biggest impact on comprehension. In contrast to the studies above, they found no effect for speech rate as an individual predictor of difficulty, although they do concede that there may have been an effect when combined with other factors.

2.3.3 Testing conditions

2.3.3.1 Single or double playback

One of the key debates around listening is how many times to listen to a recording. Buck (2001, pp. 170-171) states arguments on both sides. He argues that single play may best replicate the authenticity (Bachman and Palmer, 1996, p. 23) and interactiveness (p. 25) of the real-life situation, where a listener would usually only hear a text once and where automatic, real-time processing would be required. On the other hand, in most real-life listening situations, rather than needing to understand the content in-depth (as is the case in many testing situations), it is not usually necessary to understand exactly what is said; instead, real-life listening tends to only require ‘approximate’ understanding. Indeed, if more detailed comprehension is required in real-life, listeners usually have a chance to request clarification.
Buck (2001) argues then that, while a double playback may compromise authenticity, a single playback may compromise interactiveness since there is no opportunity to interrupt and ask the speaker to repeat something or clarify it as necessary.

A compromise may be found in the Finnish National Language Certificates (Buck, 2001, p. 171), in which detail listening questions are played twice but only once for gist listening questions, or the British Council’s Aptis exam (British Council, 2019), in which test takers can choose to listen twice if they are doing the computer-based test. Of course, one important argument in favour of having the opportunity to listen a second time in a testing situation is that we should always try to ‘test for best’ (Weir, 2005, p. 54) and a person’s chances of success can always be hampered by the unfortunate cough or a scraping of a chair.

Another argument in favour of allowing a second playback is that, during the first playback, test takers have the time to normalise to the speaker’s accent, tone and speech rate (Pisoni, 1997, cited in Field, 2015, p. 6).

When investigating how test takers listen during a test in which they hear the recording twice, Buck (1990) found that test takers listened differently between the two playbacks: the first playback tended to result in local listening as test takers located the answers to the questions while the second playback was on a more global level, presumably as they checked or revised their answers.

In a very small-scale study, Field (2009, cited in Field 2015, p. 7) investigated the effects of single play versus double play on listening test outcomes. He found that a second playback was only of benefit to test takers with an intermediate level of proficiency (IELTS scores of 5.0-6.5). However, the opportunity to listen for a second time did not benefit lower or higher-level learners. The former did not improve their scores as they could not decode enough to make any difference while the latter were good enough to find the correct answer the first time through so they did not need to listen for a second time. However, as he points out, this
was only a very small-scale study so the evidence to support the findings for the higher and lower level test takers was not sufficient due to the small numbers at each end of the scale.

In a later, larger study, Field (2015) found that that double play on a lecture task not only benefitted candidates at all levels of proficiency (low, intermediate and high) but that candidates increased their score ‘in a way that was roughly proportionate and that sustained and heightened the distinctions between the proficiency bands’ (Field, 2015, p. 32). Furthermore, contrary to what has been found in other studies (Cervantes and Gainer, 1992; Berne, 1995, both cited in Field, 2015, p. 6), his study revealed that a second playback did not make the test markedly easier.

With regard to the use of audio input, Field (2008, p. 159) also argues that a second playback compensates for the lack of ‘visual and contextual clues’ in an audio recording. Buck (2001, p. 72) argues that if ‘visual information would be available in a target-language use situation, making it available in the test situation is likely to improve both situational and interactional authenticity’.

Clearly, there are arguments on both sides; however, I favour the double play argument as I believe that this replicates the possibility of being able to ask for repetition during a test while in the TLU domain. Furthermore, this is more authentic especially in the current climate where more academic input is being made available online so students have control over how many times to listen.

2.3.4 Input format

Many tests, especially high-stakes tests, still employ the use of audio texts. However, as visual and audiovisual input have become increasingly widely used in the language classroom, this has also been reflected in tests of listening comprehension (Read, 2002, p. 107).

Several studies have investigated the effect of different input media from different angles. The areas of research covered include the impact of context vs content visuals (Ginther, 2002),
the differences in scores across different input formats (Brett, 1997; Coniam, 2001; Gruba, 1993), the use of still images versus video input in computer-based tests (Ockey, 2007; Suvorov, 2008; 2013), performance on audio vs video input medium (Aldera, 2015; Batty, 2015; Brett; 1997; Chang, Lei and Tseng, 2011; Chen, Wang and Xu, 2014; Gruba, 1993; Pardo-Ballester, 2016; Progosh, 1996; Wagner, 2008; 2010a; 2013), the use of subtitles (Sydorenko, 2010), test takers’ interaction with video (Brett, 1997; Gruba, 1994; Ockey, 2007; Suvorov, 2013; Wagner, 2007) and the use of captions in listening comprehension (Leveridge and Yang, 2013; Montero Perez, Peter and Desmet, 2014; Sydorenko, 2010; Vandergrift, 2004). The research in these areas reveals varying results with some studies finding a benefit in the use of video and others finding audio more useful.

Similarly, studies into audio versus video or multimedia input formats have yielded differing results. Some studies have shown that the use of audiovisual and non-verbal input can have a positive effect on performance (Sueyoshi and Hardison, 2005), on vocabulary acquisition (Aldera and Mohsen, 2013, in Aldera, 2015, p. 1985), and on test taker perceptions (Cubilo and Winke, 2013; Ockey, 2007; Wagner, 2010b) although test taker perceptions are not always in line with scores (Pardo-Ballester, 2016; Suvorov, 2008).

Looking at the impact of context vs content visuals, Ginther (2002) found, not surprisingly, that content visuals, which contain information that supports the audio are more helpful than context visuals, which might show a group of people having a conversation, for example, whereas when comparing the use of still images with video interactions, Ockey (2007) found that test takers only interacted minimally with still images whereas the interaction with the video was much more varied.

Suvorov (2008) investigated the impact of context visuals on a listening text using a single photograph, audio-only input and video input. His results revealed that test takers performed significantly worse on the video-mediated lecture than on the other two lecture input formats, where there were no significant differences. However, when the input was a dialogue, he
found no statistically significant difference between the video-mediated input and the other two input formats. On the contrary, he found that video might even contribute favourably to students’ performance.

However, it should be noted that the photograph-based and video-based visuals were context visuals rather than content visuals, which may have proved less useful and effective than if they had been content visuals.

Regarding test takers’ preferences towards the input type, he found that 44 per cent (n=15) of the 34 test-takers preferred listening without any visuals while 35 per cent (n=12) said that they preferred the use of video. Only 21 per cent of test takers (n=7) opted for the use of photographic visuals.

In terms of test takers’ actual performance compared to their preferences, these did not always overlap: the only statistically significant difference here was found between audio and video scores for those who preferred the audio-only input format: in this case, test takers performed ‘significantly better on the audio-only listening passages’ (Suvorov, 2008, p. 48-49). Although Suvorov provides useful insights into test takers’ preferences and their performance, this was only a relatively small-scale study so a larger study may lead to other results.

Conversely, Pardo-Ballester (2016) found that students preferred video-based input compared to audio input in her study looking at the impact of different input formats (audio-only vs video-based input) on performance on inference questions in web-based listening tests of Spanish. The students involved were intermediate level students in the third and fourth semesters of their Spanish courses (Spanish 201 and Spanish 202 respectively), which were both offered in ‘face2face blended’ and ‘online-hybrid’ (p. 93) formats.

The results indicated that the higher level group (Spanish 202) performed better on video-based than audio-based inference questions while the opposite was true for those in the lower
group (Spanish 201). Similarly, Progosh (1996) investigated test takers’ preferences in a test of listening comprehension at a university in Japan and found that his test takers preferred video over audio only.

Like Suvorov (2008), Pardo-Ballester (2016) also investigated test takers’ perceptions of their performance and found clear preferences for working with video over audio across both groups. However, in line with Suvorov’s (2008) findings, the test takers’ preferences did not always coincide with the reality of their results: the Spanish 201 group was more likely to answer inference questions correctly if the input was audio only, despite their preference for video while the Spanish 202 group’s preferences were in line with their results as they were less likely to get the correct answer when presented with audio-only input. She suggests that this could have been because the lower level students had had less exposure to the language and they may have struggled with the cognitive overload of having to cope simultaneously with listening, looking at the visuals, and reading and answering the items.

One shortcoming of Pardo-Ballester (2016) is that, although the sample was large (n=246), it consisted only of intermediate level students so the results may not be generalisable to beginner and advanced level learners. Nonetheless, based on these findings, she advocates using video input in both the teaching and testing of listening.

With regard to the effects of video on English language comprehension, Chen, Wang and Xu (2014) investigated the performance of four groups of test takers who were all exposed to 6 different listening comprehension excerpts in different input modes (audio or audio + video). They also investigated test takers’ preferences regarding the input formats. They found that the effect of video varies depending on language proficiency level, with higher level learners performing well in the video mode while lower level learners performed worse under this condition. They also found that when the video includes information which is relevant to the questions, test takers performed better on these tasks. With regard to the use of video and
audio, only around half of their test takers reported that they only listened and did not pay attention to the video. However, they suggest that this could be because they were not used to having access to the video. Furthermore, these are only self-reported claims by the test takers and the researchers do not have any empirical evidence such as eye-tracking or recording of test takers to draw on to corroborate these claims. In line with other studies (Coniam, 2001; Cubilo and Winke, 2013; Ockey, 2007), some of their participants found the use of video helpful while others considered it a distraction. Findings concur with other studies (Robin, 2007; Vandergrift, 2004) that lower level learners performed better with the help of the written text; however, when this support was removed, they struggled due to an inability to form the schema necessary for listening skill.

Wagner (2010a) also found that students performed significantly better on a listening test containing video input compared to the control group who only had access to the audio input while Coniam’s (2001) findings regarding input formats were mixed. In his study, some of the test takers who took the audio-based test said they would have preferred the video input yet the video group said they did not perceive of any advantage through having access to the video. On the contrary, some felt that the video was distracting, leading Coniam (2001) to conclude that listening tests should only use the audio input format.

On the other hand, Batty (2015) found ‘a very slight facilitative effect’ (p. 16) for the use of video in a video versus audio listening test, but found no significant interaction between delivery format and text type, nor between delivery format and proficiency level thus concluding that the choice of input format be left to the test designer. Nonetheless, he argues that the increased face validity afforded by the use of video, which is more authentic than audio only, may be reason enough to recommend its use.

Researching test taker behaviour in video based listening tests, Wagner (2007) looked into the extent to which test takers made eye contact with the video input. His findings indicated that test takers spent 72 per cent of the time watching three dialogue texts compared to 67
per cent of the time for the mini-lectures. He suggested that this might be due to the 'context-embedded nature of the dialogues’ compared to the 'context-reduced’ nature of the lectures (p. 76).

In contrast to many other studies, Wagner (2007) did not correlate test takers’ scores with the amount of time test takers spent watching the video, which would have been useful given the two different types of texts used in the study.

Building on work by Wagner (2007) and his own 2008 study, Suvorov (2013) carried out an eye-tracking study to investigate how test takers interacted with context and content visuals in both audio-based and video-based, multiple-choice academic listening tests. He also collected information regarding their perceptions on the use of the two input formats as well as comparing the effect of input format on performance. He found that although the type of visual did not impact on scores, the use of eye-tracking software revealed that there was indeed a significant difference between the two types of visuals in terms of viewing behaviour, with a statistically significantly higher mean fixation rate for content videos than for context videos. Participants reported finding content-based visuals useful for a number of reasons, mainly related to the lecture, including aiding understanding of the content (97 per cent), because they were related to the talk (52 per cent), they helped to answer the questions (30 per cent), and helped participants to take notes (9 per cent). In terms of the topic, content visuals helped both those that were familiar with the topic and those that were not (30 per cent). With regard to the context visuals, the usefulness tended to relate to the speaker, with 21 per cent of participants saying that seeing the speaker helped them to focus and another 21 per cent saying that the speaker’s movements attracted their attention and help them to better understand the content.

However, as with other studies looking at the use of video and audio input (Coniam, 2001; Cubilo and Winke, 2013; Ockey, 2007), Suvorov (2013) found that some considered the video input distracting. He also found a statistically significant correlation between test taker’s
perceptions of the usefulness of the visual information and their scores. When test takers perceived the visual input to be helpful, they usually answered the item correctly whereas if they found the visual unhelpful, the items were often answered incorrectly with content videos proving to be significantly more helpful than context videos.

However, although this was an academic listening test, the study focussed only on listening and not on integrated listening-into-writing tasks. Nonetheless, the findings of the study contribute useful insights to our understanding of the influence of different types of visuals in listening comprehension. Firstly, the difference between the focus on the speaker (with the context visuals) and the lecture (for the content visuals) shows that these two different input formats appear to make a difference what students focus on. However, perhaps the most important insight is the confirmation that the content visuals aided comprehension, which is a good argument for using content visuals in teaching.

Mayer’s (2001) Cognitive Theory of Multimedia Learning states that the auditory and visual channels can process matching representations of input. The theory comprises three key aspects: the existence of dual channels, limited capacity for processing, and active processing. The latter refers to the ability to choose the ‘most relevant information as input during information processing’ and combine this with other knowledge (Al-Shehri and Gitsaki, 2010, cited in Leveridge and Yang, 2013, p. 202). This means that, as the audio input becomes too difficult (or is delivered too fast) to process, the listener turns to the visual input mode for support. While this may help comprehension in the shorter term (Leveridge and Yang, 2013, p. 202), the danger is that, instead of developing strategies to help developing real-time processing skills, listeners start to rely on the visual support and become unable to process the aural input when the support is removed thereby causing a ‘short-circuit [in] the development of productive strategies’ (Vandergrift, 2004, p. 10).
Aldera (2015) researched the impact of multimedia on listening comprehension. The study, which was only a quantitative study, investigated performance on audio only versus multimedia input (audio + visual animation) over a series of classes. He found that the group exposed to multimedia scored more highly on the post-tests than the audio only group. Using ANOVA, Aldera (2015) found that the experimental group also performed better in a delayed post-test. In the questionnaire findings, his students expressed a preference for multimedia input. He concludes that his study ‘shows that the effects of multimedia in developing listening skills are greater than those of the use of audio only’ (p. 1987). I should like to point out, however, that Aldera does not appear to recognise that this may not have developed listening skills as the multimedia was a visual animation so the experimental group may have performed better due to the help that was gained by having access to the visual channel. In other words, they may have understood a lot purely from the visuals irrespective of the audio. Thus, it should not necessarily be claimed that this improved the listening skills.

Nonetheless, Aldera’s findings are in line with Brett (1997) who also found the greatest success in the use of the multimedia input compared with audio and video input. The reasons for this include the varied sample sizes in the study, the different proficiency levels of the participants or the varied stimuli. On the contrary, Chang, Lei and Tseng (2011) found that the visual input in their study (text) did not benefit schema construction in the long-term. This study looked at the impact of sound versus text plus sound and its effect on cognitive overload. The participants in this study were randomly allocated to one or other of the treatment groups and took part in two listening tests (one ‘ubiquitous learning’ test (p. 638) and one ‘extended listening’ test). They analysed the findings using MANCOVA and found that the cognitive load in the single mode was higher than the double mode. In line with other studies mentioned above, they found that the double mode outperformed the single mode and that there was an inverse relationship between listening comprehension and cognitive load. In other words, as listening comprehension increased, cognitive load decreased. This is likely to be a result of the
issues around automaticity of processing outlined in section 2.2 above. However, it should be noted that, in this study, access to the text did not support long-term retention (unlike the findings of the delayed post-test in Aldera’s study above). While this is an interesting study, there are a few points to highlight. One is that when they talk about their participants, they do not actually say what level the students are so it is difficult to draw any real conclusions regarding the results linked to proficiency. Another aspect is that the study is not a counterbalanced measures design so the participants are only exposed to one form of input. Finally, the ‘extended texts’ were actually only actually 180-220 words.

While both Brett’s (1997) and Chang, Lei and Tseng’s (2011) study include the use of multimedia - comprising audio, video and text – as well as audio only and video input, they only focussed on listening tests rather than integrating listening and writing. Several other studies have investigated the impact of captions (Vandergrift, 2004; Montero Perez, Peter and Desmet, 2014; Leveridge and Yang, 2013), and the use of subtitles (Sydorenko, 2010).

Montero Perez, Peter and Desmet (2014) carried out a study to investigate performance when presented with video input without captions, with captions and with keyword captions on a listening test. They found that the ‘with full captions’ group outperformed the other two groups on the global comprehension but there was no statistically significant difference on the detailed comprehension questions. However, participants expressed a strong desire for captions in a post-test questionnaire.

Leveridge and Yang (2013) also investigated caption support using a self-designed Caption Reliance Test (CRT) which was used to investigate individual learners’ reliance on captions. In contrast to the findings of Montero Perez, Peter and Desmet (2014), they found that there were individual variations in reliance on the captions but that lower-level test takers appeared to rely on them more heavily. Furthermore, they found a negative correlation between reliance on the captions and L2 ability as measured by the listening test. This raises the question of
whether using captions could actually be detrimental to learners’ progress over the longer term.

Pujolá (2002, cited in Leveridge and Yang, 2013) found that students were answering ‘listening’ comprehension questions by reading the captions rather than listening to the text. In other words, instead of using their listening skills, they were simply reading from the screen. Of course, when visual support is provided through captions, subtitles or written visual input, this is likely to happen. Therefore, the test designer who uses such support needs to consider to what extent this impacts on the construct being tested. In an academic environment, many lectures are accompanied by visual support such as Power Point. Therefore, the way students interact with this support compared to aural-only input can provide us with information about how these two forms of input contribute to success in lecture listening. This appears to be a gap in the literature which this study addressed to inform our understanding of what works better in the teaching situation and therefore how such visual support impacts on the construct of academic listening.

This section has outlined several studies which have compared performance on listening tests when test takers are presented with input in different formats. The results of such research are inconclusive with some test takers performing better using audio while others perform better using video input and still others are most successful when the input is presented in a multimedia format. Similarly, test takers’ preferences are varied with some students preferring video inputs while others find these distracting. To conclude then, since test takers’ preferences for audio versus video as well as content video versus context video have thus far been mixed, the different impacts of seeing the speaker, only hearing the speaker and the effect of both diagrammatic and textual information should be investigated further to see to what extent the test takers are influenced by these different types of input and whether they need to be considered part of the construct of (academic) listening.
2.4 Academic listening

2.4.1 Note-taking

Listening to lectures and taking notes, and participating in class discussions have been identified as listening tasks that learners are required to undertake at university (Ferris and Tagg, 1996; Westbrook and Howell, 2011).

However, understanding academic lectures has long been a challenge for students (De Carrico and Nattinger, 1988, p. 91). Focusing specifically on lecture comprehension, Flowerdew and Miller (1997) highlight four features of authentic lectures: ‘features of spoken language, the interpersonal strategies, discourse structuring and integration of listening with other media’ (p. 32) while Flowerdew (1994, p. 11) presents a number of features that are important for students needing to follow lectures. He points out that understanding lectures requires the knowledge of ‘specialist subject matter’ and the ability to distinguish the important information from the less important and to take notes on the content.

With regard to note-taking, James (1977, cited in Flowerdew, 1994, p. 11) breaks this down into a number of steps: the first two involve processing the incoming information by decoding and understanding it, then identifying key points and noting these down at appropriate times, quickly and clearly. This information may also need to be integrated with information from other sources such as handouts, textbooks (pp. 11-12) or, in the case of my study, with textual and diagrammatic input in a PowerPoint presentation. Note-taking is therefore not a passive activity but one in which listeners need to receive input, process it and produce an output which will be meaningful at a later date, all in real time while the speaker continues to speak. Clearly, then, if aural processing as discussed in 2.2 above is not automatic, this makes note-taking much more challenging.

A few studies have therefore investigated note-taking in lectures. Researchers have looked at what helps students to take notes (Chaudron and Richards, 1986; DeCarrico and Nattinger, 1988). Not surprisingly, several studies have found that discourse which is clearly signposted
can be beneficial for both L1 and L2 students (Chaudron and Richards, 1986; Rickards et al., 1997).

Chaudron and Richards (1986) found that discourse markers which signal major changes of topic had a greater impact on students’ ability to remember the content than sentence and clause-level discourse markers while DeCarri and Nattinger (1988, p. 91) investigated the lexical phrases used in academic lectures and recommend raising awareness of such discourse markers to aid students’ comprehension and to help them better ‘organise and interpret’ the flow of information’ (original emphasis).

As Piolat, Olive and Kellogg (2005) point out, note-taking from lectures is cognitively more challenging than taking notes from a reading text and ‘places more demands on working memory resources’ (p. 303) yet it is ‘widely accepted as a useful strategy for augmenting student attention and retention of academic discourse’ (Carrell, 2007, p. 1), assuming, of course, that the students are able to understand the lecture and take notes.

In an EAP listening test, asking the test takers to take notes and use these to answer questions about the content after the event (rather than answering comprehension questions while listening) would appear to reflect more closely the TLU domain, thereby increasing both the authenticity and interactiveness of the task. Consequently, much of the research on assessing academic listening has focused on this (Carrell, 2007; Chaudron, Loschky and Cook, 1994; Dunkel, 1988; Faraco, Barbier and Piolat, 2002).

---

4 Bachman & Palmer (1996, p. 23-24) state that authenticity ‘relates the test task to the domain of generalization to which we want our score interpretations to generalize’. They differentiate between situational authenticity, which relates to the interaction between the test task characteristics and the TLU domain, and interactional authenticity, which relates to the extent to which the test taker’s language ability is involved in completing the task.

5 Interactiveness is concerned with the interaction between the test taker and the task. The more the test taker has to engage their ‘language knowledge, metacognitive strategies, topical knowledge and affective schemata’ (Bachman & Palmer, 1996, p. 25-26), the more interactive the task.
Nonetheless, assessing notes can be problematic because some students take notes in their L1 (Koren, 1997, cited in Hayati and Jallilifar, 2009, p. 102) so this is difficult when the researcher does not speak that L1.

Chaudron, Loschky and Cook (1994, p. 89) found that the quality of notes varied considerably amongst their test takers but there was ‘no strong or consistent relationship between [their] quality and quantity measures and comprehension’.

Carrell (2007) found that test takers only noted down approximately 20 per cent of the information required to answer multiple choice questions based on a lecture. This is in line with findings from Dunkel (1988) whose students also failed to note down the information tested in the post-listening test. Carrell (2007, p. 42) also found that participants tended to be relatively inefficient in their note-taking, rarely using abbreviations, symbols or paraphrasing. Instead, they tended to ‘write down, in full spelling, content words used by the lecturer’. However, there was a positive correlation between abbreviations and test performance while Liu (2001, cited in Carrell 2007, p. 43) found an inverse relationship between the measure ‘words in full spelling’ (original emphasis) and performance on a post-listening test. As Carrell points out, ‘abbreviations’ is the opposite of words written out in full so her findings lend support to those of Liu.

Furthermore, Carrell (2007) found some positive correlations (although these were not consistent across all groups) between total annotations and performance on multiple choice listening comprehension tasks whereas Liu (2001 in Carrell, 2007, p. 43) found no significant relationship between the two measures. On the other hand, Dunkel (1988) found an inverse relationship between the two in a study looking at L1 and L2 students’ lecture notes compared to listening test performance. Her study revealed that ‘terseness of note taking (involving the recording of lecture propositions) rather than mere quantity’ (p. 270) was a determining factor in retaining information and subsequent success on the post-listening quiz.
On the other hand, when investigating note-taking and performance in TOEFL mini-lectures, Carrell, Dunkel and Mollaun (2004) found that listeners performed better when they were allowed to take notes on the short lectures (2½ minutes) but, on the longer lectures (5 minutes), they performed about the same irrespective of whether they were allowed to take notes or not. However, since even the 'long' lecture is very short, research into working memory and note-taking should look at performance on somewhat longer lectures.

Faraco, Barbier and Piolat (2002) investigated note-taking performance when students were presented with a 12-minute academic lecture, measuring the quantity of notes as a percentage of the total number of words in the input. The results revealed significant positive correlations between the notes and test scores on a follow-up listening comprehension test but the researchers found that paraphrased, reformulated notes correlated negatively with listening comprehension performance, possibly due to the cognitive load of paraphrasing some parts of the lecture leading them to miss other parts of the input.

All these studies inform our understanding of the effect of note-taking, yet only Carrell (2007) went beyond using the information for listening comprehension tests and also investigated performance on an integrated listening-into-writing task (see Section 2.8.4). As such, how students take notes and then use those notes in a follow-up writing task is a potential area of interest for a future study. For example, although the findings reveal that there were positive correlations between notes and test scores on a listening comprehension test, it is not clear if this would also be the case for a follow-up writing task, given that reformulated notes impacted negatively on performance in the listening comprehension test. Therefore, this is another gap in the literature could be useful to address.

2.5 Writing

2.5.1 Models of writing

Research into writing has investigated writing from three different angles. These are: writing as a product, which focuses on the text; writing as a process, which focuses on the writer;
and writing as social interaction, which focuses on the reader (Hyland, 2016, p. 3). This section will provide a brief overview of these three approaches.

### 2.5.1.1 Writing as a product

The first model of writing, text as a finished product, was the dominant one for many years. Within this paradigm, there are several perspectives. One of these considers text as discourse (Hyland, 2016, p. 6). In this case, discourse refers to a ‘socially recognised way of using language’ (p. 9) and considers:

> the purpose and functions linguistic forms serve in communication ... These factors draw the analyst into a wider perspective which locates texts in a world of communicative purposes and social action, identifying the ways that texts actually work as communication (p. 6).

However, in order to write in a specific genre, we have to first identify the ‘organisational patterns’ (p. 9) which constitute different genres. In doing this, there is a risk of oversimplifying a given genre or assuming (possibly incorrectly) that genres do not differ across different discourse communities (p. 11). As such, reproducing a text in a given genre is not as straightforward as merely identifying and reproducing a given pattern of textual organisation.

Another approach is that texts are:

> ... autonomous objects which can be analysed and described independently of particular contexts, writers or readers. Texts have a structure, they are orderly arrangements of words, clauses and sentences, and by following grammatical rules writers can encode a full semantic representation of their intended meanings (Hyland, 2016, p. 4).

In this model of writing, text is seen as being ‘removed from context and from the personal experiences of writers and readers’ (p. 4) and, instead, becomes purely mechanical, with the
focus being on form. In product-oriented approaches, the process of writing is often seen as a linear process, comprising four sequential stages: pre-writing, composing, revising and editing (Tribble, 1996, p. 38). However, such linear, text-based models have been criticised for being too simplistic and straightforward. Raimes (1985, p. 229) argues that writers ‘do not follow a neat sequence of planning, writing and then revising’ because writing is recursive rather than linear. Zamel (1983, p. 167) points out that the linear, product-based approach to writing has a focus on form over content with accurate language being more important than the purpose for writing.

Nonetheless, this approach to writing has been useful as it has led to a wide range of corpus-based research. Researchers have investigated what features of text learners produce at different levels of proficiency (DeCock, 2011 cited in Hyland, 2016, p. 4; Parkinson and Musgrave, 2014; Aull and Lancaster, 2014) while others have looked at differences in changes across text lengths (Kim, 2014) and between independent and integrated texts (Cumming et al., 2005a). We shall return to this research when discussing text difficulty in Section 2.9.

Other research has considered what types of writing tasks are used in integrated tasks. Ascención Delaney (2008) found that the most common types of written tasks used for an integrated task were the summary or the essay response task.

To sum up, Hairston (1982, p. 78) posits that the those who adhere to this text-based approach believe that writers know what they are going to write and simply need to put this into linguistic form and that by teaching editing, they are teaching writing. This is clearly a rather oversimplified view of writing so she further argues that, in addition to looking at the final product, we must also understand how the text develops through the different stages and what cognitive processes the writer goes through while writing so it is necessary to ‘examine the intangible process, rather than ... evaluat[ing] the tangible product’ (p. 84). It is to this writer-centred process approach that we shall now turn.
2.5.1.2 Writing as a process

The second influential model of writing is concerned with the writer and the steps they go through to create the text, rather than focusing on the text itself (Walker and White, 2013, p. 61). This approach recognises that a writer may produce several drafts of a piece of writing which they edit before reaching a final version.

Over the years, a number of researchers (Scardamalia and Bereiter, 1987; Field, 2004; Flower and Hayes, 1981; Flower et al., 1986; Grabe and Kaplan, 1996; Hayes and Flower, 1980; Hayes, 1996; Kellogg, 1996) have described the cognitive processes involved in writing.

Hayes and Flower (1980, cited in Shaw and Weir, 2007, p. 35) see writing as a ‘problem-solving’ activity. Their (L1) model includes three phases: planning, translating and reviewing. The planning process involves gathering, or generating, ideas then organising those ideas to achieve the goals of the task. In the translation phase, the ideas generated in the previous phase are converted into language. Finally, the writer reads and edits the text to improve it in the reviewing phase. These three phases are linked to the task environment, which considers the writing assignment itself, including the topic and the audience, on the one hand, and the text thus far, on the other. The writer’s long-term memory, responsible for providing knowledge about the topic and the audience as well as being used for storing writing plans, feeds into the planning phase to help with the generation of ideas (Hayes and Flower, 1980; Weigle, 2002).

The Flower and Hayes model (Flower and Hayes, 1981, p. 370) is very similar. It also starts with the task environment, but this is broken down into the rhetorical problem, which includes topic, audience and exigency, and the text thus far. The reviewing phase is broken down into evaluating and revising (p. 370). Evaluation, as the name suggests, is the process of evaluating the written text while revision refers to the changes to be made (Flower and Hayes, 1981).
Unlike earlier models in which writing was considered to be a simple linear process, a cognitive process model such as the Flower and Hayes model emphasises the recursive nature of writing (Flower and Hayes, 1981, p. 376) and represents a major step away from the linear approach by focussing on mental processes which are hierarchically organised and which can ‘occur at any time during the composing process’ (p. 367).

The Flower and Hayes (1981) model underwent various iterations through the 1980s (Becker, 2006, p. 26) and in 1986, Flower et al. (1986, p. 23-26) present a model which focuses on the theory of revising a text. The model was divided into Processes on the one side and Knowledge on the other. Task definition is the first process. This is followed by evaluation which involves reading to ‘comprehend, evaluate and define’ the problems (p. 24). Both task definition and evaluation feed across to and are fed into from the ‘Knowledge’ side in terms of ‘goals, criteria and constraints for the texts and plans’ (p. 24). Evaluation also feeds across to the representation of the problem, which may be ‘ill-defined’ or ‘well-defined’ (p. 24). Evaluation is followed by strategy selection and leads to goal setting. The strategies selected may be rewriting or revising and these feed into two other processes, namely, redrafting or paraphrasing, on the one hand, or entering the ‘means-end table’, on the other. The ‘means’ here are ways of dealing with any problems which may have come to light while the ‘ends’ are the problems which have been spotted (p. 26). The next step is to modify the text and / or plan while the ‘means-end table’ also feeds across to ‘procedures for improving the text’ (p. 24). This ‘theory of revision’ (p. 26) sees revision as the interaction between the processes outlined above and knowledge.

In a later work, Hayes (1996, cited in Shaw and Weir, 2007, p. 35) omits the various stages of writing and concentrates on the ‘essential components of the writing process’ (p. 35) and how these components are linked together by cognitive processes.
Hayes and Flower (1980), Flower and Hayes (1981) and Hayes (1996) all focus on L1 writing, while Grabe and Kaplan’s (1996) work investigates L2 writing. Like Hayes (1996), Grabe and Kaplan (1996) also consider the interaction between components of processing and their model, like the models outlined above (Hayes and Flower, 1980; Flower and Hayes, 1981) is also non-sequential. This rather complex model is made up of seven components, each of which has a number of sub-components, all of which interact with each other and themselves, and which ‘must somehow coalesce as multiple interacting strands’ (Grabe and Kaplan, 1996, p. 62). These seven components are:

1. Syntactic structures
2. Semantic senses and mappings
3. Cohesion signalling
4. Genre and organizational structuring to support coherence interpretations
5. Lexical forms and relations
6. Stylistic and register dimensions of text structure

However, this model has been criticised for failing to distinguish between ‘resources stored in the long-term memory’ and ‘the operations of short-term memory’ (Shaw and Weir, 2007, p. 36).

Kellogg (1996) proposes a model comprising three processes:

1. Formulation, which is further broken down into translation and planning
2. Execution, divided into programming and executing, and
3. Monitoring, which includes reading and editing (Kellogg, 1996)

Field (2004, p. 329-331) builds on the work of Flower and Hayes (1980) and Kellogg (1996) with his 5-part model which includes macro-planning, organisation, micro-planning,
translation, monitoring and revising. The first step involves gathering ideas, identifying the goals of the writing and considering how to successfully achieve the task. In the second step, the writer puts the ideas generated in order, finding links between the different points. The micro-planning step is where the planning of the text at both the sentence and paragraph level occurs in the writer’s head, all the while taking into account the goals set in the macro-planning step. However, it is not until the fourth step that the abstract ideas in the writer’s head are developed as linguistic text on the paper. In the monitoring phase, the writer checks the text both at the mechanical level, looking at spelling, grammar and so on, and also at the macro-level in terms of checking that the text is meeting the writer’s expectations and developing the argument appropriately. Finally, revising, as the name suggests, is when the writer implements the changes made further to the previous step.

In terms of monitoring, the extent to which this is done depends on the writer: the less skilled writer is more likely to focus on surface level issues such as spelling and grammar while the more skilled writers may focus on the macro-organisation of the text (Flower et al., 1986).

As can be seen from the above, there is a good deal of overlap between all of these models. Consequently, they are all useful for our understanding of writing. However, with the exception of the work of Hayes and Flower (1980), Flower and Hayes (1981) and Flower et al. (1986), the one thing that all of them appear to be missing is the distinction between how novice and experienced writers produce a text.

Hyland (2002, p. 26 cited in Shaw and Weir, 2007, p. 43) posits that the amount of planning varies between experienced and novice writers but also between L1 and L2 writers, with L1 writers tending to plan more than L2 writers. Eysenck and Keane (2005, p. 418, cited in Shaw and Weir, 2007, p. 43) concur that it is the planning process which distinguishes between writers of differing levels of expertise.
Discussing how mature writers differ from immature writers, Scardamalia and Bereiter (1987, p. 143) point out that ‘the principal difference between mature and immature composing is in how knowledge is brought into the writing process and in what happens to knowledge in that process’. They propose two contrasting models of writing which illustrate the differences between the two types of writers. These are ‘knowledge telling’ and ‘knowledge transforming’.

Knowledge telling is a way of generating a text when presented with a topic and a genre with which the writer is familiar but without the need to plan and organise ideas as would be the case in the Hayes and Flower (1980) model presented above. Instead, so-called ‘topic and genre identifiers’, such as key words related to the topic, are identified and these trigger information in the memory which serves as ideas and content for the writing task. There is a ‘built-in tendency toward relevance’ and knowledge tellers tend to manage to stay on topic without having to plan for cohesion and coherence. The text is built up by ideas being generated until there are no more ideas left to write about (Scardamalia and Bereiter, 1987, p. 144-145).

Knowledge transforming, on the other hand, is a far more complex problem-solving process, which involves analysing the problem and setting goals. The analysis and goals then feed into two problem spaces, both of which contain knowledge and operations. The first problem space is the content problem space. Here, the ‘knowledge states’ may be considered to be beliefs, while the operations, in this case, are ‘the inferring, the hypothesizing, and so on that lead from one state of belief to another’ (p. 146). The other problem space is the rhetorical space. The ‘knowledge states’ in the rhetorical problem space are the different ‘representations of the rhetorical situation’ whereas the operations are those which lead to changes in the text or the goals of the task, or the interaction between the two. The content problem space is therefore concerned with the ‘ideational content of discourse’ while the rhetorical problem space focuses on achieving the goals of the task by considering how the reader may react to the ideational content (p. 146). The rhetorical content space translates problems encountered
there into sub-goals for the text which must be achieved in the content problem space. Such problems might be the need to explain something more clearly, give examples or justify reasons for a particular point. This may mean that the writer needs to change or amend not only the text but also their ideas. It is this dialogue between the rhetorical problem space and the content problem space which leads to the transformation of knowledge, and this occurs when producing the content, that is, in the content problem space (p. 147).

It should be noted that the knowledge transformation model does not simply disregard the knowledge telling model; instead, the latter becomes part of the former in the guise of sub-processes on the way to achieving the goal.

At first glance, the two models may seem counter-intuitive: one might think that it is the skilled writer who does not need to plan as they are more experienced while the novice writer would be the one who, due to a lack of experience, may need to reconsider and change their ideas as they write.

Indeed, it may be the case that the experienced writer who has done a similar task previously does, in fact, go through the processes associated with the knowledge telling model because they have previously been through the knowledge transformation processes when completing such a task previously. Alternatively, they may be able to achieve the task satisfactorily through ‘sophisticated use’ of the knowledge transforming model (Grabe and Kaplan, 1996, p. 124).

On the other hand, even with the knowledge transformation model, the act of retrieving information initially may occur in the same way as the knowledge telling model (Scardamalia and Bereiter, 1987, p. 147) and even some experienced writers do indeed produce a first draft in the same way as a novice writer might. However, in contrast to the novice writer, they will then revise the text extensively (p. 149). Thus, as Hyland (2016, p. 17) notes:
[a] knowledge telling model addresses the fact that novice writers plan less often than experts, revise less often and less extensively, and are primarily concerned with generating content from internal resources. Their main goal is simply to tell what they can remember based on the assignment, the topic, or the genre (p. 17) whereas skilled writers go beyond what novice writers do by analysing and reflecting on the task in hand and setting goals to achieve this, then systematically changing the text as necessary to ensure they meet the goals set (p. 17). The Scardamalia and Bereiter (1987) models illustrate how a given writing task can vary in difficulty among equally experienced writers. If one person has experience of doing the specific task or has background knowledge which the other writers do not possess, that writer will have less information to process in the problem spaces compared to the other writers so the task will be somewhat less complex for them than for the others (Grabe and Kaplan, 1996, p. 125).

As Hyland (2016, p. 17) points out, the knowledge telling and knowledge transforming models can help us to understand why inexperienced writers may struggle with complex writing tasks so learners should be encouraged to undertake tasks of varying degrees of difficulty and involving different genres to help develop their skills. On the other hand, the models demonstrate how experienced writers reflect more on their writing and consider the audience in their writing. He argues, however, that the models do not provide any insights into how writers develop from knowledge tellers to knowledge transformers.

The process approach to teaching writing has been criticised for the importance it places on writing several drafts, which may not be beneficial for examination essay writing, and its reliance on peer feedback, which may provide learners with an ‘unrealistic view of their abilities’ (Horowitz, 1986, p. 446). He also argues that the inductive focus of the approach may not be appropriate for all writers or academic tasks (p. 446). Similarly, Purves (1986, p. 39 quoted in Horowitz, 1986, p. 446) maintains that, although disciplines vary in their norms
such that each discipline is a different ‘rhetorical community’, the process approach to teaching writing would seem to suggest that one field is much the same as another.

Nonetheless, Scardamalia and Bereiter’s (1987) models, along with Field’s (2004) model, and the work of Hayes and Flower (1980), Flower and Hayes (1981) and Flower et al. (1986) have provided us with an in-depth understanding of the cognitive processes involved in writing.

To summarise, the process approach differs considerably from the product approach discussed in Section 2.5.1.1 above in that the process approach looks at the text from the perspective of the writer while the product approach is concerned with the text as an object (Hyland, 2016). Research into writing as a process has informed us that both L1 and L2 process writing models see writing as a non-sequential series of cognitive processes (Grabe and Kaplan, 1996; Hayes and Flower, 1980; Flower and Hayes, 1981; Scardamalia and Bereiter, 1987) that focus on the writer and how the text comes into being (Walker and White, 2013, p. 61). We have also learnt that L1 writers plan more than L2 writers (Hyland, 2002, p. 26 cited in Shaw and Weir, 2007, p. 43), who, in turn, plan and revise more than inexperienced L2 writers (Hyland, 2016; Scardamalia and Bereiter, 1987). While experienced writers revise on a macro level and revise globally, novice writers tend to focus on the micro level, revising at the word or sentence level (Hyland, 2016).

2.5.1.3 Writing as social interaction

As we have seen in Sections 2.5.1.1 and 2.5.1.2 above, the product approach to writing considers the text from the perspective of a product whereas the process approach considers writing mainly from the perspective of the writer. Although both approaches provide us with useful insights into how a text comes into being, neither approach focusses specifically on the perspective of the reader.

This third approach to writing considers writing as a ‘profoundly social act’ (Sperling, 1996, p. 55) in which the text, the writer and the reader interact. Writing 'like language in general, [is]
a meaning-making activity that is socially and culturally shaped and socially purposeful’ (p. 55). In a later work, Hayes (1996, p. 5, quoted in Weigle, 2002, p. 19) considers writing to be ‘a social artifact and is carried out in a social setting. What we write, how we write, and who we write to is shaped by social convention and by our history of social interaction’.

With regard to academic writing specifically, Coffin et al. (2003, p. 10) argue that this can be considered social in three ways:

First, student writing is always embedded within relationships around teaching and learning and these relationships influence, not least, the extent to which students come to write successfully in higher education. Second, the conventions governing exactly what constitutes ‘appropriate academic writing’ are social to the extent that these have developed within specific academic and disciplinary communities over time. Third, student academic writing is a social practice in that the writers, students, are learning not only to communicate in particular ways, but are learning how to ‘be’ particular kinds of people: that is: to write ‘as academics’, ‘as geographers’, ‘as social scientists’. Thus, academic writing is about personal and social identity (Coffin et al., 2003, p. 10).

Kress (1989, cited in Grabe and Kaplan, 1996, p. 136) posits that ‘all language use is a matter of making discourse’. Grabe and Kaplan (1996, p. 136) also stress the importance of language in making meaning. They argue that language and content need to be integrated appropriately by the writer to create meaning and thus to create discourse. As these discourses become integrated and established in the ‘social functioning of groups’, they begin to be ‘recognized as genres which serve functional purposes in communication’ (Kress, 1989, cited in Grabe and Kaplan, 1996, p. 136).

Hamp-Lyons and Kroll (1987, p. 8) emphasise the importance of the context in this approach, stating that writing is ‘an act that takes place within a context, that accomplishes a particular purpose, and that is appropriately shaped for its intended audience’.

44
Similarly, Hyland (2016, p. 21) argues that this approach also takes account of what the reader expects to see in a given text, taking ‘the notion of context (original emphasis) beyond features of the composing situation to the purposes, goals and uses that the completed text may eventually fulfil’ (p. 21). As such, the writer must make assumptions about what the reader already knows whereas the reader predicts the text based on what they assume the author’s purpose to be and ‘each presupposes the sense-making capabilities of the other’ (Nystrand, 1989, p. 73).

This approach to writing therefore differs from the previous two approaches in that it is not one-directional (as in the writer producing the text). Instead, the text is situated in between the writer and reader and mediates each one’s purpose (p. 74). Thus, text meaning is conceptualised not simply through the writer alone but ‘in terms of the interaction between writer and reader purpose’ and ‘in terms of its semantic potential’ (p. 75).

Nystrand (1989, p. 77-78) suggests three essential acts that the skilled writer performs. These are: ‘initiating written discourse’, ‘sustaining written discourse’ and using the option to elaborate on the text. In the first case, the skilled writer establishes ‘a mutual frame of reference’ between themselves and the reader by clarifying the genre, topic and tone of the text. In the second of these three acts, the writer introduces new information and then tests this new information for reciprocity. The skilled writer is able to sense when the reader will understand the text or when explanation, exemplification or elaboration will be needed and can then provide this as necessary through constantly monitoring how the text may be received by the reader. These elaborations constitute the third act, whereby this is further divided into three possible types of elaborations: ‘genre elaborations’, which are used to explain the type of text; ‘topical elaborations’, which give further information about the topic of the discourse and ‘local elaborations’, which provide additional clarification for the content of the text (p. 78).
In this model, therefore, the writer always keeps the audience in mind. However, the concept of ‘audience’ is not straightforward. In order to understand the requirements of one’s audience, the writer must have an understanding of the ‘appropriate genre, content, stance and style’ required of the text (Hyland, 2016, p. 23).

To summarise, the social-interactive approach focuses on the reader-writer relationship, which is dependent on the writer having not only the linguistic resources to achieve the task but also an awareness of the audience, the context and the specific genre necessary in a given discourse community. The writer consistently pays attention to the context in which the text will be read and monitors the developing text, revising it as necessary, to meet the needs of the reader.

2.5.2 Conclusion
Bringing together the insights from the three approaches to researching and teaching writing, Tribble (1996, p. 43) argues that a writer needs four types of knowledge to produce a text. These are: the language system, content, writing process knowledge, and context. The language system knowledge refers to the knowledge of the linguistic systems necessary to achieve the task. The product approach (see Section 2.5.1.1) exemplifies the importance of this aspect. Content knowledge is knowledge of the subject or topic area while writing process knowledge is the understanding of the processes involved in preparing for and completing the task. We saw the need for both of these in the section on ‘Writing as a process’ in Section 2.5.1.2 above. Finally, context knowledge takes account of ‘the context in which the text will be read’ (p. 43) and the important relationship between the reader and the writer, which was discussed in Section 2.5.1.3 above.

Clearly, then, a writer who does not possess all four types of knowledge may struggle to write whereas a writer who does possess all four types of knowledge will likely be, or become, a skilled writer. Consequently, in teaching and testing academic writing, we will want to ensure that learners have all of this knowledge.
In a language testing environment, however, one issue which may arise with writing tasks is that test takers do not have the content knowledge required to complete the task (see Section 2.6.5 below). If this is the case, it is also unlikely that they will have the language. The lack of content could therefore have an adverse effect on the writer and their performance in the test. Given that we want to ‘test for best’ (Weir, 2005, p. 54), this should be borne in mind when considering whether to give independent (or discrete) writing tasks, which require the writer to produce the content from their background or topic knowledge, or integrated writing tasks, in which the content (and, to some extent, language) is provided through the input. The considerations around independent or integrated test tasks will be discussed in the next Section.

2.6 Independent tasks and real-world academic writing tasks

2.6.1 Background

Traditional language teaching and testing has centred around the four discrete skills of listening, speaking, reading and writing. Some large-scale, international English tests, such as the IELTS test or Cambridge B2 First, test learners according to discrete skills. The advantage of testing the skills separately is that test takers can get information about which skills they performed well on and which one(s) they may need to work on. In addition, visa regulations in some countries such as the UK require test takers to provide evidence of their ability in all four skills. Consequently, testing the skills discretely means that such evidence can be made readily available on the results transcript.

However, as language skills are not always employed in isolation, this may not reflect ‘the complex reality of communication’ (Council of Europe, 2018, p. 30), particularly in an academic environment, where very little academic writing is done as an independent task without reliance on other skills (Horowitz, 1986); instead, university writing usually requires students to work with other sources, such as reading and / or listening texts (Weigle, 2004). Therefore, the use of integrated test tasks ‘require examinees to write ... in ways that more authentically
resemble the types of performance needed for academic studies …’ (Cumming et al., 2005a, p. 6).

Bachman and Palmer (1996) argue that it is ‘possible’ (and I would argue necessary) ‘to identify distinguishing characteristics of language use tasks and to use these characteristics to describe a language use domain’ (p. 44) because the purpose of language testing is to ‘make inferences that generalize to those specific domains’ (p. 44). In other words, test tasks should replicate the TLU domain. Thus, without a clear understanding of the TLU domain, it is difficult to consider the construct needed to assess suitability for a given domain. Therefore, in the following sections, I first outline research looking at the requirements of real-world academic tasks and then discuss research findings related to independent writing tasks. After that, I consider the research on independent versus integrated tasks before focussing on integrated tasks.

2.6.2 ‘Real-world’ academic tasks

In the academic environment, writing ability has long been considered among the ‘fundamental facets of language ability for successful academic achievement’ (Soleimani and Mahdavipour, 2014, p. 132). In the past, academic writing ability was assessed through indirect forms of assessment such as multiple-choice items. However, direct writing assessment started to replace indirect writing assessment around 50 years ago (Hamp-Lyons, 1991) as this was seen as providing more valid evidence of the test taker’s writing ability (Camp, 1993). Examples of direct writing tests in which test takers are required to produce language are the International English Language Testing System (IELTS) test and the Test of English as a Foreign Language (TOEFL). In both of these tests, test takers are required to produce an impromptu, independent writing task in the form of an essay ‘based on personal experience or opinion’ (ETS, 2020).

Clearly independent, impromptu tasks have more face validity than indirect writing tasks, yet the validity of independent writing tasks has been questioned by some researchers (Cho, 2003;
Leki and Carson, 1997; Weigle and Friginal, 2015). As mentioned in Section 2.6.1 above, to evaluate this, it is necessary to understand what types of tasks are required in the TLU domain. Consequently, several studies have sought to investigate what is required of real-life academic tasks and found that the integration of tasks is an essential skill for the tertiary environment (Hale et al., 1996; Horowitz, 1986; Westbrook and Howell, 2011).

In a survey of 108 lecturers at two UK universities, Westbrook and Howell (2011) found that ‘listening to lectures and taking notes’ was the second most frequently stated listening task - after ‘taking part in tutorial discussions’ (91.5 per cent) - with 90.4 per cent of the lecturers surveyed stating this as one of the tasks their students need to be able to do. In the same survey, they found that writing essays, reports, projects and summaries were among the written tasks expected of university students.

Horowitz (1986) carried out a survey of academic writing tasks. The survey included 54 writing tasks from 29 courses across 17 departments of a US midwestern university. He identified seven different categories of tasks: summarising / reacting to a written input text, writing an annotated bibliography, reporting on a ‘specified participatory experience’, connecting theory and data, writing case studies, synthesising across several sources, and writing up a research project (p. 449). He concluded that, with the exception of ESL classes, and English composition, literature and creative writing, ‘the academic writer’s task is not to create personal meaning, but to find, organize, and present data according to fairly explicit instructions’ (p. 455).

Building on Horowitz’s (1986) study, Hale et al. (1996) investigated the types of writing tasks that are assigned in degree programmes at both undergraduate and postgraduate level, and across different institutions and different disciplines including Business, Chemistry, Civil Engineering, Computer Science, Psychology, Economics, English and History. The authors note that some of the disciplines selected are ones ‘in which international students specialize’ while others represented the ‘core undergraduate curriculum’ that both home and international
students take (p. 39). In comparison to other studies of university writing tasks, which have used surveys to investigate the types of tasks required (Westbrook and Howell, 2011), Hale et al.’s (1996) study involved comparing examples of actual writing tasks assigned to students. They classified the assignments according to whether they were completed in or out of class, the length of the task, the genre, the cognitive demands placed upon the student, the rhetorical nature of the task and the pattern of exposition, which relates to the form that the task takes, e.g. problem-solution, definition, analysis, etc. In terms of genre, they identified ten different genres:

1. Essay
2. Library Research Report
3. Report of Experiment / Observation without Interpretation
4. Report of Experiment / Observation with Interpretation
5. Summary (includes annotated bibliography without comment)
6. Case study
7. Plan / Proposal
8. Documented Computer Program
9. Book Review
10. Unstructured Writing (Hale et al., 1996, p. 10)

In terms of cognitive demands, they identified three different categories. These were retrieving / organising; applying / analysing / synthesising / evaluating; and ‘uncertain’. The first category here relates to lower-level cognitive demands while the second one relates to higher-level cognitive demands. They found that higher-level cognitive demands were involved in many of the writing assignments and significantly more frequently for assignments completed outside of class than in class.
Exposition was required of all tasks while narration or description were ‘involved in zero percent (rounded) of the assignments’ (p. 39). With regard to the genres identified, Library Research Papers and Reports of Experiments / Observations with Interpretation were particularly frequent as were Summaries, Plans / Proposals, Book Reviews (p. 45) and Essays (p. 47).

It should be noted that this study only had eight participating institutions so a wider-scale study may yield different results. Similarly, the authors are keen to stress that the results obtained relate to the specific courses in the study and should therefore not be considered generally representative of the disciplines involved. Nonetheless, this study highlights the fact that the tasks involve require higher-order thinking skills and integrate output tasks with some sort of input.

The findings from the research outlined here illustrate the integrated nature of university assignments. However, many language tests still employ independent writing tasks. Therefore, it is important to understand how the constructs tested in independent writing tasks compare to those of real world tasks. This will be discussed in the next section.

2.6.3 Independent writing tasks and ‘real-world’ performance
As shown in the previous section, many real world tasks require the integration of different skills; however, direct tests of writing often involve impromptu essays which do not require test takers to refer to outside sources. As such, there is a need to understand the differences between such independent writing tasks and authentic university assignments.

Moore and Morton (2005) compared the IELTS Task 2 with assignment tasks from two Australian universities. Like Hale et al. (1996), they also classified the assignments according to different categories: ‘genre’, ‘information source’, ‘rhetorical function’ and ‘object of enquiry’ (p. 47). Within each of these categories, they identified further sub-categories into which they allocated the different tasks. They found that around 60 per cent of the tasks were based around the loosely defined genre of an essay. While the IELTS task does indeed share some
similarities with the essay genre in a university environment, they argue that the former should not be considered ‘an appropriate model for university writing’ (p. 43) since there are important differences between the IELTS task and a university essay. In terms of language functions, they found that the two most commonly elicited functions in real-world academic tasks, namely, description and summarization, were not usually elicited by the IELTS task. Other differences include the fact that assignments that first year university students are required to complete rarely involve spontaneous writing, and opinions, where permitted, should not be supported by evidence which is anecdotal or pertaining to personal experience; instead evidence provided should be based on research findings. They also found that university writing was based more on theory and analysis than ‘real world’ situations. However, they consider the most important difference between the university assignments and the IELTS Task 2 essay to be the fact that the former is ‘intimately related to processes of reading’ (p. 64) whereas the latter appears more closely related to ‘public non-academic genres [such] as the letter to the editor or newspaper editorial’ (original emphasis) (p. 64).

However, the IELTS Task 2 may not be as far from real world academic writing as Moore and Morton (2005) might lead us to believe. MacDonald (1994, p. 187 quoted in Sutton, 1997, p. 48) identifies four types of academic writing that writers produce on the way from being novices to becoming ‘disciplinary insiders’. These are:

1. Non-academic writing
2. Generalized academic writing concerned with stating claims, offering evidence, respecting others’ opinions, and learning how to write with authority
3. Novice approximations of particular disciplinary ways of making knowledge

In terms of MacDonald’s (1984, p. 187 quoted in Sutton, 1997, p. 48) continuum above, this would place the IELTS Writing Task 2 at stage one on the continuum ‘non-academic writing’.
This suggests that such writing is indeed a step on the way to becoming an academic writer and not completely disconnected from it. Indeed, Moore and Morton (2005) concede that the IELTS Task can provide an opportunity to help students develop their literacy skills.

Turner et al. (2009) compared the IELTS test with the English for Academic Study (EAS) programme at their university and found similarities between the IELTS writing tasks and the EAS writing tasks with regard to the rating criteria used and some of the functions elicited such as giving reasons, providing solutions and making comparisons. However, there were also several differences, most notably that the EAS programme included a 300-word essay in which students were expected to use information referenced from source texts, and a 2000-word group research report. Additional criteria that the students were assessed on included use of register, use of citation and the ability to summarise and paraphrase as well as quoting directly and indirectly. Like the Moore and Morton (2005) study, which was based on responses from only two universities, this study was only done on a small-scale, comparing the IELTS test to the programme at their own university. As such, it should be borne in mind that these studies may not be entirely representative of higher education institutions more widely.

Priyanti (2017) evaluated the IELTS writing task with reference to Bachman and Palmer’s (1996) qualities of test usefulness. She concluded that there was more of a focus on reliability and practicality while lacking construct validity as a result of construct underrepresentation (Messick, 1989), which may impact on the test’s predictive validity. Finally, she argues that the test lacks authenticity since the tasks correspond only in small part to real world tasks.

While these studies may highlight the shortcomings of the IELTS Writing Task Two compared to real world university tasks, they do not compare performance on IELTS writing tasks to performance on real-world tasks. Consequently, although Moore and Morton (2005) argue that such tasks should not be a model for university writing, it is not actually clear to what extent performance on these tasks would reflect success or otherwise in university writing assignments. Similarly, Turner et al. (2009) did not include any investigation of the academic
success or otherwise of students who had completed the EAS programme compared to using an IELTS test only for entry purposes even though the university in question accepts both types of students. This is unfortunate, given that the data were most likely available and would have provided useful insights into our understanding of the predictive validity of the IELTS Writing Test.

In a study comparing successful performance on university writing with high-scoring performances on independent tasks in the TOEFL iBT test, Weigle and Friginal (2015) found that the TOEFL essays elicited different linguistic patterns to those produced in successful academic tasks. While the TOEFL essays did exhibit some of the linguistic features of writing in humanities subjects, the test essays were found to be more narrative and tended to feature ‘expressions of opinion and stance’ (p. 34) whereas the academic writing tasks, particularly those from natural science subjects tended to be associated more with features of informational and procedural writing, thereby demonstrating that university disciplinary writing differs significantly from the independent test essays and supporting the claim that the impromptu test essay ‘is limited in terms of its generalizability to other academic genres’ (p. 37).

In contrast to the other studies outlined above, which compared independent writing tasks to university writing tasks from a limited number of universities, Weigle and Friginal’s (2015) study used 125 high-scoring, timed, impromptu essay test papers from the TOEFL iBT corpus and native speakers (by way of comparison with L2 learners), as well as 125 discipline-based writing papers from each of the four academic fields in the Michigan Corpus of Upper-level Student Papers (MICUSP) corpus, which includes 829 student papers from across 16 disciplines in four academic fields.

Weigle and Friginal (2015, p. 37) state that their findings correspond to MacDonald’s (1994, p. 187 quoted in Sutton, 1997, p. 48) continuum as the papers from the MICUSP corpus were produced by ‘advanced undergraduate and graduate students’ whereas the TOEFL papers
were from first-year undergraduate native-speakers and ‘international students, many of whom had met minimum proficiency standards for university entrance but were still relatively unfamiliar with academic discourse’ (Weigle and Friginal, 2015, p. 37).

In contrast with Moore and Morton (2005), whose results reflect Stage 1 on the continuum proposed by MacDonald (1994, p. 187 quoted in Sutton, 1997, p. 48), Weigle and Friginal (2015) maintain that the argumentative essay ‘corresponds to Stage 2’ (p. 37) of the continuum. As such, they suggest that argumentative essay tasks may indeed be useful for assessing whether students possess the basics of academic writing, which they need in order to progress to more academic, discipline-based writing. While such tasks may indeed provide a stepping stone to academic writing, such direct writing tests which include independent, impromptu essays, do not appear to correspond to the requirements of the more integrated skills-based university assignments outlined in Section 2.6.2 above.

2.6.4 Independent writing tasks and authenticity
The studies outlined above highlight the fact that there appears to be something of a mismatch between the constructs tested in independent writing tasks and those of real-world academic writing tasks. Consequently, a number of researchers have criticised independent tasks for their lack of authenticity (Cho, 2003; Hamp-Lyons and Kroll, 1987; Read, 1990; Weigle, 2002) and for not ‘adequately tapping into the academic writing construct’ (Gebril, 2009, p. 508). Indeed, if a test purports to test academic writing ability, it should require test takers to demonstrate their ability to engage in L2 writing processes such as ‘planning, monitoring, and revising’ (Cohen and Upton, 2007, cited in Barkaoui, 2015, p. 1).

Weigle (2002) argues that the impromptu essay represents a ‘compromise’ between teachers and psychometricians because the former ‘see writing as a complex, multi-faceted process and assessment as something which must be closely integrated with instruction’ whereas the latter are concerned with validity and reliability, which can be more easily measured ‘as the sum of discrete, … component parts’ (p. 240). She states that there are ‘at least four ways’ in
which the 'timed impromptu essay' is inauthentic. One argument is that university assignments are usually graded on the basis of their content rather than language; another is that the raters are usually known to the test taker in the academic environment, which means that they may better understand the requirements of the task. The third point she makes is that real-world academic writing is generally not speeded so test takers can write and revise at their own pace and finally, impromptu writing tasks are inauthentic due to the lack of background knowledge, which in the real-world is gained through interaction with reading, listening and/or speaking (p. 52).

Consequently, integrated tasks have become increasingly popular in recent decades (Gebril and Plakans, 2009, p. 48). Cumming et al. (2000, p. 27) suggest that authenticity is perhaps the most important justification for including integrated tasks in tests which assess students’ writing ability in academic environments as they expand the construct of writing, ‘aligning it appropriately with the kinds of writing that are integral to academic settings’ and will ‘assess examinees’ writing performance on a range of rhetorical functions integral to academic contexts (e.g. summarizing, reporting, categorizing, analysing)’ (p. 27).

2.6.5 Independent writing tasks and background knowledge

Weigle (2002) is not the only researcher to criticise independent writing tasks for failing to involve background knowledge. Other researchers have criticised such tasks for assuming that all test takers have the same background knowledge about a topic. As this is unlikely to be the case, a lack of topic knowledge may have an impact on scores, thus leading to construct-irrelevant variance (Gebril, 2009; Guo, 2011; Weigle, 2004).

Investigating the impact of background knowledge on writing performance, Lewkowicz (1994) compared the performance on an academic writing task between one group who were provided with background reading texts and one group who had no background reading. She found that, although the background reading provided ideas for the writing, they did not always lead to improved quality whereas those students who had no background reading may
have had fewer ideas to write about but tended to expand on these ideas more fully. She therefore argues that independent tasks should be practised and tested due to the likelihood that students may encounter them outside of the academic environment. With regard to the use of background reading texts, she found that the better students distinguished themselves from the weaker ones by demonstrating a better ability to summarise the input and paraphrase it. She concludes that, while providing background reading texts may not improve students’ writing skills, such tasks contribute to fairness among test takers because they ‘ensure that, in terms of subject knowledge all start equally, at least in terms of the information available to them’ (Weir, 1983, p. 387 quoted in Lewkowicz, 1994, p. 204) and may help them with academic writing tasks since they reflect real-life university tasks.

Read (1990) also argues that test takers are likely to perform better when presented with a topic that they are familiar with rather than an unfamiliar one. Similarly, he argues that test takers should be given a topic which they ‘have enough relevant information on, or opinions of, to be able to write to the best of their ability’ (p. 110).

Weigle (2002, p. 52) argues that when students are required to write an impromptu essay on a general topic, this may be problematic as test takers may be lacking background knowledge and will therefore be unable to write confidently about the topic (p. 92). Conversely, if they have dealt with a topic through reading or listening input, they should, in theory, be ‘well equipped (sic) in terms of background knowledge and schemata to write about it’ (p. 52).

This supposition is supported by Tedick (1990), who compared performance on a general essay topic compared to an essay on a topic related to the test takers’ degree subject. She found that there was a highly significant difference between the scores with a ‘marked increase’ in the scores for the subject-specific essays (p. 132). The results also revealed a ‘significant main effect of topic as measured by overall length’ (p. 134) as lower and intermediate level learners produced longer texts for the subject-specific essay, although the advanced level learners’ texts were almost identical in length across both essays. She
concludes that this means that the advantage of subject knowledge was more beneficial to lower and intermediate level learners than to the more advanced ones.

These findings concur with other research about topic knowledge and text length carried out by Winfield and Barnes-Felfeli (1982) in which two groups of test takers (one group of Spanish speakers from Latin America and one group of non-Spanish speakers from different countries) were asked to reproduce the content of two short texts - one about Don Quixote and one about Japanese ‘Noh’ theatre – which they read silently as the researcher read the texts aloud to the participants. The input was presented in a counter-balanced measures design. Their findings show that the text length and grammaticality produced by the Spanish group were higher for the Don Quixote text than the mixed background group but lower than the mixed background group for the ‘Noh’ theatre text. The mixed background group maintained similar levels across both texts. This suggests that topic familiarity had an influence on both fluency of writing and grammatical accuracy.

Further, the authors found that ‘intrusions’ or additional information produced by the Spanish-speaking group, which was not in the original text was generally correct for the Don Quixote text but for the ‘Noh’ theatre text, such ‘intrusions’ represented ‘cultural interpretations or cultural miscues of the information’ (p. 377). They conclude that writing topics should be based on “known contexts’ or personal experiences’ (p. 377). They also highlight the need for students to be familiarised with the context, language and the cultural values of the target culture before they are asked to write about the topic.

On the other hand, Wolcott (1998, p. 27-28) argues against using a personal topic as it may evoke strong emotions and cause both the writer and examiner discomfort if the experience the test taker is writing about is not a happy one. Furthermore, such texts are likely to be narrative in nature, which may not adequately represent the range of genres required in higher education. Consequently, a wider range of genres should be tapped into and the type of
writing that students produce should be reflective of what they would need in the classroom (Carlson and Bridgeman, 1986, p. 141 quoted in Wolcott, 1998, p. 28).

The potential threat to construct validity brought about by a lack of background knowledge and the mismatch between independent writing test tasks and ‘real-world’ university writing tasks must be addressed if test takers are to be given the opportunity to demonstrate the best of their ability and be prepared for the academic environment. It seems reasonable then to avoid personal subjects and to heed the advice of researchers who suggest that input should be provided for reasons of fairness (Weir, 1983 quoted in Lewkowicz, 1994) and to reflect the task requirements of the TLU domain (Carlson and Bridgeman, 1986, p. 141 quoted in Wolcott, 1998, p. 28). As such, there appear to be good reasons to use integrated tasks in the EAP domain. The next section will therefore examine the research around integrated writing tasks.

### 2.7 Integrated writing tasks

#### 2.7.1 Defining integrated test tasks

McDonough and Shaw (2003, cited in Gholami and Alinasab, 2017, p. 127) consider integrated skills as ‘conjunctive skills to be practiced in the classrooms in order to make teaching and learning truly communicative and authentic’ (p. 127). On the other hand, Hartley (2007, p. 317) states that, in the field of academic writing, he ‘cannot conceive of writing … without considering reading, speaking and listening’ because the four skills are interrelated.

Cumming (2014, p. 224) states that integrated skills assessment is based on an ‘interactionist theory of human communication in which knowledge is constructed through the interpretation and expression of relevant ideas through multiple media’. Not only is it imperative to possess such abilities in order to function effectively in an academic environment but these abilities should also serve as ‘guiding principles’ when designing language assessments for these contexts.

However, in operationalising these principles, there are a number of different interpretations of what constitutes an integrated task with several researchers providing definitions for
integrated tasks (Ascención Delaney, 2008; Cumming et al., 2005a; Knoch and Sitajalabhorn, 2013; Plakans, 2009a; 2012) which range from quite simple definitions such as tasks which combine one or more skills (Plakans, 2012, p. 249) to more detailed definitions such as the one found in Cumming et al. (2005a, p. 34) which posits that integrated tasks ‘require complex cognitive, literate, and language abilities ... to produce written compositions that display appropriate and meaningful uses of and orientations to source evidence’.

Responding to calls for a ‘coherent manifesto for the design of integrated writing tasks’ (Cumming, 2013, p. 4), Knoch and Sitajalabhorn (2013) explore the integrated writing construct and review a range of task types to extract the features that they consider essential for a task to be integrated. On the basis of this review, they propose a more comprehensive, 6-part definition for integrated writing tasks, as follows:

... test takers are presented with one or more language-rich source texts and are required to produce written compositions that require (1) mining the source texts for ideas, (2) selecting ideas, (3) synthesising ideas from one or more source texts, (4) transforming the language used in the input, (5) organizing ideas and (6) using stylistic conventions such as connecting ideas and acknowledging sources. (Knoch and Sitajalabhorn, 2013, p. 306)

The definition above is in line with Spivey’s (1984, cited in Spivey and King, 1989, p. 9) explanation of discourse synthesis whereby writers select ideas from the input texts, then organise the ideas and connect the ideas from the input with their own, transforming the content to produce their own texts, which, in turn, corresponds to Scardamalia and Bereiter’s (1987) knowledge transforming model outlined in Section 2.5.1.2 above.

2.7.2 Integrated writing task types
Various tasks are used in integrated tests. Many integrated tasks comprise reading and writing (Campbell, 1990; Keck, 2006; Plakans and Gebril, 2012; Shi, 2004; Weigle and Parker, 2012) but some employ visuals, such as a graph, as the basis for writing (Yang, 2012). Others integrate both reading and listening into writing (Cumming et al., 2005a; 2005b; Plakans and
Gebril, 2013; Sawaki, Quinlan and Lee, 2013; Zhu et al., 2016) while others include listening-into-writing tasks (Carrell, 2007; Cubilo and Winke, 2013; Rukhthong and Brunfaut, 2020).

Some integrated tasks require test takers to make use of only one input text while others employ several. The input can take the form of a visual, as in the IELTS Task 1; a reading text, for example, from a course textbook; a listening text, such as a university lecture in the EAP environment; or a combination of these. Knoch and Sitajalabhorn (2013, p. 301) point out that reading-into-writing tasks, which require test takers to write a text based on the content of the reading input are the most common integrated tasks. One reason why such tasks are common could be, as Weir (2014, p. 6) points out, the research literature concludes that ‘a knowledge-transforming, integrated reading-into-writing task can address construct validity concerns better than the more common, independent writing-only, knowledge-telling task type’.

The types of writing tasks used most commonly include summary tasks (Baba, 2009; Li, 2014; Wolfersberger, 2007; Yu, 2007, 2008, 2009, 2010), which can vary in length from a one-sentence summary of a text as in the Pearson Test of English Academic to much longer, more complex open-ended tasks marked according to holistic or analytical criteria, as in the TOEFL iBT (Cumming, 2014, p. 224), or essay response tasks (Ascención Delaney, 2008), in which test takers are required to summarise the input and provide a response. Alternatively, if several input texts are provided, test takers may be asked to provide a synthesis of the different input texts or to discuss opposing views of a problem presented in the input texts (Knoch and Sitajalabhorn, 2013, p. 301-302). As mentioned above, in some cases, most notably the TOEFL iBT, the writing task involves both listening and reading input. Here, test takers summarise the lecture content and relate this to the reading text (p. 302).

Knoch and Sitajalabhorn (2013, p. 302) note that integrated tasks ‘require learners to transform the language of the input material to accomplish the writing task’; therefore, the input material is ‘crucial’ to the task.
As Weigle (2002, p. 94) states, ‘[t]here are valid reasons for either providing or not providing stimulus material, and ultimately the choice depends on the definition of the construct ...’ Similarly, the choice of writing task selected for the output will also depend on the construct. The next section will look at the research into summary writing tasks as this is the task that was used for the output in both the exploratory study and the main study in this thesis.

2.7.2.1 Summary writing

The popularity of the summary writing task decreased in the 1970s but has seen a resurgence in recent years due to the need to ‘mirror the real-life demands that are made on students in an academic context in order to enhance test validity’ (Weir, 2014, p. 6). Reading-into-writing summary tasks have been found to demonstrate both context and cognitive validity (Shaw and Weir, 2007).

In arguing the case for introducing a summary writing task as part of the IELTS test, O’Kane (2017, p. 52) also warns about some of the potential issues around summary writing. He argues that such a task can lead to difficulties when assessing reading and writing as it is not clear where the border between the two skills lies, nor indeed, whether it should be considered a reading test or a writing test as there are arguments on both sides. He also highlights the potential issues of cultural bias and comparability across test versions. While these points need to be considered for an integrated task, it could be argued that this is the case when selecting any reading or listening input so these issues are not limited to a summary writing task.

Another point that O’Kane (2017, p. 52) raises is the considerable cognitive load that summary writing can impose. Both internal and external factors can impact on cognitive load (Hidi and Anderson, 1986; Kirkland and Saunders, 1991; Yu, 2009). These include the length of the text to be summarized, whether students have access to the text while writing the summary, and the complexity of the text to be summarised (Hidi and Anderson, p. 473). Summarising longer texts is more difficult due to the need to make more decisions about what to include or exclude from the summary whereas summarising a shorter text is likely to be easier as learners can
select one or two topic sentences to convey the meaning of the text (p. 475). However, Kirkland and Saunders (1991, p. 106) found that for L2 students, summarising shorter texts was more difficult as they were constrained by the language and structure of the shorter texts.

Access to the source text has also been found to affect cognitive load. Hidi (1984b, cited in Hidi and Anderson, 1986) suggested that different operations are executed while summarising depending on whether the source text is available during writing. If available, students may refer back to check details but this may also lead to increased source text borrowing. However, despite the increased cognitive load of not having the text available while writing, students with no access were found to be better at recalling the content one week later than those who did (Hidi 1984b; 1985, both cited in Hidi and Anderson, 1986). Although this research refers to reading texts, the same may (or may not) be true of listening. Students who have access to the recording may play parts several times and note down more words from the text than if there is no access after the initial playback(s).

Text complexity is also an important factor in summary writing. The readability should be at an appropriate level and a clear structure will aid comprehension (DeCarrico and Nattinger, 1988; Yu, 2009; Hidi and Anderson, 1986).

Another factor which may affect summary writing is L2 proficiency. An adequate level of comprehension, reading skills, grammar and vocabulary are ‘fundamental to successful summarizing’ irrespective of the skills and strategies that a learner may possess in their L1 (Kirkland and Saunders, 1991, p. 108). Yu (2009) investigated the impact of different types of input text on test takers’ ability to summarise the input. His participants summarised three texts, each of which varied in terms of discourse features and organisation. In contrast to Kirkland and Saunders (1991), his findings revealed that the source text had a bigger influence on performance than did L2 proficiency. He also found that factors such as macro-organisation, topic knowledge and the number of unfamiliar words in the text also impacted on participants’ ability to summarise the input.
Summary writing is a complex, high-level cognitive skill which learners acquire over time (Kirkland and Saunders, 1991; O’Kane, 2017; Weir, 2014) so, while several factors may affect the cognitive load, summary writing is, nonetheless, ‘essential in an academic setting’ (Kirkland and Saunders, 1991, p. 105). Students are often required to summarise input, either as a learning aid or as a way of incorporating information from other sources into assignments (Kirkland and Saunders, 1991; O’Kane, 2017) so ‘[o]ne of the skills which learners need for academic writing in a second language is knowing when and how to summarize in their own words’ (Walker and White, 2013, p. 68).

Indeed, Tarone and Yule (1989, cited in O’Kane, 2017) suggest that students often lack the ability to evaluate, select and synthesise so O’Kane (2017) suggests that summary writing might be ‘one means of redressing this deficiency and of identifying candidates whose skills are not yet at the requisite level for the rigours of academic discourse’ (p. 52). This is a strong argument in favour of such integrated tasks in university entrance tests as students may have good reading and / or listening skills, and good writing skills. However, despite having good levels of proficiency in the discrete skills, mediating between the input and the output may prove challenging so this is often where students struggle and fall into the plagiarism trap.

Summary writing consists of reading (or listening to) and understanding a text, establishing and selecting the main ideas, putting these into note form then reconstructing the notes to make a coherent text using the student’s own words (León et al., 2006; Weir, 2014; Wolfersberger, 2007). Consequently, poor summary writing is often the result of poor content selection (Hare and Borchardt, 1984, cited in Wolfersberger, 2007, p. 49) and an inability to combine information from the source text(s) into one single, abridged version of the text (Hare, 1992, cited in Wolfersberger, 2007, p. 49). In contrast, according to León et al. (2006, p. 617), two fundamental features of a good summary are ‘[s]ynthesis and coherence’. Therefore, it is ‘more complicated’ than simply reading a text (p. 617). In fact, summarising
the main ideas of a text represents ‘one of the more demanding levels of processing activity’ in which university students have to engage (Weir, 2014, p. 6).

Ascención Delaney (2008) and Brown (2004) both distinguish between a summary and an essay response task. Brown (2004, p. 214) points out that the former requires ‘a synopsis or overview of the text’ whereas the latter requires the student to ‘provide [their] opinion on the text as a whole or on some statement or issue within it’. This highlights the fact that the response task goes beyond mere comprehension of the text as required of a summary, which, in turn, means an added cognitive load.

Ascención Delaney (2008, p. 147) looked into the reading-into-writing construct as measured by a summary and an essay response task. Her findings suggest that the tasks were, in fact, two different aspects of the reading-into-writing construct, leading her to conclude that a student who can write a summary may not be able to perform other reading-into-writing tasks such as response essays. In fact, by comparison with the summary, the response essay proved more challenging for lower-proficiency learners. She points out that the essay also ‘seemed to involve more critical thinking in constructing the task representation’ (p. 147) so those with less experience in performing such tasks and / or those with low language proficiency may be affected by this in terms of how they approach the task.

Authenticity, therefore, is a key feature of summary writing tasks. They have been found ‘a) to be particularly germane to university-bound ESL/EFL students, b) to be more realistic than essay writing in their uses of English for academic purposes, and c) to be in accordance with recent research and theory demonstrating the interrelations of writing with reading and aural skills in academic settings’ (Cumming, Kantor and Powers, 2001, p. 65).

Other research into summary writing has looked at the impact of lexical proficiency on summary writing (Baba, 2009; Leki and Carson, 1994), lexical diversity (Yu, 2013a), presentation mode (online vs paper-based) (Yu, 2010), writing processes of L1 and L2 learners
Source text use has received a good deal of attention in integrated tasks so this will be dealt with in more detail below (see Section 2.8.3).

There are strong arguments in favour of summary-writing as tasks which reflect the TLU domain but evaluating a summary can be challenging. Imao (2001, p. 184 quoted in Brown, 2004, p. 214) outlines four criteria for assessing a summary. A summary:

1. Expresses accurately the main ideas and supporting ideas.
2. Is written in the student’s own words; occasional vocabulary from the original text is acceptable.
3. Is logically organized.

Brown (2004, p. 214-215) highlights the fact that only the first of these relates to reading (or, in the case of a listening-into-writing summary task, listening) while the other three criteria relate to writing. It is therefore crucial that reading / listening and writing are not considered separate constructs; instead, there must be an interaction between the two (or three). If the first criterion is ignored, the text produced may not demonstrate any understanding of the input text while the other three criteria are necessary to evaluate the extent to which the task has been completed accurately and successfully, that is, the extent to which the test taker has produced a text which accurately reflects the content of the original.

Brown (2004, p. 215 proposes a simple holistic scale for summarising and responding to reading. However, as this was based around reading and only covered comprehension of the input text, I decided this was not an appropriate scale for this study. While it would have been desirable to design my own specific task rating scale, this would have required more time and resources than could be managed within this study. Instead, I opted for a practical solution.
which involved using readily available scales but, as part of the rater training, raters were asked to bear in mind the extent to which the output reflected the propositional content of the input. Analyses of source text use (see Sections 3.4.2.2 and 5.4) provided evidence of the extent to which the test takers had understood and reproduced the input successfully.

2.7.3 Criticisms of integrated tasks

Even though, ‘writing ... is not (at least not in its academic uses) a stand-alone skill but part of the whole process of task response and creation’ (Hamp-Lyons and Kroll, 1987, p. 19), several researchers have criticised integrated tasks for their multidimensional nature (Charge and Taylor, 1997; Cumming et al., 2005b; Sawaki, Quinlan and Lee, 2013). Writing about the tasks developed for the change from the ELTS exam, the predecessor to IELTS, to the IELTS exam in the mid-1990s, Charge and Taylor (1997, p. 375-376) justified the removal of integrated tasks from the ELTS exam, arguing that the thematic link between the reading and writing tasks was causing issues of construct validation because the risk of penalizing candidates with lower reading abilities for a poor performance in the writing test was unfair. They also stated that the differing approaches to the use of the source texts by candidates was an issue since some candidates relied heavily on the source text while others barely referred to it. Still others appeared confused as to whether they should express their own opinions or refer to the source texts. However, these supposed defects reflect precisely the issues that many students have when dealing with source texts in real-life academic environments and thus, this change away from integrated tasks reflects a move away from testing the construct which is operationalised in the TLU domain. It is therefore somewhat disappointing that this change was implemented precisely at a time when other major testing organisations had recognised the importance of integrated tasks and were designing such tasks for introduction into their tests. It could be argued that implementing this change fuelled the criticisms around the lack of construct validity that we have seen directed towards independent writing tasks in recent years.
Nonetheless, integrating information from other sources is indeed challenging due to the need to employ different modalities. Thus, one challenge of integrated testing is that lower level learners may have difficulty understanding the input, selecting the key information required and / or organising the output (Sawaki, Quinlan and Lee, 2013, p. 93).

Likewise, when investigating the concurrent validity of the integrated speaking and integrated writing tasks for the TOEFL iBT, Cumming et al. (2005b) found that the most frequently raised concern was that test takers would be disadvantaged if they did not understand the input texts (p. 29). However, the teachers interviewed in Cumming et al.’s study suggested that this issue could be mitigated by providing texts on topics that test takers would be both familiar with and interested in. Perhaps more importantly, they recommended that examinees be made familiar with the requirements of the task and practise this before taking the exam as this would help them to perform better.

Another concern that some teachers in the study by Cumming et al. (2005b) expressed was that the tasks may be cognitively and intellectually more demanding than the previous tasks yet the target population used in the trial were students who were either just about to enter or who had just entered university. Consequently, this may suggest that such tasks actually discriminate well between those who have and have not yet acquired the academic and linguistic skills required for academic study (pp. 29-30).

Summarising the research on integrated testing, Cumming (2013) outlines the ‘promises and perils’ of these types of test tasks. He argues that they:

(a) confound the measurement of writing abilities with abilities to comprehend source materials; (b) muddle assessment and diagnostic information together; (c) involve genres that are ill-defined and so difficult to score; (d) require threshold levels of abilities for competent performance, producing test results that may not compare neatly across different ability levels; and (e) elicit texts in which the language from
source materials is hard to distinguish from examinees’ own language production (p. 1).

These are all valid points; however, integrated tasks are - perhaps most - frequently used for academic English tests so it should be borne in mind that these tasks do in fact replicate the TLU domain in which test takers are required to comprehend source materials and extract information from them. Furthermore, threshold levels of competence are required for entry to university. As such, these tasks may not be suitable for very low proficiency learners but such learners would not be admitted to university either.

Cumming (2013) also presents what he refers to as the ‘promises’ of integrated tasks, not least of which is the fact that the such writing is authentic, ‘challenging’ and ‘responsible to specific content’ (p. 2). He also states that integrated tasks ‘counter test method or practice effects associated with conventional item types’ (p. 2). Therefore, while there may be some fair criticisms of integrated tasks, it could be argued that such tasks, when designed well, will give us a much better insight into how a given test taker would be placed to ‘survive’ in the multiliteracy academic environment and can provide diagnostic information which can be used to plan teaching in the EAP classroom, which will have beneficial effects for students embarking on a university degree course.

2.8 Independent vs integrated writing tasks

2.8.1 Independent vs integrated writing tasks

Investigating students’ writing needs in an academic environment, Leki and Carson (1994, p. 96) acknowledge that there does seem to be a need for undergraduates to move away from knowledge-telling to knowledge-transforming. Consequently, they suggest that practising writing ‘in and of itself’ (Leki and Carson, 1994, p. 97) does not sufficiently prepare students for their degree programmes because composition tasks such as those usually practised in EAP classes ‘may not require knowledge transformation and therefore may do little to promote ESL students’ participation in the academic culture’ (p. 97). They therefore contend that EAP
writing classes should move away from opinion and experience-based tasks towards tasks that require the integration of opinions and experiences with evidence from other sources (p. 95).

More recently, Gholami and Alinasab (2017), called for more source-based writing practice in writing courses. Recognising that integrated writing tasks are frequently-used in the academic environment, but rarely practised in writing classes, they investigated the effect of practising source-based writing, and the impact that this had on scores on independent and source-based integrated essays in a writing course. The study comprised two groups: the ‘hybrid’ writing approach group, which combined both independent essay writing practice with source-based writing practice, and an ‘independent writing practice only’ group, which served as a control group. They found that the ‘hybrid’ group outperformed the control group on integrated tasks but not in the independent tasks. This was only a small study with 20 participants, so the results should be treated with caution; however, it is likely that, even in a bigger study, students who practise integrated writing will perform better than a group which has no practice in this.

This tension between the independent writing task, which provides information about a student’s ability to write but may limit their performance due to the lack of topic knowledge or which may not prepare them for the academic environment, and an integrated EAP writing task, in which input is provided to ensure everyone has the same knowledge base to start from and to better reflect an authentic academic task, raises the question of whether these two types of essays tap into the same constructs.

The findings of research investigating the differences between independent and integrated writing tasks are mixed with a number of studies suggesting that the two types of writing measure the same construct (Gebril, 2009; Lee and Kantor, 2007) while others argue that independent and integrated tasks test different constructs (Cumming et al., 2005a; Plakans, 2008). Guo, Crossley and McNamara (2013) found that independent and integrated tasks
share some construct coverage for discourse features but this ‘does not fully overlap’ (p. 230) because the two types of task ‘tap into different elements of writing’ (p. 234).

Cumming et al., (2005a) compared performance on a TOEFL independent writing task, a reading-into-writing task and a listening-into-writing task. They found significant differences between the three types of text with the integrated essays tending to be shorter than the independent ones but the former included both longer words and clauses and a wider range of words and clauses. However, it should be borne in mind that test takers were given longer to write the independent tasks (30 mins) and were told in the prompt that ‘an effective response will contain a minimum (my emphasis) of 300 words’ (p. 13) whereas for the integrated reading-into-writing test, they had 25 mins and were told that ‘an effective response will be between 175 and 200 (my emphasis) words’ (p. 13). For the integrated listening-into-writing test, test takers had only 15 mins and were told that ‘an effective response will be between 125 and 200 (my emphasis) words’ (p. 13). Therefore, by giving a range for the length for the two integrated tasks as well as less time, this is likely to have limited how much test takers wrote. They also found that the integrated essay tasks tended to be less argumentatively oriented but, in contrast to the independent essays, included the use of source texts in paraphrased, summarised or verbatim reproductions.

In a small, qualitative study with only 10 participants, Plakans (2008) looked at composing processes across independent and integrated tasks on a university placement test. Although this was only very small-scale so the results may not be generalisable more widely, she found that the independent writing task required ‘more initial and less online planning’ while the integrated reading-into-writing task elicited ‘a more interactive process’ (p. 111) and was the test takers’ preferred task.

Looking at score generalisability, Gebril (2009) compared scores on reading-to-writing and independent writing tasks and found equal reliability in the scores while Esmaeili (2002), on
the other hand, found that students performed significantly better when there was a thematic connection between their reading and writing tasks compared to when this was not the case.

Similarly, Read (1990) carried out research looking at performance on three different types of writing tasks: independent tasks, in which no input or guidance is provided; guided tasks, which provide guidance during the writing process by providing, for example, a graph or a picture; and experience tasks, which allow students to learn about a topic before being asked to write on the topic. The resulting correlation coefficient was not high and highlighted a possible impact of test taker variables. Some test takers were proficient, but slow, writers, who, if not under time constraints, might have produced better essays. This is an important consideration because in a post-graduate situation in particular, test takers are not generally required to write under such time constraints. As a result, this raises questions about the validity of such timed writing tests because test tasks should reflect authentic TLU tasks to provide validity evidence for the scores on which decisions are based. In this case, the score may not be giving a result which reflects or predicts test takers’ true abilities. Nonetheless, Read (1990, p. 113) suggests that, ‘when the writing tasks are linked with earlier reading and listening tasks, [they] may represent a better simulation of the process of academic study than simply giving a stand-alone writing test’.

Despite these differences between independent and integrated tasks, research looking at performance on the IELTS Task 1 - an example of an integrated test which uses only a visual as input - and Task 2 - the independent essay task – found no statistically significant differences between the two tasks (Ahmadi and Mansoordeghchan, 2014; Sheibani and Ahmadi 2018). However, this could be related to the fact that they are both part of the same test and students may have practised these types of tasks many times before while preparing to take an IELTS test.

While a good deal of research has investigated the differences between independent and integrated tasks for the TOEFL test (Barkaoui, 2015; Brown, Iwashita and McNamara, 2005;
Cumming et al., 2005a; 2005b; Guo, Crossley and McNamara, 2013; Lee and Kantor, 2007), the lack of research investigating the differences between the independent essay task for IELTS and other, text-based integrated writing tasks (as opposed to the visual input provided by the IELTS Task 1) is conspicuous by its absence. It appears then that this is an area which could usefully be investigated.

2.8.2 Factors affecting performance on integrated writing tasks
Aside from comparing differences between independent and integrated tasks and critiquing integrated tasks, other research has looked at variables that influence performance on integrated writing tasks. These include task difficulty in integrated test tasks, focusing on prompt characteristics (Ahmadi and Mansoordeghan, 2014; Cho, Rijmen and Novak, 2013; Mickan, Slater and Gibson, 2000), the factors that account for the variance in scores (Ascención Delaney, 2008; Cumming, 1989; Soleimani and Mahdavipour, 2014) while other research has focused on the strategies (Yang, 2012) and processes (Plakans, 2008; 2009b) involved in integrated writing, and source text use (Cumming et al., 2005a; Gebril and Plakans, 2009; 2013, Neumann, Leu and McDonough, 2019; Plakans and Gebril, 2012; 2013; Shi, 2004; 2012; Weigle and Parker, 2012; Wette, 2017). These will be discussed in the coming sections.

2.8.2.1 Task difficulty in integrated tasks
When comparing task difficulty in IELTS Tasks 1 and 2, Ahmadi and Masoordeghan (2014) found a significant effect on performance in the IELTS Task 2 (independent essay) but no significant difference for IELTS Task 1 (integrated visual-into-writing task). Cho, Rijmen and Novák (2013) found that language proficiency contributed to score variance, along with ‘distinctness of ideas within the prompt and difficulty of ideas in the passage’ (p. 513). They also looked at whether the perceived task difficulty of the prompt contributes to performance. All test takers were asked about their perceptions of the difficulty of the prompts but the authors did not compare test takers’ perceptions with actual performance. This seems like a
missed opportunity to assess the extent to which the perceptions converge or diverge with the reality of test takers’ performance.

Investigating the effect of task prompts on the writing sub-test of IELTS, that is both the integrated visual-into-writing task (Task 1) and the independent essay (Task 2), Mickan, Slater and Gibson (2000) found that pragmatic and lexico-grammatical features influence test takers’ understanding of the prompts, which in turn affects their ability to produce satisfactory responses.

On the other hand, Ascención Delaney (2008) looked at the factors that affect scores on a reading-into-writing summary task and a reading-into-writing response essay and found that there was only a weak relationship between scores on the reading-into-writing test and reading ability while there was no relationship between test scores and writing scores. Consequently, she claims that reading-to-write ability involves ‘an interaction of these skills affected by internal and external factors’ (p. 147). This is likely where other academic and mediation skills (Council of Europe, 2018) play an important role in task achievement. In her study, she did, however, find that there was a minimal effect on performance on the reading-to-write tasks caused by language proficiency and level of education, with native speakers and advanced level students outperforming intermediate students and EFL students, and postgraduates outperforming undergraduates. This is in line with findings from Cumming, (1989) who found that language proficiency was an ‘additive factor’ (p. 81) which improved the overall quality of writing but did not affect composing processes. Writing expertise, however, did indeed relate to such processes.

2.8.2.2 Test-taker strategies and processes in integrated writing
Several studies have investigated the processes and strategies involved in integrated writing tasks (Bridges, 2010; Plakans, 2009b; Yang and Plakans, 2012; Yang, 2012).

In a study using think-aloud protocols, Plakans (2009a) investigated the processes involved in completing a reading-into-writing task. She based the interpretations on Spivey’s (1984, cited
in Spivey and King, 1989) definition of discourse synthesis (see Section 2.7.1) above, which comprises the three processes of selecting, organising and connecting ideas from source texts. Plakans (2009b) found that four of the six participants in the study used discourse synthesis in their writing with test takers tending to focus more on connecting and organising than selecting. She also found that language proficiency may have had an impact on writers’ processes, as the lower proficiency learners struggled with a lack of vocabulary. While this study reveals the value of discourse synthesis for integrated writing tasks, a larger study would be necessary to confirm (or otherwise) the findings of such a small study.

Bridges (2010) identified six composing processes engaged in by test takers completing the IELTS Task 1, which is considered an integrated task by a number of researchers (Ahmadi and Mansoordehghan, 2014; Bridges, 2010; Sheibani and Ahmadi, 2018; Yang, 2012) due to the provision of input in the form of a graph. These were: macro-planning, organizing, micro-planning, translating, monitoring, and revising. He found that test takers engaged in organizing less than the other processes, thus leading him to conclude that this task may actually be more of a knowledge-telling than a knowledge-transforming task (Bridges, 2010, p. 32).

Similarly, Yang and Plakans (2012) looked into strategy use and performance on an integrated reading-listening-writing task. Using structural equation modelling, their study revealed that the construct of integrated writing comprises three factors. The first of these, self-regulatory strategy use (SEFLS) ‘had an executive control over other types of strategy use’ (p. 80). The other two strategies, discourse synthesis strategy use (DSS) and ‘test-wiseness’ strategy use (TWS) had a positive and a negative impact respectively on test takers’ performance. They suggest that such an integrated writing task requires both receptive and productive skills as well as ‘regulation skills’ (p. 80) to manage the interaction between the different skills.

The negative impact of test-wiseness strategies is evidence that scores are not artificially inflated through such strategies and thus they do not contribute to construct-irrelevant
variance in this task. Another point that Yang and Plakans (2012) highlight is that the ‘test takers’ L2 writing ability was not proportionate to their ability to use sources appropriately’ (p. 94) so a test taker with a high score for content and language may, in fact, receive a low score for verbatim source use. This is an important point as it highlights how source text use is a separate (academic) skill which is not directly linked to language proficiency and, as such, should be taken into account in the design of a task-specific rating scale.

Yang (2012) investigated test taker strategies in an integrated task based on visual input containing a graph. She found that the task generally involved test takers in the processes of ‘graph comprehension, graph interpretation, and graph translation strategies’ (p. 174) but writers experienced difficulties centred on knowledge and use of lexis. This finding highlights one of the advantages of language-rich input texts compared to graphs as input, namely, that the textual input ‘serv[es] as a language repository’ (Plakans and Gebril, 2012, p. 18).

Furthermore, it should be pointed out here that Knoch and Sitajalabhorn (2013, p. 304) argue that only tasks which involve a ‘significant proportion of language’ being provided as part of the input should be considered integrated tasks and, for a task to be an integrated task, the source text language should be ‘used and transformed to complete the writing task’. They further argue that the ‘stimulus materials need to provide sufficient language (either in written or audio format) to allow writers to produce sufficient text to be rated by assessors’ (p. 304).

2.8.3 Source text use in integrated writing tasks

It should now be clear from the evidence above that the use of source texts in academic tasks is a fundamental skill required of students at university. There are a number of reasons why the use of source texts is advocated. The first, as we have already seen in Section 2.6.5, is that of fairness so that everyone has access to the same information for the written output (Weir, 1993, p. 135, quoted in Lewkowicz, 1994, p. 204). Another is that in many disciplines, students have to extract pertinent information from oral and written input, which they then transform into their own extended writing assignment (Weir, 1983, p. 378, quoted in
Lewkowicz, 1994, p. 204). However, source texts also serve several other useful functions including being a springboard for generating ideas (Plakans and Gebril, 2012; Leki and Carson, 1994) and a source of language (Plakans and Gebril, 2012, p. 18).

Source text use is a key factor contributing to the construct validity of integrated tests and success in academic writing, yet many students struggle to acquire this skill. Consequently, the use of source texts in academic English has received considerable attention in the research. Studies have looked at how students develop their ability to use sources and the challenges they face in doing so both in terms of larger-scale quantitative and qualitative studies (Neumann, Leu and McDonough, 2019) and longitudinal, qualitative research (Davies, 2013), although as is perhaps already apparent from the studies cited above, the vast majority of research into integrated testing looks at reading-into-writing tasks (Campbell, 1990; Charge and Taylor, 1997; Cumming, Lai and Cho, 2016; Davies, 2013; Keck, 2006; Leki and Carson, 1994; 1997; Neumann, Leu and McDonough, 2019; Plakans and Gebril, 2012; 2013; Shi, 2004; 2012; Weigle and Parker, 2012; Wette, 2017; Wolfersberger, 2007; Yu, 2009) with a number of other studies looking at reading and listening-into-writing, mainly focussed on the TOEFL iBT (Cumming et al., 2005a; 2005b; Plakans and Gebril, 2013). However, there is very little research on literature on listening-into-writing by comparison.

When investigating source text use, researchers have looked at both L1 and L2 expository writing (Campbell, 1990; Weigle and Parker, 2012), summary writing (Keck, 2006) or both (Shi, 2004) and the results have been mixed.

Campbell (1990) investigated source text use in L1 and L2 expository writing. Distinguishing between quotations, copy (exact and near) and paraphrase, she found that most students copied from the source and, although nearly half of L2 students acknowledged the authors, only 16 per cent of L1 students did this. Conversely, Moore (1997, cited in Shi, 2004) found that L2 students were less likely to cite the sources.
Keck (2006) differentiates between types of paraphrase slightly differently to Campbell (1990) using 'Near Copy', 'Minimal Revision', 'Moderate Revision' and 'Substantial Revision' (p. 261) to code the summaries in her study. She found that L2 learners had significantly more 'Near Copies' than the L1 learners, who revised moderately and substantially significantly more.

Shi (2004) included both the type of source use (copy, modification or reformulation) and citation type (quoted, referenced or no reference). The study investigated the differences between a summary writing task and an opinion task by both L1 and L2 students. The findings revealed more direct source text use among the students who completed the summary task compared to the opinion task. Like Moore (1997, cited in Shi, 2004), Shi (2004) also found that the Chinese learners tended not to reference the sources for either task.

Other studies have looked at source text use compared to proficiency level (Cumming et al., 2005a; Plakans and Gebril, 2009; 2012; Weigle and Parker, 2012).

Weigle and Parker (2012, p. 118) amended Shi’s (2004) system to include the length of string (the number of consecutive words from the source text), as well as source text type (direct copy, a modified quotation, that is, a quotation which was copied incorrectly in some way, or a reformulation) and citation type (quotation, referenced or not referenced). They consider a match as being a string of at least two content words ‘from the same clause’ or at least three consecutive words ‘that form a syntactic constituent’ (Weigle and Parker, 2012, p. 121-122). This study revealed that textual borrowing did not appear to be a major issue for most test takers with only a few of the lower scoring test takers borrowing considerable amounts from the text. The study also showed that there was no significant difference in the amount of textual borrowing across participants of different levels.

Weigle and Parker’s (2012) findings are partly in line with Johns and Mayes (1990), who compared idea units in summaries written by two different proficiency levels. Johns and Mayes (1990) also found that lower proficiency test takers copied more and struggled to differentiate
between key points and details. However, there were ‘few significant differences between high-and low-proficiency ESL university students in the idea unit manipulation which produces summary products’ (p. 265). This may have been due to the level of complexity of the input text, which was a lower-intermediate business text designed to be within the linguistic ability of all the participants. As such, this text may not have been suitable to allow for differentiation between the two levels of proficiency. Similarly, the requirement for the length of the test was 20 per cent of the original so selecting what to delete and include may not have been too difficult. However, in contrast to Weigle and Parker (2012), Johns and Mayes (1990, p. 253) found significant differences across the two proficiency groups in terms of replication of sentences from the source text and ‘combinations of idea units taken from two or more punctuated sentences’.

In a reading-into-writing study, Plakans and Gebril (2012, p. 18) found that proficiency affected input text comprehension but was not related to ‘source use functions’.

As part of a larger study into the discourse produced in prototype reading-into-writing and listening-into-writing TOEFL tasks compared with independent essay tasks, Cumming et al. (2005a) looked at the differences in source text use in the integrated tasks. It should be borne in mind that the audio input in these prototype TOEFL test uses only audio-based listening texts. They found significant differences in source text use between the reading-into-writing test and the listening-into-writing test at proficiency levels 3 and 5, that is the lowest and the highest of the three groups in the study: as examinees’ proficiency increased, verbatim reproductions decreased for the reading-into-writing task but the converse was true for the listening-into-writing task. The more proficient test takers were able to summarise and synthesise ideas from the input texts coherently whereas the least proficient ones were unable to comprehend the input texts well enough to be able to paraphrase or summarise. The mid-range test takers tended to produce a mixture of verbatim and paraphrased information. In the integrated listening-into-writing task, the least proficient writers did not use information
from the text directly, unlike the mid-range proficiency test takers. The weaker ones also used formulaic phrases to structure their texts. The more proficient ones, on the other hand, used information from the text but summarised or paraphrased it.

Cumming et al. (2005a) provide several reasons for these findings: 1) they are due to the test takers’ comprehension of the source texts, 2) test takers had access to the reading text as they wrote but not the listening text, 3) there could have been some effect resulting from the task characteristics, for example, the rhetorical organisation of the input texts, and 4) the length of the texts: the reading-into-writing task was 25 mins while the listening-into-writing task was only 15 mins.

On the other hand, Plakans and Gebril (2013) investigated performance on the TOEFL iBT, which integrates listening and reading into writing. They looked at the importance of ideas used from the source texts, ideas reproduced from different input texts (reading and listening) and verbatim source text use. They found that more proficient learners used both source texts and used the listening text as instructed in the prompt. On the other hand, in line with Cumming et al. (2005a), they found that lower-level learners ‘relied heavily on the reading texts for content and direct copying of words and phrases’ (p. 217). However, the correlation between test takers’ ability to select main ideas and verbatim source text use was very low, suggesting that those who copied were not successfully selecting and copying the main ideas.

They also highlight the fact that, in contrast to studies investigating reading-into-writing only, the inclusion of the listening skill in this study ‘has a significant impact on score, more so than other features that may be more commonly associated with integrated writing, such as verbatim source use’ (p. 227). In line with Field (2008, p. 115), they suggest that the impact of the listening on scores may be a result of this form of input tapping into ‘higher proficiency writers’ automaticity in language comprehension’ (Plakans and Gebril, 2013, p. 227) whereas the reading input provided both content and language support to lower level learners who
were not able to access the listening text to the extent necessary to complete the task satisfactorily.

Both the study by Plakans and Gebril (2013) and the one by Cumming et al. (2005a) used audio input only for the listening input so, while there appears to be clear evidence of the reliance on reading texts by learners at lower proficiency levels in a reading- and listening-into-writing test, it is not clear whether the use of visuals, for example, Power Point slides with key points on them would provide the same level of support for the listening input in a listening-into-writing test. Indeed, the suggestion to include Power Point as a visual aid was made by two of the informants in Cumming et al.’s (2005b) teacher-verification study of the prototype TOEFL tasks as a way to ‘facilitat[e] examinees’ comprehension’. The impact of such visual aids in a listening-into-writing test appears to be a gap in the literature which could usefully be addressed.

Although most of the studies outlined above focus on reading-into-writing to investigate the use of source texts, which undoubtedly contributes to our knowledge base in terms of the way students interact with written sources, it appears that very little emphasis has been placed on source text use in listening-into-writing tasks. Given how much research has been done on the use of source texts, it is somewhat surprising that such a gap exists, yet it does. Therefore, more research into this area would complement the work done on reading-into-writing tests.

2.8.4 Listening-into-writing research
It is not just the research into source text use in listening-into-writing tests that is scarce; there is a dearth of research into these types of integrated tasks in general. As outlined in Section 2.3.4 and Section 2.4 above, a good deal of work has looked at the impact of different input formats in listening tests (Aldera, 2015; Batty, 2015; Brett, 1997; Coniam, 2001; Gruba, 1993; Leveridge and Yang, 2013; Montero Perez, Peter and Desmet, 2014; Ockey, 2007; Pardo-Ballester, 2016; Suvorov, 2008; 2013; Vandergrift, 2004; Wagner, 2007; 2010a) and the work around academic listening has focussed on note-taking (Carrell, 2007; Chaudron,
Loschky and Cook, 1994; Chaudron and Richards, 1986; DeCarrico and Nattinger, 1988; Flowerdew, 1994; Flowerdew and Miller, 1997; Hayati and Jallilifar, 2009; Piolat, Olive and Kellogg, 2005) yet there is very little on integrated listening-into-writing tests. Nonetheless, a number of studies have looked at listening combined with reading into writing, (Plakans and Gebril, 2013; Sawaki, Quinlan, and Lee, 2013; Zhu et al., 2016), and a few studies have investigated listening-into-writing (Cubilo and Winke, 2013; Cumming et al., 2005a; 2005b; Rukhthong and Brunfaut, 2020).

As mentioned above, Cumming et al. (2005a) investigated the differences in discourse produced in independent TOEFL essays compared to integrated reading-into-writing and listening-into-writing tasks and found significant differences in the discourse variables between the independent and integrated tasks.

Cumming et al. (2005b) looked at the concurrent validity of several prototype tasks for the TOEFL test. These included independent writing tasks, listening-into-writing, and reading-into-writing among others. Feedback received from the ESL instructors who participated in the study suggests that the students’ performance on these tasks is in line with their performance in class.

Sawaki, Quinlan and Lee (2013) also investigated reading and listening into writing on the TOEFL iBT. Test takers read a text and listened to a lecture then summarised the links between the two input formats. Using a Confirmatory Factor Analysis, the study investigated the relationship between the underlying constructs in the written responses produced by the participants, who were divided into lower and higher level groups. A significant difference was found between the two groups for the three factors they analysed, namely, ‘Comprehension, Sentence Conventions, and Productive Vocabulary’ (p. 90). The study also revealed that the Listening factor (along with those for Content and Vocabulary) were ‘strongly associated’ with the Comprehension factor. They conclude that content in integrated writing responses does differentiate between learners of different levels. While this study investigates the impact of
the input and aspects of the language produced, their grouping variable is higher- and lower-level learners rather than distinguishing between input media, since the TOEFL iBT only uses one input medium.

Plakans and Gebril (2013) (see Section 2.8.3) investigated source text use in the TOEFL iBT. They found that listening text use correlated most highly with integrated writing score. On the other hand, they found that there was a negative correlation between listening text and reading text use. Furthermore, a regression analysis revealed that listening text use ‘had the highest predictive value’ (p. 225) of all variables used in the study. These findings may be more pronounced because test takers only listened to the text once so lower level learners may have performed better if they had been allowed a second playback. Alternatively, in line with Field’s (2015) findings, a second playback may have benefitted all learners, making the distinctions between levels more pronounced. In any case, providing an opportunity for learners to perform better is an argument in favour of a second playback.

Zhu et al. (2016) compared performance on an independent listening test with performance on integrated listening-reading-writing tasks in Chinese for L1 Chinese high school students. Test takers listened to two recordings on one topic then answered 14 multiple-choice and short answer items designed to tap into the following six listening competences: ‘memorization’, ‘explanation’, ‘summarization’, ‘elaboration’, ‘evaluation’ and ‘creation’ (p. 172). The integrated task comprised an audio text featuring two students giving opposing opinions on a topic and 5 texts (including diagrams) which were related to the topic of discussion. Test takers had to assume the role of one of the speakers and write a text expressing their opinion on the topic.

Clearly, this is a slightly different type of task to other integrated writing tasks, which tend to focus on lecture input and academic texts with a follow-up essay or summary. Nonetheless, they found a statistically significant correlation between scores on the two tests, yet a follow-up factor analysis revealed that ‘the listening competence factor had no commonality with the
integrated test’ (Zhu et al., 2016, p. 177). They also carried out a regression analysis to investigate whether the indicators used for the listening test were predictive of performance in the integrated listening-reading-writing test and found some limited evidence of predictability of the listening test to the integrated task. Both ‘evaluation’ and ‘creation’ were found to have low predictive power but did predict scores on the listening-reading-writing task significantly. On the basis of the results, they suggest that ‘higher-order thinking skills are more important for successful performance in the Listening-Reading-Writing Task than lower-order thinking skills’ since the tasks require test takers to evaluate the opinions presented in the input texts and to develop ideas of their own related to the input (p. 179).

In one of the very few studies looking at listening-into-writing (without reading input), Rukhthong and Brunfaut (2020) investigated the listening processes and strategies that ESL learners use in an integrated listening-into-writing task and then looked into the differences between task modalities and performance levels. 72 test takers completed four tasks, which involved them listening to four excerpts from academic lectures and completing two oral content summaries and two written content summaries. 60 of the participants responded to a background questionnaire while 12 of them took part in stimulated recalls for the four tasks. Their findings suggest that more successful listeners use both high and low-level processes while less successful learners only engaged in low-level processes. Successful listeners are also able to assess the input in real-time, deciding what is important or not, and monitor comprehension more effectively than lower level learners. With regard to lower level processes, the higher level learners, not surprisingly, were found to have a higher degree of automaticity and accuracy in phonemic processing, word recognition and parsing. They also appeared to be more accurate on tasks that required inferencing skills. They concluded that listening should be considered one of the constructs in listening-to-speak and listening-to-write tasks.
Rukthong (2015) also investigated test takers’ perceptions of the tasks in the above-mentioned study. She found that 90 per cent of test takers either agreed or strongly agreed that the listening-to-write summary tasks were authentic and 80.8 per cent believed that the tasks assess the ability required in an academic environment. Over two-thirds of participants believed that such tasks assessed their language level fairly. However, when she compared participants’ perceptions of various aspects of task difficulty with their test scores, she found only a few significantly strong correlations between the two (p. 266). She also found both positive and negative correlations for other aspects of task difficulty but many were not significant. She concluded that this could be due to the fact that despite finding the tasks difficult, the test takers were still able to perform them successfully if ‘their cognitive processing worked effectively’ (p. 309).

As part of a larger study looking at the impact of note-taking strategies on performance on the LanguEdge test, Carrell (2007) looked at the impact of note-taking on performance in an integrated listening-into-writing test in which students listened to short lectures (4-6 minutes each in length) delivered at speech rates of between 143 and 208 words per minute. They found significant, positive correlations between the number of content words and performance on the integrated listening-into-writing test. They also found that, in line with Faraco, Barbier and Piolat (2002), test takers paraphrased very little. They suggest that this may have been due to the increased cognitive load caused by trying to paraphrase while taking notes. However, it may also be that test takers did not have time to write out as much as they wanted because the speed of delivery may also have affected their ability to take notes in those lectures which had a higher speech rate.

Cubilo and Winke (2013) looked at the impact of different input formats (audio with still pictures vs video) on students’ note-taking strategies and performance on a follow-up integrated writing task. Using a counter-balanced design to control for potential order or content effect, the study involved participants each writing two essays, one for each input
format, and completing a post-task survey to investigate test takers’ perceptions of the usefulness of the video input. Although the findings did not reveal any significant differences between the input formats in four of the five rating criteria (‘content’, organization’, ‘vocabulary’ and ‘mechanics’ and in overall scores (p. 382), test takers received a significantly higher score for the criterion, ‘language use’, for the essays which had been supported by video input.

The lack of a significant difference between the two formats could be (at least in part) accounted for by the fact that the video included context-only visuals (that is, they saw the lecturer who was speaking so could pay attention to non-verbal information but had no visual input such as PowerPoint slides). Such content visuals may have produced different results.

On the other hand, test takers’ notes revealed a difference between the two formats in terms of the quantity of notes made, with test takers making significantly fewer notes when presented with the video input. Rather than this being due to their observing the speaker, the qualitative analysis revealed that this was due to the cognitive overload of listening, watching the video and taking notes at the same time. Another finding from the qualitative analysis was that the video was perceived to be faster even though the speed of delivery for both inputs was the same. Nonetheless, the authors did find that the nonverbal cues in the video input aided retention of information for some test takers.

This appears to be the only study which looks at listening-into-writing using audio and video input with a follow-up writing task. However, as mentioned, the video input was not content-based while the audio input used still pictures and it is not clear what the impact of the latter was.

With regard to test takers’ perceptions of the input, in line with other studies outlined above (Coniam, 2001; Ockey, 2007; Pardo-Ballester, 2016; Suvorov, 2008), the findings were mixed. 70 per cent (n=28) of the 40 participants said that video was their preference. The reasons
given include the fact that the visual cues aided comprehension, video is more authentic and it was more engaging. Conversely, 25 per cent (n=10) said they preferred the audio only input, mainly because they found it easier to focus on the content, while the other two students said that they closed their eyes while listening. Participants were also asked if they found the video helpful or distracting. Here, 63 per cent said they found it helpful, which is not as many as preferred the use of video over audio and still pictures but this may have been due to the fact that the input was only context-based. 11 participants said they found it distracting, one of whom had said they actually preferred the video input. This suggests that while the use of video input may be preferable, it can also be as distracting at the same time.

As can be seen from the studies outlined above, there have been several studies looking at the integration of reading, listening and writing (Cumming et al., 2005a; 2005b; Sawaki, Quinlan and Lee, 2013; Zhu et al., 2016) but the research specifically on listening-into-writing is very limited, restricted to the listening processes and strategies used by test takers completing integrated listening-into-writing tests (Rukhthong and Brunfaut, 2020) and, while Carrell’s (2007) study looked at integrated listening-into-writing, the only study which has investigated the impact of audio and video listening inputs on performance in an integrated EAP listening-into-writing test has been the study by Cubilo and Winke (2013). While this study investigates the impact of different input formats, the researchers use context video instead of content video and use still photographs with the audio input, which may have mitigated, at least to some extent, the effect of the different input formats.

Furthermore, while Cubilo and Winke (2013) and Carrell (2007) both look at the effect of notes on performance, neither study investigates verbatim reproductions from the source text to see the extent to which test takers engage with the source text. As such, this is an important gap in the literature. In the current climate, in which more and more input is being provided in both audio and audiovisual format, both face-to-face and online, it is important to investigate the impact of the source texts on the productive task.
2.8.5 Rating integrated tasks

Some studies have also considered the impact of integrated tasks on rating and raters (Chan, Inoue and Taylor, 2015; Cumming, Kantor and Powers, 2001; Gebril and Plakans, 2014).

Cumming, Kantor and Powers (2001) investigated raters’ decision-making behaviour when rating independent and integrated TOEFL essays using think-aloud protocols. They found that the integrated tasks led to raters paying more attention to language at the lower levels, and rhetoric and organisation at the higher levels. They also found that for the integrated writing tasks, the raters needed to consider not only the ‘formal aspects of composing and argumentation’ (p. 49) but also how test takers used the source texts and how well they achieved the task.

Gebril and Plakans (2014) also used think aloud protocols to investigate raters’ decision making processes. Their raters also considered distinguishing between source text use and the students’ own words to be challenging. Raters spent a considerable amount of time judging the extent to which the source material had been integrated appropriately into student writing. As such, rating scales and rater training need to consider these issues.

Chan, Inoue and Taylor (2015) report on their mixed methods study which set out to redevelop and validate rubrics for assessing reading-into-writing skills. In line with Knoch and Sitajalabhorn (2013), they stress the need for rating scales for integrated writing tasks to account for the specific features of such tasks.

Knoch and Sitajalabhorn (2013) state that a rating scale for an integrated task should include descriptors which measure test takers’ ability to ‘mine, select, organise and connect ideas’ in the written output and should also account for the transformation of the language from that used in the source text to that used in the output. They also suggest that the inclusion of descriptors which measure test takers’ ability to appropriately acknowledge the sources is worthwhile in an academic writing task.
2.8.6 Integrated testing and plagiarism

Due to the difficulties experienced by students of integrating sources into their texts (see Section 2.7.3 and Section 2.8.3), other areas investigated in relation to source text use include why international students plagiarise (East, 2005), problems that international students have with citation (Keck, 2006; Pecorari, 2003; Shi, 2012), what constitutes plagiarism (Pecorari and Shaw, 2012), how to support students in avoiding plagiarism (Westbrook and Holt, 2015), and the effect of language level on plagiarism (Lee, 2010; Schmitt, 2005). Unsurprisingly, the latter studies suggest that lower level candidates struggle more with plagiarism and paraphrasing. This could be due to a lack of vocabulary (and grammar), which impacts on their ability to decode and manipulate the input.

2.9 Text difficulty

2.9.1 Factors contributing to text difficulty

Green, Ünaldi and Weir (2010) outline a number of factors that can contribute to text difficulty in reading texts. These include the length of the input texts, lexical and grammatical characteristics, the impact of the presence or absence of cohesive devices, genre, topic, the readers’ background and cultural knowledge, and the concreteness or abstractness of the text (pp. 194-197).

As mentioned in Section 2.3 above, Bloomfield et al. (2010) outline the factors that contribute to difficulty in listening. These can be broken down into three areas: characteristics of the listener, the text and the conditions in which the test is taken. Listener characteristics include working memory (see Section 2.3.1.1), anxiety (see Section 2.3.1.2), and L2 proficiency (see Section 2.2.1). Regarding the text itself, text characteristics include text length (see Section 2.3.2.1), complexity, genre and text organisation, and auditory features such as accent (see Section 2.3.2.2), and speech rate (see Section 2.3.2.3). The testing conditions include considerations such as number of playbacks (see Section 2.3.3.1) and note-taking (see Section 2.4.1) (Bloomfield et al., 2010, p. 1). Since most of these characteristics have already been
dealt with in Section 2.3 of this literature review, this section will focus on text difficulty in terms of linguistic complexity and how to measure this.

2.9.1.1 The role of vocabulary in text difficulty
Vocabulary has been found to be linked to comprehension in all four skills. ‘Lexical coverage’ (Adolphs and Schmitt, 2003) refers to ‘the percentage of known words in a text’ (Van Zeeland and Schmitt, 2013, p. 457).

Research into reading comprehension has found that readers can adequately understand a text if they have 98 per cent lexical coverage (Van Zeeland and Schmitt, 2013, p. 459). However, what is considered adequate varies. Laufer (1989, cited in Van Zeeland and Schmitt, 2013, p. 459) considers a score of 55 per cent on the test in her study as ‘adequate’ while Hu and Nation (2000) set the boundary between inadequate and adequate comprehension at a score of 70 out of 124 on their reading test (approximately 56.5 per cent) (p. 416).

Laufer and Ravenhorst-Kalovski (2010) investigated the relationship between L2 ‘learners’ vocabulary size, the lexical text coverage that their vocabulary provides and their reading comprehension’ (p. 15) and found that vocabulary knowledge contributes incrementally to reading comprehension, accounting for 64 per cent of the variance in reading score (p. 26). As such, they conclude that vocabulary is likely to be the ‘major factor’ in reading comprehension (p. 26) and consider 95 per cent lexical coverage, that is a vocabulary size of 4000-5000 words, as the minimal coverage (p. 26) and 98 per cent as optimal, requiring a vocabulary size of 6000-8000 words (p. 25). They point out, however, that reading more technical texts may require ‘knowledge of more low frequency words’ than the vocabulary sizes suggested (p. 26).

Similarly, vocabulary has also been found to play ‘an essential role’ in oral production (Koizumi 2013, p. 1, quoted in Text Inspector, 2019). Adolphs and Schmitt (2003) cite research which found that ‘2000 word families provide the lexical resources to engage in everyday spoken English discourse’ (Schonell et al., 1956, cited in Adolphs and Schmitt, 2003, p. 425). However,
in their study, Adolphs and Schmitt (2003) found that the vocabulary size required to achieve 96 per cent lexical coverage 5000 words but it is unclear as to how this number relates to 2000 word families (p. 425).

More importantly for this study, vocabulary has also been found to play an important role in listening (Bonk, 2000; Stæhr, 2008; 2009; Van Zeeland and Schmitt, 2013) and writing skills (Laufer and Nation, 1995; Stæhr, 2008), although it is recognised that other factors play a role in both listening ability (Van Zeeland, 2013, p. 1, quoted in Text Inspector, 2019) and writing ability, too (Laufer and Nation, 1995; Stæhr, 2008).

Bonk (2000) found that comprehension of listening texts improved as familiarity with lexis increased. At three levels of proficiency (low, intermediate and high), learners rarely achieved good levels of comprehension if they understood less than 75 per cent of the lexis while good comprehension was achieved frequently when learners understood 90 per cent and above (p. 14). ‘Inferior’ comprehension was defined by a score of 1-2 out of 4 on the holistic rating scale used whereas ‘Good’ comprehension was defined by a score 3-4) (p. 21). He concludes that the study does not give rise to ‘a particular lexical threshold for achieving good comprehension across all learners’ (p.24) as results were varied. He suggests that the use of listening strategies may have helped some participants but almost a quarter of participants struggled to understand connected speech so this must be a goal for learners who want to use the language ‘for purposes other than speaking to each other in class’ (p. 28).

Stæhr (2008) found that, although vocabulary size was only moderately correlated with listening comprehension, the former still explained ‘a significant and substantial portion of the variance’ in the listening test scores (p. 139). He concluded that, although a vocabulary size of 6000-9000 words is necessary for ‘optimal reading and listening’ (p. 150), knowing ‘the most frequent 2000 or 3000 words in English’ (p. 150) would considerably help low-level learners. In a later study, he found that there was a significant correlation between vocabulary knowledge and listening comprehension (Stæhr, 2009). He also found that, in line with
findings from reading research (Laufer and Ravenhorst-Kalovski, 2010; Van Zeeland and Schmitt, 2013), 98 per cent lexical coverage is necessary to cope with the demands of the listening text in his study, which was taken from the Cambridge Certificate of Proficiency in English (Stæhr, 2009, p. 577).

Stæhr’s (2008) findings are also in line with Van Zeeland and Schmitt (2013), who also found that 2000-3000 word families is adequate for comprehension of ‘most spoken and written text’ (p. 150). They found that almost full comprehension was achieved with a lexical coverage of 100 per cent while very high comprehension was achieved when learners had 98 per cent lexical coverage. However, they suggest that ‘if less stringent comprehension rates are acceptable’ then this appears to be achievable with a lexical coverage between 90-95 per cent (p. 474).

With regard to writing and vocabulary, when validating the Lexical Frequency Profile (LFP) tool as a way of measuring ‘lexical richness’, Laufer and Nation (1995, p. 307) found vocabulary size can be expected to reflect productive use of language in a writing test. Similarly Stæhr (2008) found that writing and vocabulary correlated strongly and significantly (0.73) whereby vocabulary size accounted for over half of the variance in above average performances (p. 148), although they concede that this may have been affected by the fact that learners were allowed to use dictionaries.

From the studies above, it is clear that learners need to have a certain level of vocabulary to access listening and reading texts as well as to be able to produce oral and written discourse. Therefore, it is likely that if the requisite vocabulary size is not present, this will hinder both comprehension and production.

2.9.1.2 Tools for assessing text difficulty
When preparing teaching materials or test input, teachers and testers need to consider the complexity of the input text and, often, may need to simplify or amend the input to ensure
that it is of the correct level. To do this, the teacher or tester traditionally has two options. These are simplification through the ‘intuitive approach’ or the ‘structural approach’ (Allen, 2009, p. 586). In the case of the former, the author of the text uses their intuition to decide on the level of complexity; in the case of the latter, predetermined lists of grammatical structures and lexis are used for the different levels of proficiency (p. 586).

However, as a result of developments in the ‘fields of text linguistics, artificial intelligence, psychology, education and sociology’ (Graesser, McNamara and Louwerse, 2011, p. 35), the possibilities for assessing the level of a text have expanded considerably in recent years, with a number of automated textual analysis tools having been developed to analyse different aspects of text difficulty. Therefore, it is no longer necessary to rely on such primitive approaches. These tools range from fairly simple lexical diversity measures such as type-token ratio (Templin, 1957, cited in McCarthy and Jarvis, 2010, p. 381) to more complex measures of lexical diversity such as the ‘measure of textual lexical diversity’ (MTLD) (McCarthy and Jarvis, 2010, p. 381) and from Readability Formulae, such as Flesch Reading Ease (Flesch, 1948) and Flesch-Kincaid Grade Level (Kincaid et al., 1975, cited in Crossley, Allen and McNamara, 2011, p. 84), to far more complex tools such as Coh-Metrix (Graesser et al., 2004), Lexile® (MetaMetrics, 2020) or Text Inspector (Text Inspector, 2019), which measure various aspects of text difficulty using numerous metrics.

Lexical diversity measures such as type-token ratio calculate the ‘range of different words in a text’ and the higher the number, the more diverse the text (McCarthy and Jarvis, 2010, p. 381). The formula for the type-token ratio is extremely straightforward:

$$TTR = \frac{\text{number of types}}{\text{number of tokens}} \times 100$$

whereby the number of types is the number of unique types of words in a text and the number of tokens is the total number of words in the text. One of the criticisms of some lexical diversity measures, in particular type-token ratio, which is the most widely known measure, is that they
can be particularly sensitive to text length (McCarthy and Jarvis, 2010, p. 381) so caution should be exercised when interpreting this measure.

MTLD, however, is somewhat more complex than type-token ratio. The ‘number of words is divided by the total factor count’ (McCarthy and Jarvis, 2010, p. 385), whereby the factor is derived from the ‘mean length of sequential word strings that maintain a given TTR value’ (p. 384). This value is calculated twice: once counting forwards in the text and again going backwards from the end of the text. These two values are then averaged to give the MTLD (p. 385). However, one caveat of MTLD is that it can only analyse texts of 100 words and above. This needs to be borne in mind when considering the results of my study as texts from lower proficiency learners were not included in the MTLD analyses (see Section 5.3.2.3 and Section 6.3.1.3).

In a validity study for the MTLD, McCarthy and Jarvis (2010) compared this measure to several other well-known measures of lexical diversity including the type-token ratio. They found that the MTLD measure demonstrates ‘convergent validity’ with other sophisticated lexical diversity tools such as VOC-D or Maas, but they point out that the same cannot be said for its correlation with type-token ratio (p. 390).

Another option for measuring the difficulty of a text is the Flesch Reading Ease (Flesch, 1948) formula. This formula calculates the readability of a text based on sentence length and word length, and produces a number between 1 and 100 so the higher the number, the easier the text. The scale is based around US school grades so ‘a reading score of 60-70 is equivalent to a grade level of 8-9’. In other words, a text of this level would be suitable for 13-15 year olds (Readable, 2020). The formula for this is also very simple:

\[
\text{FRE} = 206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW})
\]

whereby ASL is the average number of words per sentence and ASW is the average number of syllables per word (Crossley, Allen and McNamara, 2011, p. 90).
Another widely-used formula is the Flesch-Kincaid Grade Level formula. Again, this formula relates to US school grades (Readable, 2020) so the number resulting from this formula, e.g. 6, means that a text would be suitable for pupils in their 6th year of school, that is, when they are around the age of 11-12. The formula for this is:

\[
(0.39 \times \text{number of words/number of sentences}) + (11.8 \times \text{number of syllables/number of words}) - 15.59 \quad \text{(Crossley, Allen and McNamara, 2011, p. 90)}
\]

Such measures are used widely – and successfully – in English as a native or first language studies, yielding correlations up to the low .90s when compared with concurrent measures such as reading test scores or teacher evaluations and using ‘large samples’ of texts (Carrell, 1987, p. 24).

However, when investigating the success of these readability formulae in L2 reading studies, the findings have been mixed. They have been criticised by some for not predicting L2 reading difficulty very well (Brown, 1998; Carrell, 1987) while others have found them to be predictive of reading difficulty (Greenfield, 1999, cited in Crossley, Allen and McNamara, 2011, p. 87). Despite these mixed findings, these readability formulae continue to be widely used, possibly due to ease of access since they are a standard feature in Microsoft Word.

Coh-Metrix is a ‘system for computing computational cohesion and coherence metrics for written and spoken texts’ (Coh-Metrix, 2017) and is capable of analysing texts ‘on over 200 measures of cohesion, language, and readability’ (Graesser et al., 2004, p. 193). It has been used in several studies to analyse text difficulty and writing quality (Aryadoust and Liu, 2015; Crossley et al., 2007; Crossley and McNamara, 2012; Guo, Crossley and McNamara, 2013; Weir, 2014).

The Lexile Framework ‘measures students and texts on the same developmental scale to seamlessly match readers to targeted texts’ (MetaMetrics, 2020). However, until recently, this
was only available for reading although there is now also a listening version (Koons and Van Moere, 2020).

Text Inspector (Text Inspector, 2019) analyses a range of measures including descriptive features such as text and sentence length, number of syllables per word and sentence, and readability measures including Flesch Kincaid Grade Level, Flesch Reading Ease, MTLD as well as providing information regarding the levels of words from various corpora (BNC, COCA), metadiscourse markers and words from the Academic Wordlist (Coxhead, 2000). Although many of the metrics that Text Inspector measures are the same as those measured by Coh-Metrix, Text Inspector differs from Coh-Metrix as it also estimates the CEFR level of a text. Such automated tools can therefore be used to analyse texts and language according to various metrics and can be used to differentiate between levels of proficiency. A number of studies have used automated text analysis tools to identify parameters which can assess the difficulty of reading texts (Dunlea, 2015; Green, Ünaldi and Weir, 2010), listening texts (Révész and Brunfaut, 2013), speaking texts (Crossley, Clevinger and Kim, 2014) and L2 writing (Bax, Nakatsuhara and Waller, 2019; Owen, Shrestha and Bax; forthcoming) as well as integrated tasks (Guo, Crossley and McNamara, 2013; O’Loughlin, 2013).

In a study comparing differences in text difficulty between IELTS reading texts and academic textbooks, Green, Ünaldi and Weir (2010) used expert judges and Coh-Metrix (Graesser et al., 2004) as well as other automated tools, namely, WebVocabProfiler (Cobb, 2003, in Green, Ünaldi and Weir, 2010, p. 197) and Wordsmith Tools (Scott, 2006, in Green, Ünaldi and Weir, 2010, p. 197) to assess the different characteristics of the texts. They found that the IELTS texts did exhibit similar features to those of academic texts although they highlight the fact that ‘the absence of significant differences on many of the measures investigated does not indicate equivalence’ (p. 207). They are also careful to point out that, although the IELTS texts fall within the range of that of the university textbooks, they are not as linguistically demanding as some of the textbooks and the range of values on many of the measures in their study was
indeed broader for the university text books. Nonetheless, the study does provide evidence that Coh-Metrix may be useful in test development to ensure that text characteristics are more consistent.

2.9.1.3 Investigating text difficulty in input texts
Text cohesion has also been found to be an important feature which contributes to text difficulty both in reading (McNamara, Graesser and Louwerse, 2012, p. 89) and in listening, for example, when listening to lectures (Chaudron and Richards, 1986) (see Section 2.4.1). Connectives provide the reader with information regarding how ideas in the text link together and, consequently, increase cohesion (Halliday and Hasan, 1976, cited in McNamara, Graesser and Louwerse, 2012, p. 89) and can therefore help with comprehension (DeCarrico and Nattinger, 1988).

Coh-Metrix measures the occurrence (per 1000 words) of ‘four general types of connectives: causal, additive, temporal, and clarification’ (McNamara, Graesser and Louwerse, 2012, p. 106). The first of these identifies a causal link between two clauses, sentences or parts of a text and is signalled by words such as ‘because’ or ‘so’ (p. 106). The second signals the fact that two parts of a text need to be joined together. These include conjunctions such as ‘also’, ‘further’ and ‘moreover’ (p. 107). Temporal connectives such as ‘before’ or ‘after’ relate to a time-based relationship in the text (p. 107) while clarification connectives signal that an example or an explanation is coming. These include ‘for example’ or ‘in other words’ (p. 107).

McNamara, Graesser and Louwerse (2012) investigated five different sources of reading text difficulty across three genres of texts (narrative, social studies and science) using Coh-Metrix. The sources of text difficulty analysed were:

- ‘word level’ – relating to the difficulty level of words
- ‘sentence level’ – relating to the length of the sentence
• ‘referential cohesion’ – referring to ‘the degree to which there is overlap or repetition of words or concepts across sentences, paragraphs, or the entire text’ (p. 91)

• ‘verb cohesion’, and

• ‘connectives’ (p. 91).

The findings revealed that word knowledge and connectives contributed most to readability in narratives but the converse was true for science texts where difficulties arose at the word level and with regard to a lack of connectives. On the other hand, science text readability was aided by sentence level features, referential cohesion and verb cohesion. The social studies texts were found to be somewhere in the middle, demonstrating high levels of difficulty at the word level and referential cohesion level but high/moderate difficulty at the sentence level and moderate difficulty with regard to both verb cohesion and connectives (p. 111). Based on these findings, the authors recommend that test developers should consider several features of text difficulty including ‘word, sentence and semantic challenges’ (p. 113).

Révész and Brunfaut (2013) investigated the impact of task input factors (including linguistic complexity and text explicitness) and task response characteristics on performance on tasks from Trinity College London’s Graded Examinations in Spoken English listening test. They also investigated test taker perceptions of the difficulty of the tasks. In terms of linguistic complexity, they looked at lexical complexity using WebVocabProfiler (Cobb, n.d., in Révész and Brunfaut, 2013, p. 46) and analysed lexical diversity using the D-formula (Malvern and Richards, 1997, in Révész and Brunfaut, 2013, p. 46) since the texts were too short to use MTLD. Coh-Metrix 2.0 (McNamara, Louwerse, Cai and Graesser, 2005, in Révész and Brunfaut, 2013, p. 46) was used to analyse concreteness of words as well as syntactic and discourse complexity. Specifically, they looked at whether the listening text characteristics predicted task difficulty and found five characteristics to be significant predictors. ‘Incidence of K1 function words’ and ‘lexical density’ each accounted for 40 per cent of the variance while ‘lexical
diversity’, ‘incidence of academic words’ and ‘causal content’ were also significant predictors of difficulty (Révész and Brunfaut, 2013, p. 50). The authors also investigated whether a relationship existed between text characteristics and expected outcome impacted on task difficulty. The findings revealed that ‘the frequency of K1 function words, frequency of formulaic expressions and lexical density’ were also significant predictors of task difficulty (p. 53). With regard to the third question, which compared test takers’ perceptions of text and task difficulty with actual task difficulty, they found strong correlations between the two. Their final research question aimed to elicit the perceived sources of the test task difficulty. Nine participants took part in a stimulated recall for this question and the findings suggest that ‘difficult pronunciation’, ‘speed of delivery’, ‘complex sentence structure’, ‘lack of explicitness’, ‘too many details’ and specific parts of the text caused comprehension issues (p. 54-55). However, ‘lexis’ was considered by all participants to be the major source of text difficulty (p. 54).

Other studies have focussed on predicting the difficulty of items in listening tests (Freedle and Kostin, 1999; Loukina et al., 2016; Rupp, Garcia and Jamieson, 2001).

Freedle and Kostin (1999) investigated whether test takers paid attention to the input texts in TOEFL minitalks. Using a number of text and item characteristics as predictor variables, their results showed that they were able to account for ‘33 per cent of the item difficulty variance’ mainly through variables that were related to the content and structure of the complete passage or a portion of it while item characteristics accounted for very little of the variance. More specifically, they found that ‘rhetorical organisers’, ‘fronted structures’, and ‘vocabulary effect’ all had a significant effect on item difficulty. However, the results of the ‘vocabulary effect’ were unexpected since the higher the number of multisyllabic words in the text, the easier the items were found to be (pp. 16-17). On the other hand, ‘average sentence length’ and ‘passage length’ were not found to be significant for the listening items.
By contrast, Rupp, Garcia and Jamieson (2001) investigated item difficulty using multiple regression analysis and classification and regression tree (CART) in a listening comprehension test. They found that text characteristics (word count, sentence length and type-token ratio) predicted item difficulty as did ‘characteristics of the interaction between the item and the text (directness of information, number of plausible distractors in the text, and item type)’ (p. 211). It is clear from the above studies that text difficulty is affected by text characteristics, particularly lexis, but also text and sentence length play a role so these should be borne in mind when designing input texts for listening and reading comprehension tests. These findings were taken into account in the design of the tasks used in this study by running the texts through Text Inspector to ensure that they were at an appropriate level for the test takers.

2.9.1.4 Text complexity in L2 writing

In addition to analysing input text characteristics, automated tools have been used to assess different aspects of text complexity in learners’ written output (Banerjee, Franceschina and Smith, 2007; O’Loughlin, 2013; O’Loughlin and Wigglesworth, 2003; Weir, 2014). Several of these studies have looked specifically at the use of discourse markers (Bax, Nakatsuhara and Waller, 2019; Owen, Shrestha and Bax, forthcoming). Others have looked at differences in linguistic features produced in independent and integrated tasks (Guo, Crossley and McNamara, 2013; Hessamy and Hamedi, 2013; Kyle and Crossley, 2016).

According to Hyland (2000, p. 109, cited in Hyland and Tse, 2004 p. 157), metadiscourse relates to ‘the linguistic resources used to organize a discourse or the writer’s stance towards either its content or the reader’. Hyland and Tse (2004, p. 169) present a ‘model of metadiscourse in academic texts’ which breaks metadiscourse markers down into two main areas: ‘interactive resources’ and ‘interactional resources’. The former aim ‘to guide [the] reader through the text’ while the latter ‘involve the reader in the argument’.

Interactive resources are further broken down as follows:
• Transitions, e.g. in addition, but

• Frame markers, e.g. finally

• Endophoric markers, e.g. see Figure

• Evidentials, e.g. according to

• Code glosses, e.g. namely

while interactional resources comprise:

• Hedges, e.g. perhaps

• Boosters, e.g. definitely

• Attitude markers, e.g. unfortunately

• Engagement markers, e.g. note that

• Self-mentions, e.g. I, we (Hyland and Tse, 2004, p. 169, abridged)

These categories are slightly different to those presented in Hyland (2004, p. 111, quoted in Bax, Nakatsuhara and Waller, 2019, p. 81), which included the following 13 categories: ‘logical connectives’, ‘frame markers’ (broken down into ‘sequencing’, ‘label stages’, ‘announc[ing] goals’, and ‘topic shifts’), ‘code glosses’, ‘endophoric markers’, ‘evidentials’, ‘attitude markers’, ‘hedges’, ‘relational markers’, ‘person markers’ and ‘emphatics’. Using these 13 categories, Bax, Nakatsuhara and Waller (2019) investigated metadiscourse marker use across student writing at B2, C1 and C2 levels on the CEFR using human raters and Text Inspector (Bax, 2012, cited in Bax, Nakatsuhara and Waller, 2019). Their findings suggest that lower level writers used more metadiscourse markers than students at higher levels; however, higher proficiency candidates produced a significantly wider range of the 13 classes of discourse markers in their analysis. They highlight the importance of analysing both individual items and metadiscourse item classes.
In a study investigating cohesive devices in both independent and integrated tasks from the TOEFL iBT test, Hessamy and Hamedi (2013) found that, in almost all cases, there were significant differences in the use of cohesive devices across the two types of tasks. The integrated tasks included more cohesive devices than the independent writing tasks and both the use of references and lexical cohesion was more successful in the integrated tasks. They also found that anaphoric referencing was more widely used than cataphoric referencing but neither substitution nor ellipsis were widely used in either text type.

Owen, Shrestha and Bax (forthcoming) analysed L2 writing in the Aptis test (British Council, 2020) in terms of metadiscourse and lexis using Text Inspector (Text Inspector, 2019). The results demonstrate that the Aptis writing test discriminates across CEFR bands A0-C. The results of the study reveal that higher-level test takers wrote more than lower-level test takers and demonstrated greater lexical diversity in their texts. They also found that the use of metadiscourse markers was lacking in the CEFR A bands while those achieving a B band on the CEFR ‘rely on metadiscourse to create coherent responses’ (p. 30). Test takers who were awarded a C band score demonstrated textual coherence through a greater variety of lexical resources and thus used fewer metadiscourse markers in their responses. This reduction in the use of metadiscourse markers at the higher bands is in line with other findings (Kennedy and Thorp, 2002, cited in Banerjee, Franceschina and Smith, 2007). However, the most powerful measure in discriminating between CEFR bands was found to be vocabulary.

This finding is in line with Kyle and Crossley (2016) who looked into the impact of lexical sophistication in independent and source-based writing on the TOEFL iBT. They found lexical knowledge to be an important predictor of essay quality. Specifically, their study found that range and bigrams were important in predicting quality in independent tasks while more sophisticated lexis tended to be present in responses to integrated, source-based tasks.

Lexis has been found to be a predictor of level in other studies, too. Guo, Crossley and McNamara (2013) found lexical sophistication to be a significant predictor in both independent
and integrated TOEFL iBT tasks, with higher-level test takers producing more low frequency vocabulary than lower-level test takers while Kim and Crossley (2018) also found that lexical sophistication (when captured as lexical decision mean response time) is more important than syntactic complexity (when captured as mean length of clauses) and cohesion (when captured as lexical overlap between paragraphs) in assessing L2 writing quality. (Kim and Crossley, 2018, p. 51)

These studies demonstrate the important role played by both metadiscourse markers and lexis in academic writing quality. However, while some of these studies address the question of metadiscourse marker use in both independent and integrated TOEFL iBT tasks, they do not address this question with specific reference to listening-into-writing (although it is recognised that the TOEFL iBT involves listening as well as reading in its integrated task). The current study aimed to address this gap by including the question of discourse markers in an IELTS independent essay task and an integrated listening-into-writing EAP test (see Chapter 3).

2.10 Conclusion
In this chapter, I have reviewed the research into listening, writing, independent and integrated test tasks as well as text difficulty. As can be seen from the above, a number of factors can influence test takers’ listening ability. The literature has outlined a number of studies that have investigated listening in terms of performance based on different input media but these have generally only looked at independent listening tests (Batty, 2015; Coniam, 2001; Pardo-Ballester, 2016; Vandergrift, 2004; Wagner, 2007), or, in the case of tests of academic listening, these have generally investigated the impact of note-taking (Carrell, 2007).

With regard to writing, I have given an overview of the processes involved in writing (Coffin et al., 2003; Field, 2004; Flower and Hayes, 1981; Grabe and Kaplan, 1996; Hayes and Flower, 1980; Hyland, 2016; Scardamalia and Bereiter, 1987) and looked at what academic writing tasks in a tertiary environment look like (Hale et al., 1996; Horowitz, 1986). In order to highlight the gaps in the literature, I then looked at how independent, impromptu writing tasks
compare to real-world academic tasks (Moore and Morton, 2005; Turner et al., 2009; Weigle and Friginal, 2015).

With regard to integrated tasks, there is a good deal of research into reading-into-writing test tasks (Gebril and Plakans, 2009; Plakans and Gebril, 2012) but there is only limited research on integrated reading- and listening-into-writing tests, mainly based around the TOEFL test (Campbell, 1990; Cumming et al., 2005a; 2005b; Keck, 2006; Plakans and Gebril, 2013) and there is even less research into listening-into-writing (Cubilo and Winke, 2013; Rukhthong, 2015; Rukhthong and Brunfaut, 2020).

Some research into integrated tasks has focused on source text use (Johns and Mayes, 1990; Plakans and Gebril, 2012; Shi, 2004; 2012; Weigle and Parker, 2012), while other studies have considered differences between independent and integrated writing tasks (Gebril, 2009; Gholami and Alinasab, 2017; Leki and Carson, 1994; Read, 1990). With regard to research comparing performance on independent and integrated tasks, a good deal of work has looked at the TOEFL test (Barkaoui, 2015; Cumming et al., 2005a; 2005b; Lee and Kantor, 2007; Sawaki, Quinlan and Lee, 2013) but research comparing the IELTS Writing Task 2 and integrated tasks seems to be conspicuously lacking.

The few studies which involve only listening-into-writing look at the construct of integrated listening-into-writing (Sawaki, Quinlan and Lee, 2013), cognitive processes (Rukhthong, 2015; Rukhthong and Brunfaut, 2020), the impact of note-taking on performance (Chaudron and Richards, 1986; DeCarrico and Nattinger, 1988) and the impact of different input formats on an integrated listening-into-writing test (Cubilo and Winke, 2013) but such studies are scarce.

Finally, I looked at the research around text difficulty, both in relation to input texts (Green, Ünaldi and Weir, 2010; Révész and Brunfaut, 2013) and L2 writing performance (Hessamy and Hamedi, 2013). This section highlighted the fact that lexis plays a major role in both comprehension of input texts (Stæhr, 2008; 2009; Van Zeeland and Schmitt, 2013) and, along
with discourse markers, the quality of essay writing (Bax, Nakatsuhara and Waller, 2019; Owen, Shrestha and Bax, forthcoming). I outlined the results of several studies comparing the use of discourse markers in TOEFL iBT independent and integrated tasks (Guo, Crossley and McNamara, 2013; Hessamy and Hamedi, 2013; Kim and Crossley, 2018; Kyle and Crossley, 2016) but again noted the lack of research around independent tasks and listening-into-writing EAP tasks specifically.

As a result of this Literature Review, a number of gaps in the literature came to light as well as a number of points to consider in the design of the study.

First of all, the review highlighted the fact that there appears to be a lack of research comparing performance on the IELTS test with performance on an integrated EAP test. As integrated EAP tests appear to replicate the TLU domain better than independent tests of writing such as the IELTS Writing Task 2, the relationship between an integrated EAP test and the IELTS test was worth investigating so this was explored as part of the exploratory study in this PhD (see below and Chapter 3).

Secondly, the research revealed that, although a good deal of research has looked at the effects of different input formats on listening comprehension tests, these have mainly (with the exception of note-taking tasks (Carrell, 2007)) focussed on listening as a discrete skill (Batty, 2015; Leveridge and Yang, 2013; Sydorenko, 2010). As such, these studies have involved test takers responding to multiple choice questions and other types of comprehension check questions such as short-answer and fill-the-gap items (Zhu et al. 2016). However, such limited production is not reflective of the demands of university study, in which test takers are usually expected to produce longer outputs based on oral and written input (Weir, 1983, p. 378, quoted in Lewkowicz, 1994, p. 204) so a more authentic task would involve more production on the part of the test taker. Therefore, test takers in my study were required to produce a summary text as part of the integrated task in both the exploratory (see Chapter 3) and the main study (see Chapter 4).
The only study which appears to address the impact of input format on an integrated EAP listening-into-writing test is Cubilo and Winke (2013). However, although this study employs both audio and video, the latter showed the speaker and ‘a projected picture of material related to the topic’ (p. 378) rather than a full, authentic Power Point presentation while the audio text was accompanied by a context-related still photo, showing ‘a snapshot from the video’ (p. 379), which may have had an impact on the results. Considering that there is already a large amount of research into input formats in discrete listening tests (Aldera, 2015; Ginther, 2002; Montero Perez, Peter and Desmet, 2014) and given the lack of research into the impact of different inputs formats on an integrated EAP listening-into-writing task, this is another gap in the literature.

Cubilo and Winke’s (2013) study compares the content of the notes made with performance on the written task but does not look at how and to what extent test takers incorporated language from the source texts. Therefore, this is an area which merits investigation.

Furthermore, although some work has been done on the discourse features and text features which affect text difficulty (Bax, Nakatsuhara and Waller, 2019; Guo, Crossley and McNamara, 2013; O’Loughlin, 2013; Owen, Shrestha and Bax, forthcoming; Révéz and Brunfaut, 2013), there does not seem to be any study that combines discourse features with different input formats in an integrated listening-into-writing EAP test. Consequently, this is another area which could usefully be examined. Therefore, in my study, I addressed this gap in the literature by comparing the discourse and text features (descriptive analyses, lexical diversity, syntactic pattern density and readability (see Section 3.3.2) produced in independent IELTS essays with the features produced in an integrated EAP test in the exploratory study (see Chapter 3). In the main study (see Chapter 4), I analysed the discourse and text features produced as a result of the different input formats.

When looking at input formats, several studies have looked at test taker perceptions of their performance and their preferences regarding audio, audiovisual and sometimes multimedia
formats (Cubilo and Winke, 2013; Ockey, 2007; Pardo-Ballester, 2016; Rukhthong, 2015; Suvorov, 2008). However, to date, the findings appear to be inconclusive. Therefore, it is worth including test taker perceptions as part of the study to see to what extent these can inform our understanding of how to enable test takers to perform to the best of their ability.

As there were several potential areas of interest, I first carried out an exploratory study (see Chapter 3) to look at several different areas of interest with a view to narrowing down on a particular area for the main study (see Chapter 4).

The exploratory study for this thesis first compared performance on the independent IELTS listening and writing tasks with performance on an EAP integrated listening-into-writing task. Then I investigated the impact of different input formats on performance on the written output from the integrated task and analysed test takers’ source text use. Finally, data regarding their input format preferences was gathered.

In order to address the gaps in the listening-into-writing literature, there was a choice to be made regarding the input that provided and the output task that test takers would complete. As many studies had already compared audio and audiovisual input, using context and content stills and video as well as multimedia, I compared audio only (without any accompanying still photos) to video input which was presented in a number of formats. These were video of the speaker only (as in Cubilo and Winke’s (2013) study), and the speaker plus a Power Point with both textual and diagrammatic input to investigate the impact of the four different types of audiovisual input (see Chapter 4 for a more detailed explanation).

With regard to the output task, the literature review found that summaries or essay response tasks were the most frequently used tasks in integrated tasks (Ascención Delaney, 2008). Both of these are worthwhile tasks for an integrated EAP test. However, the essay response task has been found to be more challenging for lower level learners and to involve more critical thinking than a summary task (Ascención Delaney, 2008). On the other hand, summary writing
is considered an essential skill in a tertiary environment (Kirkland and Saunders, 1991; O’Kane, 2017). As this study also involved lower level learners, I decided to use the summary task for the written output on the integrated test task.

To conclude, this Literature Review highlighted several gaps in the research literature regarding integrated listening-into-writing tests. Based on these gaps, the following questions were investigated in the exploratory study (see Section 3.3):

1. Is there a relationship between listening ability (measured by the IELTS listening test) and performance on:
   a) an independent writing test?
   b) an integrated EAP listening-into-writing test?
2. To what extent does performance on an independent EAP writing task (IELTS Academic Writing Task 2) predict performance on an integrated EAP listening-into-writing test?
3. How does the input format of an academic lecture affect performance on an integrated EAP listening-into-writing task?
4. How do learners’ perceptions of their performance compare with observed performance on an integrated EAP listening-into-writing task?

The findings from the exploratory study and the gaps highlighted in the literature review then led to the following questions being addressed in the main study (see Section 4.3.1):

1. How does performance on an integrated listening-into-writing EAP task differ when test takers are presented with lecture input in the form of a podcast or a vodcast?
2. To what extent are test takers influenced by the written word in an EAP lecture compared to the spoken word?
3. How do test takers’ perceptions of their performance on an integrated EAP listening-into-writing test presented in two different input formats compare to their observed scores as awarded by trained raters?
In the next chapter, the methodology for the exploratory study will be outlined and the results presented and discussed.
3. Exploratory study

3.1 Introduction
This chapter will discuss the research questions, methodology, results and discussion for the exploratory study. The changes made ahead of the main study will then be discussed along with the justifications for these changes.

3.2 Research approach
In line with Dörnyei’s (2007, p. 173) typology for mixed methods research, the current study took a ‘QUAN + qual’ approach. This means that both quantitative and qualitative data were collected concurrently (Dörnyei, 2007, p. 172). However, the quantitative approach is the dominant one while the qualitative analysis is used to triangulate the quantitative findings.

The mixed methods approach allows ideas from a Positivist and Interpretivist approach (Thomas, 2013, p. 111) to be used. A Positivist approach involves looking at ‘things that [could] be quantified and counted’ and analysing ‘variables, decided on in advance …’ while an Interpretivist approach involves analysing ‘perceptions, feelings, ideas, thoughts as … observed’ and ‘emergent patterns’.

Opinions about a mixed methods approach are varied among researchers. While some ‘traditionalists’ argue that the qualitative and quantitative paradigms are ‘at cross purposes’, others argue that the two approaches are indeed ‘compatible’ and the ‘methodological choice should always be based on what is useful in answering your questions, regardless of any philosophical or paradigmatic assumptions’ (O’Leary, 2014, p. 147).

Mackey and Gass (2005, p. 164) point out that the combination of qualitative and quantitative methods should not be seen as two opposing ends of a dichotomous continuum but rather as ‘complementary means of investigating the complex phenomena at work in second language acquisition’. Indeed, one advantage of mixed methods research is that it allows the researcher to get the best of both worlds by maximising the strengths of both methods while at the same time reducing some of the weaknesses of each one (Dörnyei, 2007, p. 45-47). This is the
primary reason for adopting a mixed methods approach in the current study. Using purely quantitative methods would be useful in highlighting significant (or not) differences between the groups in the study and the input formats used. However, it would mean that the deeper analysis of investigating how and in what ways the data varied would be missing. As such, the reader would only be getting a part of the picture. On the other hand, using only qualitative analyses would not enable any inferences to be drawn as the research would not highlight whether there were any statistically significant results from the study.

Furthermore, mixed methods research can provide additional evidence to improve validity by triangulating data obtained from the different types of research (Dörnyei, 2007, p. 45-47). Triangulation is a way of looking at a research question through different lenses. By using several data sources, a research question can be investigated from several perspectives (Dörnyei, 2007, p. 165). This was one of the major reasons why ‘mixing qualitative and quantitative methods emerged and evolved in social research’ (Hashemi and Babaii, 2013, p. 829).

In the exploratory study, the quantitative analysis gave the overall picture but the qualitative analyses were triangulated with the quantitative analyses to add further detail and provide examples to support the quantitative analysis. The qualitative approach can reveal the causes of the results of the quantitative analysis and provide evidence to support the results. On the other hand, the quantitative approach can reveal more about how generalisable the results are to the wider EAP environment.

3.3 Exploratory study
The Literature Review highlighted a number of gaps in the research around integrated EAP listening-into-writing tasks. Therefore, before carrying out the main study, an initial study was conducted to investigate those gaps and highlight areas for further study. As such, the exploratory study was intentionally complex to investigate several potential areas of research with the intention of revealing one to pursue in more detail for the main study. At the same
time, the exploratory study served to investigate the usefulness of the research instruments and to identify problematic aspects of the experimental design and the tools used.

In the exploratory study, there were four research questions:

1. Is there a relationship between listening ability (measured by the IELTS listening test) and performance on:
   a) an independent writing test?
   b) an integrated EAP listening-into-writing test?

2. To what extent does performance on an independent EAP writing task (IELTS Academic Writing Task 2) predict performance on an integrated EAP listening-into-writing test?

3. How does the input format of an academic lecture affect performance on an integrated EAP listening-into-writing task?

4. How do learners’ perceptions of their performance compare with observed performance on an integrated EAP listening-into-writing task?

It was expected that there would not be any significant differences between performance on the IELTS academic listening test and the integrated EAP test because the participants in the trial were already studying at university and, to enter university, they needed to have a minimum of IELTS 4.5 in each of the skills. Consequently, although some test takers may have a ‘spiked’ profile across the skills, it was expected that most would have a fairly similar profile across all skills, resulting in a significant correlation between the IELTS listening test and the IELTS writing test. Similarly, it was expected that there would be no significant difference between the IELTS academic writing test scores and the integrated EAP scores as both of these are productive skills tests and engage many of the same cognitive processes required for writing such as organisation, planning and translation.

With regard to performance on the integrated task, it was anticipated that lower-level students would reproduce fewer propositions than stronger students but mid-range students would
have the largest number of identical propositional matches because they were likely to understand more than the weaker students (and therefore reproduce more key points) but would be less able than the stronger students to paraphrase the content of the lecture. Therefore, the assumption was that the stronger students would paraphrase more, thereby better demonstrating their ability to not only understand but also reproduce the content using their own voice.

With regard to performance across test formats, it was envisaged that students would perform best when presented with textual and aural information (in line with the findings of Locke, 1977).

Finally, the expectation relating to students’ feedback questionnaires was that the higher-level candidates would be able to evaluate their performance accurately while the lower level candidates were likely to have a less well-defined understanding of their performance. This expectation was based on the study of Kruger and Dunning (1999, p. 1121). They found that study participants whose scores on ‘tests of humor, grammar and logic’ resulted in their being in the lowest quartile (12th percentile) actually rated themselves as being in the 62nd percentile. However, as their skills increased, so did their metacognitive competence and this enabled them to better understand the ‘limitations of their abilities’.

3.3.1 Participants
Twenty students – ten from each of two UK universities - took part in the exploratory study. 14 participants were female and 6 were male. They were aged 18-40 with one participant over 40, and were of a number of nationalities: Omani, Spanish, Taiwanese, Chinese, Japanese, Romanian, Iraqi and Thai.

Despite the fact that the participants should have had an IELTS level of at least 4.5, they had been rated by their lecturers as having an overall IELTS level of between 3.5 and 6.5 and were studying on either a Foundation course or a Pre-Master’s course.
All participants were given an information sheet about the study (see Appendix 1) and informed consent was sought from all participants (see Appendix 1). All data was anonymised by using a coding system for identification. Table 3.1 below shows the breakdown of nationality, age and gender.

Table 3.1: Breakdown of participants in the exploratory study

<table>
<thead>
<tr>
<th></th>
<th>Nationality</th>
<th>Age</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>B01</td>
<td>Omani</td>
<td>26-30</td>
<td>Male</td>
</tr>
<tr>
<td>B02</td>
<td>Omani</td>
<td>22-25</td>
<td>Male</td>
</tr>
<tr>
<td>B04</td>
<td>Spanish</td>
<td>22-25</td>
<td>Female</td>
</tr>
<tr>
<td>B05</td>
<td>Chinese</td>
<td>22-25</td>
<td>Male</td>
</tr>
<tr>
<td>B06</td>
<td>Taiwan</td>
<td>22-25</td>
<td>Male</td>
</tr>
<tr>
<td>B07</td>
<td>Japanese</td>
<td>31-40</td>
<td>Female</td>
</tr>
<tr>
<td>B08</td>
<td>Japanese</td>
<td>31-40</td>
<td>Female</td>
</tr>
<tr>
<td>B09</td>
<td>Taiwan</td>
<td>18-21</td>
<td>Female</td>
</tr>
<tr>
<td>B10</td>
<td>Omani</td>
<td>26-30</td>
<td>Male</td>
</tr>
<tr>
<td>B11</td>
<td>Omani</td>
<td>26-30</td>
<td>Male</td>
</tr>
<tr>
<td>R01</td>
<td>Thai</td>
<td>31-40</td>
<td>Female</td>
</tr>
<tr>
<td>R02</td>
<td>Thai</td>
<td>31-40</td>
<td>Female</td>
</tr>
<tr>
<td>R03</td>
<td>Taiwan</td>
<td>26-30</td>
<td>Female</td>
</tr>
<tr>
<td>R04</td>
<td>Taiwan</td>
<td>22-25</td>
<td>Male</td>
</tr>
<tr>
<td>R05</td>
<td>Thai</td>
<td>31-40</td>
<td>Female</td>
</tr>
<tr>
<td>R06</td>
<td>Romanian</td>
<td>26-30</td>
<td>Female</td>
</tr>
<tr>
<td>R07</td>
<td>Iraqi</td>
<td>&gt;40</td>
<td>Female</td>
</tr>
<tr>
<td>R08</td>
<td>Chinese</td>
<td>26-30</td>
<td>Female</td>
</tr>
<tr>
<td>R12</td>
<td>Thai</td>
<td>31-40</td>
<td>Female</td>
</tr>
<tr>
<td>R13</td>
<td>Thai</td>
<td>26-30</td>
<td>Female</td>
</tr>
</tbody>
</table>

3.3.2 Procedure

There were 4 steps in the exploratory study. The following section describes the procedures for each.
1. **Independent IELTS academic listening test (25 mins)**

Test takers took IELTS Listening Test 3, a 40-item test from the *Cambridge IELTS 3, with answers* book of past IELTS papers (UCLES, 2002). The time allocated for this test was 25 mins and it was marked in accordance with the marking key provided in the book (see Appendix 2). An academic listening test was considered appropriate as it would reflect the ‘target language use (TLU) domain’ (Bachman and Palmer, 1996: 18). This was done to provide a measure of their discrete listening ability to answer RQ1. The results from this test were correlated with the results of the IELTS writing test and the integrated EAP test to answer RQ1.

2. **Independent IELTS academic writing test (40 mins)**

Test takers wrote a Task 2 essay from an IELTS past paper to provide a measure of their writing ability, which, in turn, would contribute towards answering RQ1a. The essay was Academic Writing Task 2 from the IELTS Scores Explained Updated 2009 DVD (IELTS, 2009) (see Appendix 3):

> 'Children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents. To what extent do you agree or disagree with this opinion?'

This essay was chosen as it is an IELTS Academic Writing Task and, being from the DVD, it had sample papers, marks and examiner comments, all of which could be used for rater training before rating the participants’ essays. The DVD also provided the public IELTS descriptors so these were used as the rating criteria.

The essays were independently rated by two raters who had undergone familiarisation with the IELTS descriptors using the public descriptors and the sample essays on the IELTS Scores Explained DVD (IELTS, 2009). In the familiarisation session, raters were asked to reconstruct the cut-up scales for each
of the four writing criteria (Task response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy). After that, they marked the three sample papers for IELTS Academic Writing Task 2 from the DVD and compared their answers to each other and to the examiner comments on the DVD.

After this, they each rated all of the IELTS tasks independently. Since there were not many test takers and it was important that the raters shared a joint understanding of the rating scale because they needed to use the same scale for the EAP essay, they compared scores and discussed any differences before agreeing on a composite score for each essay.

The composite scores from this test were correlated with the results of the IELTS listening test (RQ1a) and the integrated EAP test (RQ2).

Both essays were also analysed quantitatively and qualitatively to investigate RQ2 further. First of all, a quantitative analysis of the number of cohesive devices used in the two types of essays was carried out. Coh-Metrix, a free text analysis software, which can be used to measure a wide range of textual features (Graesser et al., 2004; McNamara et al., 2014), was also used to analyse the essays quantitatively. 13 analyses were selected for analysis. These were broken down as follows:

*Descriptive analyses*

- DESPC: Paragraph count, number of paragraphs
- DESSC: Sentence count, number of sentences
- DESWC: Word count, number of words
- DESPL: Paragraph length, number of sentences in a paragraph, mean
- DESSL: Sentence length, number of words, mean
- DESWLSy: Word length, number of syllables, mean
- DESWLLit: Word length, number of letters, mean
Lexical diversity

- LDTTRa: Lexical diversity, type-token ratio, all words

Syntactic pattern density

- DRNP: Noun phrase density, incidence
- DRVP: Verb phrase density, incidence
- DRPVAL: Agentless passive voice density, incidence

Readability

- RDFRE: Flesch Reading Ease
- RDFKGL: Flesch-Kincaid Grade Level

The means for these analyses were calculated for both the IELTS Writing task and the EAP Writing task to measure the extent to which these reveal a relationship between the two types of essays. Coh-Metrix results for the IELTS essay were correlated with the Coh-Metrix results for the EAP essay. The qualitative analysis compared which types of cohesive devices were used in the IELTS Writing and the EAP essay.

3. Integrated EAP listening-into-writing test (75 mins) (see Appendix 4)

In this part, test takers listened to a short EAP lecture (approx. 12 mins long) (see Appendix 5) on the topic of culture shock, which the researcher developed for the purposes of the research. The lecture was divided into four sections and each section was presented in a different mode:

- **Kinaesthetic and aural** – test takers saw the lecturer, thus enabling them to make use of the lecturer’s body language, gestures and facial expressions;
- **Aural only** – the test takers heard only the audio track;
- **Visual (diagram) and aural** – the test takers heard the lecturer and saw only a PowerPoint slide with a diagram of the aspects being spoken about;
d. Visual (textual) and aural – test takers heard the lecturer but saw a PowerPoint slide with textual information in the form of key words relating to the aural input.

James (1986, p. 41 in Long, 1990, p. 68-69) found that texts which are longer than three minutes can lead to sensory overload unless the test takers find the topic interesting. Bearing this in mind, each section was kept to between 3 and 4 mins in length (see Table 3.2 below):

Table 3.2: Word length and time for input sections

<table>
<thead>
<tr>
<th>Part no.</th>
<th>No. of words</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>406</td>
<td>3:14</td>
</tr>
<tr>
<td>2</td>
<td>430</td>
<td>3:45</td>
</tr>
<tr>
<td>3</td>
<td>562</td>
<td>4:02</td>
</tr>
<tr>
<td>4</td>
<td>555</td>
<td>3:54</td>
</tr>
</tbody>
</table>

In line with recommendations from the literature regarding the number of times a listening input should be played (Buck, 1990; 2001; Pisoni, 1997, in Field, 2015), the lecture was played twice.

The task for the test takers was twofold:

a. in the first part, test takers were required to take notes on the short lecture. Participants were given one minute to read the instructions for each section and had 30 seconds to check their answers before the second playback of each section.

b. In the second part, they were asked to write a 500-word essay on the topic of the lecture, which was about culture shock. Test takers were given the title of the essay at the beginning of the test to help them in knowing which key points to listen out for:

‘What is culture shock and how can we reduce the effects of culture shock?’
Since there are a number of potential issues with asking preparatory course students (as in the case of the foundation or pre-Master’s students in this exploratory study) to take notes, including the fact that the participants may not yet be adept at taking notes and / or spotting signposting phrases in lectures (DeCarrico and Nattinger, 1988), scaffolding was provided to all test takers in the form of ‘guided’ notes pages in the test booklet which guided them through the content of the lecture (see Appendix 4). This support also served to ‘level the playing field’, so that everyone had access to the same level of input regarding what to take notes on. However, the notes were merely there to assist with writing the essay and were not graded due to the issues outlined in the Literature Review concerning the difficulty of marking notes (Chaudron, Loschky and Cook, 1994; Hayati and Jalilifar, 2009, p. 102). Nonetheless, they were collected in as it was recognised that they may shed some light on areas of difficulty with the writing task.

In addition to the ‘guided’ notes pages, the participants were provided with an outline for the essay which followed the contents of the notes pages. The note-taking part lasted 30 minutes, after which 45 minutes were allowed for writing the essay. Although the length requirement of the essay was longer than the IELTS essay, it was anticipated that students would be able to write more as they would have been provided with a great deal of scaffolding through the guided notes pages and they would have all the content from the input lecture so would not need to spend time generating ideas, planning and organising as would be the case for the independent IELTS essay. On the other hand, the timing was similar to the 40 minutes recommended for IELTS Task 2, so would facilitate comparisons between the two.
The topic of culture shock was chosen to avoid the issue of topic bias (Lee and Anderson, 2007, p. 308). Culture shock is a general academic topic which would be relevant to all students who are currently studying or intend to study abroad and, in fact, it is likely that the test takers in the exploratory project would all be experiencing culture shock to some extent. Therefore, the topic would be both useful and educational (and in turn, interesting) to all participants without unduly advantaging or disadvantaging students from one particular discipline.

These essays were also independently rated by the same two raters who had undergone familiarisation with the IELTS descriptors using the public descriptors and the sample essays on the IELTS Scores Explained DVD (IELTS, 2009) because the public descriptors would again be used to mark these essays. After marking the essays, the raters compared their results and agreed on composite marks for all the essays.

The scores for this test were correlated with both the IELTS listening task and the IELTS writing task using SPSS (IBM Corp., 2013) to answer RQ1b and RQ2 respectively. To answer RQ2, the essays were also analysed qualitatively and quantitatively. The quantitative analysis focused on the number and type of cohesive devices used in the two types of essays and the percentage of propositions that were reproduced from the four sections of the lecture. Again, Coh-Metrix (Graesser et al., 2004; McNamara et al., 2014) was also used to analyse the essays quantitatively. The same 13 analyses selected for analysis of the IELTS essays were used for the EAP essays to allow ease of comparison. These were broken down as follows:

Descriptive analyses

- DESPC: Paragraph count, number of paragraphs
- DESSC: Sentence count, number of sentences
As mentioned in task 2 Independent IELTS academic writing test, above, means were calculated for the Coh-Metrix results for comparison with the IELTS essay. After that, the results for the IELTS essay were correlated with the results for the EAP essay to examine whether there was any relationship between the two essay types for these analyses.

The qualitative analysis for RQ2 compared which types of cohesive devices were used in the IELTS Writing and the EAP essay.

To answer RQ3, the EAP essays were again analysed quantitatively and qualitatively. The quantitative analysis focused on the number of words written and the percentage of propositions reproduced in each of the four sections. The qualitative analysis investigated which propositions were reproduced from the four sections of the EAP essay.
Finally, to answer RQ4, the sections in which the test takers reproduced the highest and lowest percentage of propositions were identified. These sections were correlated with the learners’ perceptions of their best and worst performances according to the questionnaire results.

4. Post-task feedback questionnaire

After the test, all participants completed a short feedback questionnaire focussing on their preferences and their perceptions about their performance in the different tasks.

The questionnaire results served, firstly, to collect personal details such as nationality, age and gender. The data from these questions are presented in Section 3.3.1 above. Other questions were used to investigate RQ4 (see Section 3.4.4). These questions asked participants to state which sections of the EAP writing task they felt they performed best and worst on and why. Test takers’ perceptions of their performance on the EAP essay were correlated with their observed best and worst performances as measured by the percentage of propositions reproduced. The remaining questions served primarily as validation evidence for the instruments used. The information collected provided feedback regarding the suitability of the tools for the study and highlighted any issues which needed to be addressed. The relevant findings from the questionnaire will be presented in Section 3.4.5 below.

Descriptive statistics (mean, standard deviation, median and mode) were calculated using SPSS (IBM Corp., 2013) and the findings were analysed qualitatively and quantitatively to investigate test takers’ perceptions regarding:

- their performance on the IELTS listening test as a reliable representation of their ability
- the difficulty of the IELTS writing test
their previous knowledge of the lecture topic and whether this impacted on their performance
whether there were any construct-irrelevant factors which affected their performance on the EAP task
the difficulty of the EAP task
any other factors which affected how much they could write in the essay.

3.4 Exploratory study results and discussion

3.4.1 RQ1: Is there a relationship between listening ability and performance on an independent writing test and an integrated EAP listening-into-writing test?
20 participants took the three tests: the IELTS listening test, the IELTS writing test and the integrated EAP listening-into-writing test. Descriptive statistics are shown below in Table 3.3.

As can be seen from the table, the scores for the IELTS listening test ranged from 8 to 29 out of 40 with a mean of 16.5 and a standard deviation of 5.96. The median score was 15 while the mode was 19.

Table 3.3: Descriptive statistics for test tasks

<table>
<thead>
<tr>
<th></th>
<th>IELTS Listening score</th>
<th>IELTS Writing score</th>
<th>EAP score</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Valid</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Mean</td>
<td>16.50</td>
<td>5.200</td>
<td>5.500</td>
</tr>
<tr>
<td>Std. Error of Mean</td>
<td>1.333</td>
<td>.1469</td>
<td>.1308</td>
</tr>
<tr>
<td>Median</td>
<td>15.00</td>
<td>5.250</td>
<td>5.500</td>
</tr>
<tr>
<td>Mode</td>
<td>19</td>
<td>5.5</td>
<td>5.5</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>5.960</td>
<td>.6569</td>
<td>.5849</td>
</tr>
<tr>
<td>Variance</td>
<td>35.526</td>
<td>.432</td>
<td>.342</td>
</tr>
<tr>
<td>Skewness</td>
<td>.324</td>
<td>.087</td>
<td>-.438</td>
</tr>
<tr>
<td>Std. Error of Skewness</td>
<td>.512</td>
<td>.512</td>
<td>.512</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-.679</td>
<td>-.664</td>
<td>1.167</td>
</tr>
<tr>
<td>Std. Error of Kurtosis</td>
<td>.992</td>
<td>.992</td>
<td>.992</td>
</tr>
<tr>
<td>Range</td>
<td>21</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Minimum</td>
<td>8</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>Maximum</td>
<td>29</td>
<td>6.5</td>
<td>6.5</td>
</tr>
</tbody>
</table>
The scores for the IELTS writing test ranged from 4.0 to 6.5 on a scale of 0-9. The mean score was 5.2 and the standard deviation was .657. The median score was 5.25 and the mode was 5.5.

The scores for the integrated EAP test also ranged between 4.0 and 6.5 on a scale of 0-9, with a mean score of 5.5 and a standard deviation of .585. The median score was 5.5 and the mode was 5.5.

To assess whether there is a relationship between performance on the IELTS listening test and the IELTS writing test on the one hand, and between the IELTS listening test and the integrated EAP listening-into-writing test on the other, the scores on the listening test were correlated with the composite (agreed) scores on the IELTS writing test and composite (agreed) scores on the integrated EAP tests respectively.

Before calculating the correlation coefficients, tests of normality were carried out. The tests of normality were not met so non-parametric tests, calculated using SPSS, are reported here. The results for the Spearman Rank Order correlation can be seen in Table 3.4.

Table 3.4: Spearman Rank Order results for test tasks

<table>
<thead>
<tr>
<th></th>
<th>IELTS Listening score</th>
<th>IELTS Writing score</th>
<th>EAP score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spearman’s rho</td>
<td>IELTS Listening score Correlation Coefficient</td>
<td>1.000</td>
<td>.353</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.</td>
<td>.127</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>IELTS Writing score Correlation Coefficient</td>
<td>.353</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.127</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>EAP score Correlation Coefficient</td>
<td>.318</td>
<td>.396</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.172</td>
<td>.084</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>
The results show that there was a non-significant correlation ($r_s = .353, p>0.05$) between performance on the IELTS listening test and the IELTS writing test. The correlation between the IELTS listening test and the integrated EAP test was also non-significant using the Spearman rank correlation ($r_s = .318, p>0.05$ (2-tailed)).

Field (2009, p. 181) recommends using Kendall’s tau with small data sets which have a large number of tied ranks so this correlation was also computed (see Table 3.5 below).

The results of this analysis are also non-significant. Kendall’s tau for the IELTS listening and IELTS writing test is $\tau b = .255$, $p>0.05$ (2-tailed) while for the IELTS listening and integrated EAP test, Kendall’s tau is $\tau b = .271$, $p>0.05$ (2-tailed).

While neither of the non-parametric tests reveal any significant correlations, the Spearman’s rho correlation coefficient is stronger for IELTS Listening and IELTS Writing than it is for the IELTS Listening and the integrated EAP test whereas Kendall’s tau is stronger for the IELTS Listening and the integrated EAP test. Field (2009, p. 182) states that Kendall’s tau is ‘a more accurate gauge’ of the correlation for the population so it is wise to interpret this result.

Table 3.5: Kendall’s tau results for test tasks

<table>
<thead>
<tr>
<th></th>
<th>IELTS Listening score</th>
<th>IELTS Writing score</th>
<th>EAP score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kendall’s tau b</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IELTS Listening score</td>
<td>Correlation Coefficient</td>
<td>1.000</td>
<td>.255</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.</td>
<td>.148</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>IELTS Writing score</td>
<td>Correlation Coefficient</td>
<td>.255</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.148</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>EAP score</td>
<td>Correlation Coefficient</td>
<td>.271</td>
<td>.313</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.132</td>
<td>.099</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>
On the basis of these results then, we can conclude that there is no significant correlation between the IELTS Listening test and the IELTS Writing and integrated EAP tests respectively. However, this could be due to the small sample size and a larger-scale study may have revealed different results.

It is not surprising that the correlation, albeit non-significant, between the IELTS Listening test and the integrated EAP test is slightly stronger than the correlation between the IELTS Listening and the IELTS Writing test since listening ability plays a direct role in the integrated EAP test but listening ability may not be so directly linked to writing so this is not the case for the IELTS Writing test. However, one reason for the fact that the correlation is not significant could be that some of the weaker participants may have struggled to understand the input. Plakans and Gebril (2012, p. 26) suggest that, in their reading-into-writing study, the lower candidates could understand the text but the stronger candidates ‘may have had an advantage of understanding the topic more completely’. Another reason may be that the test takers may not yet have been very good at taking notes (Chaudron, Loschky and Cook, 1994).

On the other hand, it is perhaps not surprising that there is not a significant correlation between the listening test and the other two tests as these latter tests are productive skills test whereas the listening test is a receptive skills test and thus does not require the test taker to produce very much language. Instead, in the IELTS listening test, the test taker has to engage a number of cognitive processes such as phoneme, syllable and word decoding, syntactic parsing or meaning construction (Field, 2008, p. 114) but their production of the language is limited to selecting an answer or producing a short, constructed response up to a maximum of three words in length. Writing, on the other hand, is a more difficult skill since the test taker has so many more cognitive processes to deal with in creating a piece of writing such as planning, organising, translating and reviewing (Shaw and Weir, 2007).
3.4.2 RQ2: To what extent does output performance on an independent EAP writing task predict output performance on an integrated EAP listening-into-writing test?

3.4.2.1 Quantitative data

To investigate the extent to which performance on the IELTS writing test can predict performance on the EAP writing test, correlations between the scores on the two tests were calculated using SPSS. Next, the number of cohesive devices reproduced across the two test types was analysed. Finally, the IELTS essays and the integrated EAP essays were analysed using Coh-Metrix. The means of the different analyses were calculated to see to what extent there was a correlation between the textual features of the two types of essays.

As discussed in the results for RQ1 above, the data were not normally distributed so non-parametric test results (Spearman Rank Order correlation coefficients) are presented. Since just a small sample of participants with a large number of tied ranks was involved, Kendall’s tau was also calculated. These correlation coefficients were calculated as part of the analyses for RQ1 above so the results can be found in Tables 3.4 and 3.5 above.

As can be seen from Table 3.4, the Spearman Rank Order correlation coefficient for the IELTS Writing and the integrated EAP test was non-significant: $r_s = .396, p>0.05$ (two-tailed). Kendall’s tau, which can be used on small samples (Field, 2009, p. 182) was also non-significant: $\tau_b = .313, p>0.05$ (2-tailed). Nonetheless, both the Spearman rank correlation and Kendall’s tau correlation coefficients were stronger between the IELTS Writing test and the integrated EAP test than between the two IELTS tests or between the IELTS Listening and EAP tests.

The next part of the results will focus on the total number of cohesive devices used in the two types of essays.

In total, 355 cohesive devices were used across the twenty IELTS essays whereas a total of 719 cohesive devices were used in the EAP essays (see Appendices 6 – IELTS Cohesive devices and 7 – EAP cohesive devices).
The mean number of cohesive devices across the 20 essays was 17.75 and the standard deviation was 8.465. The median was 17.5 and the mode was 18. The lowest number of cohesive devices for the IELTS essays was in essay R3 (n=1); however, this student had to leave after 15 minutes so the essays with the next lowest number of cohesive devices were students B1, B8 and R3, who all used 11 cohesive devices each. The highest number of cohesive devices was found in essay R8 (n=35).

For the EAP essay, the mean number of cohesive devices used in the 20 essays was 35.95 and the standard deviation was 10.308. The median was 34 and there were two modes: 30 and 32. The lowest number of cohesive devices was in essay B1 (n=15) and the highest number was in essay B2 (n=53).

Tests of normality were carried out on the number of cohesive devices on the two types of essay but the assumptions were not met due to one outlier; therefore, non-parametric test results are reported here. Table 3.6 shows the results of these tests.

As can be seen, there was a statistically significant correlation between the number of cohesive devices on the IELTS tests and the integrated EAP tests: \( r_s = .463, p<0.05 \) (two tailed). Due to the small sample size and the fact that there was more than one mode, Kendall’s tau was also calculated. As expected, this was slightly weaker but was still statistically significant: \( \tau_b = .356, p<0.05 \) (2-tailed). Cohen (1988: 79-81) describes correlations between \( r = .30 \) to \( .49 \) as medium strong. As such, both of these demonstrate a medium strength relationship (Cohen, 1988: 79-81) between the two tests in terms of the cohesive devices used.
Table 3.6: Non-parametric test results for cohesive devices

<table>
<thead>
<tr>
<th>Kendall's tau_b</th>
<th>IELTS - total no. of cohesive devices</th>
<th>EAP - total no. of cohesive devices</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correlation Coefficient</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>20</td>
</tr>
<tr>
<td>EAP - total no. of cohesive devices</td>
<td>Correlation Coefficient</td>
<td>.356*</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.033</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>20</td>
</tr>
<tr>
<td>Spearman's rho</td>
<td>IELTS - total no. of cohesive devices</td>
<td>Correlation Coefficient</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>20</td>
</tr>
<tr>
<td>EAP - total no. of cohesive devices</td>
<td>Correlation Coefficient</td>
<td>.463*</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.040</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>20</td>
</tr>
</tbody>
</table>

* Correlation is significant at the 0.05 level (2-tailed).

Next, the Coh-Metrix analyses were carried out. As outlined in Section 3.3.2, the analyses were broken down into 4 sections: descriptive analyses, lexical diversity, syntactic pattern density and readability.

The descriptive analyses were:

- DESPC: Paragraph count, number of paragraphs
- DESSC: Sentence count, number of sentences
- DESWC: Word count, number of words
- DESPL: Paragraph length, number of sentences in a paragraph, mean
- DESSL: Sentence length, number of words, mean
- DESWLSy: Word length, number of syllables, mean
- DESWLIlt: Word length, number of letters, mean

Table 3.7 below shows the descriptive statistics for the 7 descriptive analyses outlined above.
Table 3.7: Descriptive statistics for Coh-Metrix analyses

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>IELTS_DESPC</td>
<td>20</td>
<td>6.00</td>
<td>1.00</td>
<td>7.00</td>
<td>3.600</td>
<td>1.35336</td>
</tr>
<tr>
<td>EAP_DESPC</td>
<td>20</td>
<td>36.00</td>
<td>3.00</td>
<td>39.00</td>
<td>10.650</td>
<td>10.71779</td>
</tr>
<tr>
<td>IELTS_DESSC</td>
<td>20</td>
<td>21.00</td>
<td>3.00</td>
<td>24.00</td>
<td>14.350</td>
<td>5.51815</td>
</tr>
<tr>
<td>EAP_DESSC</td>
<td>20</td>
<td>52.00</td>
<td>9.00</td>
<td>61.00</td>
<td>33.850</td>
<td>15.08319</td>
</tr>
<tr>
<td>IELTS_DESWC</td>
<td>20</td>
<td>378.00</td>
<td>56.00</td>
<td>434.00</td>
<td>264.000</td>
<td>98.65885</td>
</tr>
<tr>
<td>EAP_DESWC</td>
<td>20</td>
<td>394.00</td>
<td>255.00</td>
<td>649.00</td>
<td>449.900</td>
<td>98.65885</td>
</tr>
<tr>
<td>IELTS_DESPL</td>
<td>20</td>
<td>8.50</td>
<td>1.50</td>
<td>10.00</td>
<td>4.371</td>
<td>2.20292</td>
</tr>
<tr>
<td>EAP_DESPL</td>
<td>20</td>
<td>10.95</td>
<td>1.25</td>
<td>12.20</td>
<td>5.036</td>
<td>3.26039</td>
</tr>
<tr>
<td>IELTS_DESSL</td>
<td>20</td>
<td>28.76</td>
<td>11.57</td>
<td>40.33</td>
<td>19.872</td>
<td>6.96034</td>
</tr>
<tr>
<td>EAP_DESSL</td>
<td>20</td>
<td>29.50</td>
<td>7.28</td>
<td>36.78</td>
<td>15.798</td>
<td>7.70590</td>
</tr>
<tr>
<td>IELTS_DESWLsy</td>
<td>20</td>
<td>.28</td>
<td>1.32</td>
<td>1.60</td>
<td>1.434</td>
<td>.07597</td>
</tr>
<tr>
<td>EAP_DESWLsy</td>
<td>20</td>
<td>.22</td>
<td>1.35</td>
<td>1.57</td>
<td>1.435</td>
<td>.06610</td>
</tr>
<tr>
<td>IELTS_DESWLit</td>
<td>20</td>
<td>1.03</td>
<td>4.00</td>
<td>5.02</td>
<td>4.560</td>
<td>.28253</td>
</tr>
<tr>
<td>EAP_DESWLit</td>
<td>20</td>
<td>.68</td>
<td>4.29</td>
<td>4.97</td>
<td>4.562</td>
<td>.21320</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

With regard to the number of paragraphs (DESPC) in the IELTS essays and the integrated EAP essays, the former had a mean of 3.6 paragraphs while the latter had a mean of 10.65 paragraphs. The standard deviations were 1.353 and 10.718 respectively. The IELTS essays ranged from 1 paragraph to 7 while the EAP essays ranged from 3 paragraphs to 39.

The IELTS essays had a mean sentence count (DESSC) of 14.35 and a standard deviation of 5.519 compared to a mean sentence count of 33.85 for the EAP essays with a standard deviation of 15.083. The IELTS essays had a minimum DESSC of 3 and a maximum of 24 whereas the EAP essays had a minimum DESSC of 9 and a maximum of 61.

The word count (DESWC) also varied considerably between the two types of essays: the IELTS essays had a mean of 264 words with a standard deviation of 82.084 while the EAP essays had a mean word count of 449.9 words with a standard deviation of 98.659. The next lowest
word count was 173. The maximum DESWC for the IELTS essay was 434. The EAP essays were considerably longer, ranging from 255 to 649 words.

To measure the difference in word length across the two essay types, the percentage increase was calculated for each participant, along with the mean (see Table 3.8 below).

Table 3.8: Percentage increase in word length across the two essay types

<table>
<thead>
<tr>
<th></th>
<th>IELTS WORD COUNT</th>
<th>EAP WORD COUNT</th>
<th>% increase in word count</th>
</tr>
</thead>
<tbody>
<tr>
<td>R03</td>
<td>56</td>
<td>510</td>
<td>810.71</td>
</tr>
<tr>
<td>B05</td>
<td>245</td>
<td>551</td>
<td>124.90</td>
</tr>
<tr>
<td>B04</td>
<td>283</td>
<td>584</td>
<td>106.36</td>
</tr>
<tr>
<td>B02</td>
<td>215</td>
<td>429</td>
<td>99.53</td>
</tr>
<tr>
<td>R05</td>
<td>243</td>
<td>467</td>
<td>92.18</td>
</tr>
<tr>
<td>R02</td>
<td>224</td>
<td>422</td>
<td>88.39</td>
</tr>
<tr>
<td>B08</td>
<td>201</td>
<td>364</td>
<td>81.09</td>
</tr>
<tr>
<td>R06</td>
<td>266</td>
<td>477</td>
<td>79.32</td>
</tr>
<tr>
<td>R13</td>
<td>259</td>
<td>439</td>
<td>69.50</td>
</tr>
<tr>
<td>R01</td>
<td>330</td>
<td>548</td>
<td>66.06</td>
</tr>
<tr>
<td>R12</td>
<td>292</td>
<td>470</td>
<td>60.96</td>
</tr>
<tr>
<td>B07</td>
<td>277</td>
<td>444</td>
<td>60.29</td>
</tr>
<tr>
<td>R08</td>
<td>434</td>
<td>649</td>
<td>49.54</td>
</tr>
<tr>
<td>R04</td>
<td>173</td>
<td>255</td>
<td>47.40</td>
</tr>
<tr>
<td>B01</td>
<td>190</td>
<td>265</td>
<td>39.47</td>
</tr>
<tr>
<td>R07</td>
<td>242</td>
<td>331</td>
<td>36.78</td>
</tr>
<tr>
<td>B09</td>
<td>290</td>
<td>394</td>
<td>35.86</td>
</tr>
<tr>
<td>B10</td>
<td>367</td>
<td>496</td>
<td>35.15</td>
</tr>
<tr>
<td>B06</td>
<td>313</td>
<td>419</td>
<td>33.87</td>
</tr>
<tr>
<td>B11</td>
<td>380</td>
<td>484</td>
<td>27.37</td>
</tr>
</tbody>
</table>

Removing the essays for test taker RDG03, who had to leave the IELTS test after 15 minutes, the mean percentage increase in word length from the IELTS essay to the EAP essay was 64.95 per cent. The minimum increase was 27.37 per cent from 380 words on the IELTS essay to 484 words on the integrated task, although it should be noted that even 380 words (which was not the longest IELTS essay) is still considerably longer than the approximately 250 word essays that one might expect of an IELTS test task. The maximum percentage increase in
word count was 124.9 per cent, with a word count of 245 on the IELTS essay, which is the approximate usual length of an IELTS essay, to 551 words for the EAP essay.

Going back now to Table 3.7, the number of sentences in a paragraph (DESPL) was also slightly different: the IELTS essays had a mean of 4.371 sentences per paragraph with a standard deviation of 2.203. On the other hand, the EAP essay had a mean of 5.037 sentences per paragraph with a standard deviation of 3.260. The number of sentences per paragraph ranged from 1.5 to 10 in the IELTS essays and from 1.25 to 12.20 in the EAP essays.

The IELTS essays contained on average 19.873 words per sentence (DESSL) with a standard deviation of 6.960. On the other hand, the EAP essays had shorter sentences: 15.799 words per sentence with a standard deviation of 7.706. The shortest sentences in the IELTS essays were 11.57 words long while the longest sentences had 40.33 words in them. The EAP essays ranged from 7.28 words per sentence to 36.78 words per sentence.

The next result is the number of syllables per word (DESWLsy). Both the IELTS essays and the EAP essays had a mean number of syllables per word of 1.435 syllables per word. However, standard deviation for the IELTS essays was 0.076 whereas the standard deviation for the EAP essays was 0.066. The range for the IELTS essays was marginally larger, ranging from 1.32 to 1.60 as compared to 1.35 to 1.57 for the EAP essays.

The last of the descriptive analyses is the word length (DESWLlt). The two essay types were very similar again for this measure. The mean word length for the IELTS essays was 4.56 letters per word compared to 4.563 letters per word for the EAP essays. The standard deviations were 0.283 and 0.213 respectively. The word length in the IELTS essays ranged from 4.00 letters per word to 5.02 letters per word while the EAP essays ranged from 4.29 letters per word to 4.97.

Tests of normality were carried out for the above analyses and the assumptions were met for DESSC and DESWLsy but not met for the other tests. As a result, Pearson’s product-moment
correlation was carried out for the DESSC and DESWLsy measures and the Spearman rank order correlation and Kendall’s tau were calculated for the other measures. The results were as follows:

- **DESPC**: there was a statistically significant, moderately strong, positive correlation between the IELTS essays and the EAP essays for the number of paragraphs according to the Spearman rank order correlation: \( r_s = .458, p<0.05 \) (two tailed). However, Kendall’s tau was not statistically significant: \( \tau_b = .344, p>0.05 \) (2-tailed).

- **DESSC**: again, there was a statistically significant, moderately strong, positive correlation between the IELTS essays and the EAP essays for the number of sentences as measured by the Pearson product-moment correlation: \( r = .473, p<0.05 \) (two tailed).

- **DESWC**: the correlations between the number of words in the two essay types were also statistically significant for both Spearman’s rho and Kendall’s tau. Spearman’s rho was a strong, positive correlation of \( r_s = .540, p<0.05 \) (two tailed) while Kendall’s tau was a moderately strong, positive correlation of \( \tau_b = .432, p<0.01 \) (2-tailed).

- **DESPL**: the correlations for paragraph length between the two types of essays were both non-significant: \( r_s = .228, p>0.05 \) (two tailed) and \( \tau_b = .169, p>0.05 \) (2-tailed).

- **DESSL**: Spearman’s rho was moderately strong \( r_s = .462, p<0.05 \) (two tailed) and statistically significant. However, Kendall’s tau was non-significant: \( \tau_b = .316, p>0.05 \) (2-tailed).

- **DESWLsy**: there was a statistically significant, strong, positive correlation between the number of syllables per word in the IELTS essays and the number of syllables per word in the EAP essays: \( r = .550, p<0.05 \) (two-tailed).
The Spearman’s rho correlation coefficient was statistically significant and strong: \( r_s = .624, p<0.01 \) (two tailed); Kendall’s tau was moderately strong but also statistically significant: \( \tau_b = .442, p<0.01 \) (2-tailed).

The next analyses using Coh-Metrix were the *Lexical Diversity, Syntactic Pattern Density* and *Readability* measures. One *Lexical Diversity* measure was carried out:

- **LDTTRA**: Lexical diversity, type-token ratio, all words

Three Syntactic Pattern Density analyses:

- **DRNP**: Noun phrase density, incidence
- **DRVP**: Verb phrase density, incidence
- **DRPVAL**: Agentless passive voice density, incidence

and two *Readability* measures were calculated:

- **RDFRE**: Flesch Reading Ease
- **RDFKGL**: Flesch-Kincaid Grade level

Table 3.9 shows the descriptive statistics for the above Coh-Metrix analyses.

The type-token ratio refers to the number of unique words in a text (types) divided by the number of occurrences (tokens) of that word in a text. A high type-token ratio means that the text is more difficult to understand as there is less repetition of words compared to a low type-token ratio, which tells us that words are repeated more frequently in a text, thus making it easier to understand (Graesser *et al.*, 2004: 198). For the IELTS test, this was 0.481 with a standard deviation of 0.781 whereas the mean for the EAP essay was 0.426 with a standard deviation of 0.569. The minimum type-token ratio for the IELTS test was 0.364 and the maximum was 0.696. For the EAP test, these values were 0.34 and 0.56 respectively.

Noun phrase density (DRNP) refers to the number of noun phrases per 1000 words. Thus, the higher the measure, the higher the number of noun phrases (Graesser *et al.*, 2004: 197). In the IELTS essays, there were on average 361.656 noun phrases per 1000 words with a
The next measure was the incidence of verb phrases (DRVP). As with the noun phrase density above, the verb phrase density refers to the number of verb phrases per 1000 words so the higher the number, the more verb phrases there are per 1000 words (Graesser et al., 2004: 197). The results for this measure show that the IELTS essays again had a slightly higher mean than the EAP essays with 254.469 and 234.384 verb phrases per 1000 words respectively. The standard deviations were 8.261 and 5.092 respectively. At the lower end, the range was very similar, with IELTS essays having 191.34 verb phrases per 1000 words compared to 192.89 for the EAP essays. However, at the higher end, there was a wider range:

![Table 3.9: Coh-Metrix content analyses](image)

The next measure was the incidence of verb phrases (DRVP). As with the noun phrase density above, the verb phrase density refers to the number of verb phrases per 1000 words so the higher the number, the more verb phrases there are per 1000 words (Graesser et al., 2004: 197). The results for this measure show that the IELTS essays again had a slightly higher mean than the EAP essays with 254.469 and 234.384 verb phrases per 1000 words respectively. The standard deviations were 8.261 and 5.092 respectively. At the lower end, the range was very similar, with IELTS essays having 191.34 verb phrases per 1000 words compared to 192.89 for the EAP essays. However, at the higher end, there was a wider range:

![Table 3.9: Coh-Metrix content analyses](image)
The IELTS essays had 315.07 verb phrases per 1000 words compared to 275.47 for the EAP essays.

The next measure was the measure of incidence of agentless passive voice (DRPVAL). Again, since this is an incidence measure, this refers to the number of occurrences per 1000 words (Graesser et al., 2004: 197). There was quite a difference between the means on the IELTS test and the EAP test for this measure. The mean for the IELTS test was 5.913 with a standard deviation of 1.556 compared to just 1.944 with a standard deviation of 0.516 for the EAP essay. For both essays, the minimum was zero occurrences and the maximum was 23.17 for the IELTS test and 6.85 for the EAP test.

The final two analyses look at the readability of the two types of essay tasks. The first analysis is the measure of Flesch Reading Ease (RDFRE). Readability formulas are measures of text difficulty. The two measures used here, Flesch Reading Ease and the Flesch-Kincaid Grade Level are just two of many but are perhaps the most popular ones. The results above show that the IELTS essays had a mean Flesch Reading Ease of 65.305 with a standard deviation of 2.285, making the texts slightly harder to read than the EAP essays, which had a mean RDFRE measure of 69.377 with a standard deviation of 2.064. The lowest (and therefore most difficult) RDFRE for the IELTS essays was 37.56 and the highest (and therefore the easiest) RDFRE was 78.73. On the other hand, the EAP essays ranged from the lowest (most difficult) measure of 43.28 to the highest (the easiest) of 82.53.

As is to be expected from the results above, the IELTS essays had a higher RDFKGL mean than the EAP essays. These were 9.088 and 7.508 with standard deviations of 0.666 and 0.680 respectively. The lowest Grade Level for the IELTS essays was 6.14 and the highest was 18.04. Both of these are higher than the EAP essays, which were 3.99 for the minimum and 16.36 for the maximum.
To measure whether there were any significant correlations between any of these measures, tests of normality were again conducted. With the exceptions of DRNP and DRVP, the assumptions were not met. For this reason, parametric tests are reported for these two measures but non-parametric tests are reported for the other measures.

For the type-token ratio (LDTTRa), the results for both Spearman’s rho and Kendall’s tau show statistically significant, strong positive correlations between the IELTS Writing test and the integrated EAP test. The correlations were $r_s = .700$, $p<0.01$ (two tailed) and $\tau_b = .536$, $p<0.01$ (2-tailed) respectively.

For noun phrase density (DRNP), verb phrase density (DRVP) and agentless passive voice density (DRPVAL) the results were all non-significant. DRNP was $r = .393$, $p>0.05$ (2-tailed); DRVP was $r = -.035$, $p>0.05$ (two-tailed) and DRPVAL was $r_s = .226$, $p>0.05$ (2-tailed) and $\tau_b = .176$, $p>0.05$ (2-tailed).

The correlations between the IELTS Writing and the EAP tests for the Flesch Reading Ease (RDFRE) and Flesch-Kincaid Grade Level (RDFKGL) were all significant. Spearman’s rho for RDFRE was a strong, positive correlation: $r_s = .510$, $p<0.05$ (2-tailed) while Kendall’s tau was a weaker yet moderately, strong positive correlation: $\tau_b = .358$, $p<0.05$ (2-tailed). For RDFKGL, Spearman’s rho was a moderately strong, positive correlation: $r_s = .475$, $p<0.05$ (2-tailed) and Kendall’s tau was a slightly weaker but still moderately strong, positive correlation: $\tau_b = .347$, $p<0.05$ (2-tailed).

To summarise the quantitative results for RQ2, there was no statistically significant correlation between the scores on the IELTS Writing test and the integrated EAP test. However, there was a statistically significant correlation between the two for the number of cohesive devices used, although test takers produced considerably more in the EAP essays than in the IELTS essays in general.
Overall, the Coh-Metrix descriptive analyses revealed that, on average, the EAP essays had more paragraphs, more sentences, higher word counts and longer paragraphs than the IELTS essays, yet the sentences in the IELTS essays were longer. The number of syllables per word and the word length were very similar across both types of essays. For the measures of lexical diversity, syntactic pattern diversity and readability, however, the results show that the IELTS essays had higher lexical density and higher syntactic pattern density, thus making them more difficult to read than the EAP essays.

There were statistically significant correlations for the paragraph count (Spearman’s rho only, not Kendall’s tau), the sentence count, the word count, the sentence length (Spearman’s rho only), number of syllables per word and word length as well as on the type-token ratio and the two readability measures but all the other measures were non-significant.

Overall, it could be argued that, although there were no statistically significant correlations between the scores on either of the IELTS tests and the EAP test, there is still a relationship between the two writing tasks in terms of the textual analysis, which revealed seven statistically significant correlations out of the 13 metrics used.

There may be a number of reasons for the lack of a significant correlation between the IELTS Writing test and the EAP test. One of the reasons could be a lack of content or topic knowledge for the IELTS test. In this test, participants had to brainstorm, plan and organise the content. However, students may lack ideas for the independent IELTS Writing task, which will impact on their scores even though they are likely to be familiar with the organisation of such an essay having taken such tests before. On the other hand, the outline for the integrated essay and guided notes pages were provided to help them to identify the structure, content and ideas, which, in turn, were provided by the input text. This may have provided too much scaffolding to some while others may not have performed to the best of their ability in the writing task as the content available to them could have been impacted by the test takers’ differing abilities in listening comprehension or an ability to write good notes. Thus, the lack
of correlation could be due to the differing task types and the demands made on the test taker by the two tasks, yet despite being unfamiliar with this type of task and only having 5 minutes longer, the test takers produced essays based on the integrated task which were considerably longer than the IELTS essays. This could have been due to the high level of scaffolding provided in the EAP test, which allowed test takers to pick out and then reproduce the key information more easily or it could be due to the fact that they did not have to spend too much time planning the content and generating ideas as this was already present after having listened to the lecture. Similarly, this could also have been helped by having the necessary vocabulary and structures available having had the input containing this language. The slightly higher means on the EAP scores and the considerably higher mean word count might provide some evidence that the support offered helped students to perform better. The advantages provided by the support available in integrated tasks are in line with other findings (Leki and Carson, 1997; Read, 1990).

For the Coh-Metrix analyses, DESPC (paragraph count), the results show that the lowest number of paragraphs for the IELTS essays was one. This is likely due to the fact that one of the test takers had to leave to go to an appointment after 15 mins. This student’s need to leave also impacted on the minimum DESWC measure for the IELTS test, which was 56. In the EAP test, however, the student’s results were more in line with the other students. On the other hand, the maximum DESPC measure in the EAP essays was 39. It should be noted that, in the EAP essays, several students used a number of bullet points and headings, which were calculated as paragraphs in each case. With regard to the lack of statistically significant correlations between the IELTS essays and the EAP essays, the DESPC measure was not significant for Kendall’s tau. This could be due to the task types: the input on the EAP test may have helped test takers to structure their content better. Alternatively, given that a number of students used bullet points and headings in this task, it may not have been clear that they were supposed to write a correctly-structured essay or it could be that they knew
they would not have enough time to write all their information out in full sentences so they used bullet points to reproduce as much as possible.

For the DESSC (sentence count) measure, the results show a maximum sentence count of 61 for the EAP essays. The minimum DESPL (paragraph length), on the other hand, was only 1.25 in the EAP essays, although the number of sentences per paragraph for the IELTS essays was 1.5, which again, was due to the student having to leave early. The minimum number of words per sentence (DESSL) was also lower for the EAP essays than for the IELTS essays. Again, all of these results could be due to the use of bullet points and headings in some of the EAP essays or a lack of understanding of paragraphing. This might also account for the non-significant correlations between the two types of essays.

Moving on to the descriptive results for the other Coh-Metrix analyses (see Table 3.7), as can be seen from the results above, the type-token ratio for the IELTS test was higher than for the EAP tests. This is perhaps because several of the words, e.g. culture shock, stages, symptoms, were used numerous times by the speaker and were therefore reproduced frequently by the test takers.

Noun phrase density and verb phrase density are slightly higher for the IELTS essay than the EAP essay. This could be due to the preparation that the students had had for IELTS essays and being taught to use nominalisations in such essays but, given the difference in the task type for the EAP task, they may have completely forgotten about the ‘rules’ of academic writing as they were focused on writing down as much of the content as possible. While this would account for the lack of any statistically significant correlations, this is only speculation and merits further investigation.

With regard to the density of verb phrases on the IELTS essays and the EAP essays, this, again, could have been affected by the use, by some students, of bullet points in the EAP essays. One possible reason for the use of bullet points could be that the stronger students
on the IELTS tests were the ones who understood more of the lecture in the EAP test but, due to time pressure, felt they did not have enough time to write down all their points in full sentences and therefore resorted to the use of bullet points to get more points down, thereby leading to a lack of verb phrases compared to the IELTS essay.

The findings for type-token ratio, noun phrase density and verb phrase density are noteworthy as other research suggests that integrated tasks provide a ‘language repository’ (Plakans and Gebril, 2012, p. 30) so one would normally expect that these tasks would have a wider range of language but this does not appear to be the case here. This might also explain the lack of a significant correlation for the incidence of noun phrases. The cause of this may be the limited scope of the input, which is very focussed on the topic of culture shock, what it is and how to avoid it. Consequently, this must be taken into account in the design of integrated tasks.

With regard to the differences in the use of the Passive Voice, one reason for the considerable differences between the two types of test could be the preparation that students had had for IELTS tests and / or the training they had had in academic writing in their university courses, which may have encouraged them to use the Passive Voice. However, another likely reason is the input text. As spoken language and written language can be considered as being on a complexity continuum (Tannen, 1982, in Luoma, 2004, p. 13), planned speech, although relatively formal, may not be as linguistically complex and formal as if the content were written in an academic textbook. This could have affected the language produced by the test takers because, in focusing on what they had written in their notes, they may have simply forgotten to apply the rules they had learnt, for example, to use the Passive Voice, which could account for the lack of a significant correlation between the two types of tasks. This change of register may be less of an issue with reading-into-writing tasks as the input is delivered in the written medium but is an important consideration which must be borne in mind when designing listening-into-writing tasks because a less formal, less complex input text may have a negative impact on the academic nature of the test taker’s output.
The Flesch Reading Ease and the Flesch-Kincaid Grade Level both confirm the fact that the EAP essays were easier to read on the whole than the IELTS essays, which could be accounted for by a combination of the results outlined above, making the texts easier on the whole.

### 3.4.2.2 Qualitative data

The qualitative analysis revealed a good deal of similarity in the use of common cohesive devices in both essays (see Appendices 6 and 7). Table 3.10 provides an overview of the cohesive devices used in both types of essays.

#### Table 3.10: Comparison of cohesive devices

<table>
<thead>
<tr>
<th>IELTS cohesive device</th>
<th>Totals</th>
<th>EAP cohesive device</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>And</td>
<td>114</td>
<td>And</td>
<td>272</td>
</tr>
<tr>
<td>Or</td>
<td>37</td>
<td>Or</td>
<td>62</td>
</tr>
<tr>
<td>Because</td>
<td>30</td>
<td>Because</td>
<td>60</td>
</tr>
<tr>
<td>For example / I'd like to give an example / For instance</td>
<td>28</td>
<td>There are (5 stages) / (Culture shock) consists of / shows / has (5 stages)</td>
<td>39</td>
</tr>
<tr>
<td>So</td>
<td>22</td>
<td>For example / For instance / such as</td>
<td>35</td>
</tr>
<tr>
<td>But</td>
<td>15</td>
<td>So</td>
<td>22</td>
</tr>
<tr>
<td>However</td>
<td>13</td>
<td>But</td>
<td>20</td>
</tr>
<tr>
<td>I think</td>
<td>12</td>
<td>In this stage / In this / that one</td>
<td>17</td>
</tr>
<tr>
<td>Also</td>
<td>12</td>
<td>First(ly) / At first / First of all</td>
<td>15</td>
</tr>
<tr>
<td>There are</td>
<td>11</td>
<td>Also</td>
<td>15</td>
</tr>
<tr>
<td>Although</td>
<td>10</td>
<td>(The) first one / stage / step is</td>
<td>14</td>
</tr>
<tr>
<td>Therefore</td>
<td>10</td>
<td>The third / fourth / fifth one / stage is ... / Stage number three / four is ...</td>
<td>14</td>
</tr>
<tr>
<td>In conclusion</td>
<td>10</td>
<td>The second (one is)</td>
<td>11</td>
</tr>
<tr>
<td>In my opinion / To my mind / In my personal view / As far as I am concerned</td>
<td>9</td>
<td>Secondly</td>
<td>11</td>
</tr>
<tr>
<td>On the other (hand)</td>
<td>8</td>
<td>Finally</td>
<td>11</td>
</tr>
<tr>
<td>Moreover</td>
<td>6</td>
<td>However</td>
<td>11</td>
</tr>
</tbody>
</table>

As can be seen from the table above, the most widely used cohesive devices in both essays were ‘and’, ‘or’ and ‘because’. In the IELTS essays, the results were:

- ‘and’ (n=114)
- ‘or’ (n=37)
• ‘because’ (n=30)

compared to:

• ‘and’ (n=272)
• ‘or’ (n=62)
• ‘because’ (n=60)

in the EAP essay.

The next most widely used cohesive devices in the IELTS essays were phrases related to giving examples (‘for example’, ‘I’d like to give an example’ or ‘for instance’) (n=28) followed by ‘so’ (n=22), ‘but’ (n=15) and ‘however’ (n=13). This pattern is very similar for the EAP essays: example phrases (n=35), ‘so’ (n=22), ‘but’ (n=20) and ‘however’ (n=11). The only exceptions here are the use of ‘also’, ‘there are’ and other phrases which were used in the EAP lecture for signposting the stages of culture shock. ‘Also’ was used slightly more frequently in the EAP essay (n=15) compared to the IELTS essay (n=11) and the phrase ‘there are’, which was used 39 times in the EAP essays, was only found 11 times in the IELTS essays.

While opinion phrases such as ‘I think’ (n=12) and ‘in my opinion’ (n=9) were used in the IELTS essays, these were not present at all in the EAP essays. Instead, there was much wider use of signposting phrases for sequencing: ‘In this stage / In this / that one’ (n=17), ‘First(ly) / At first / First of all’ (n=15), ‘(The) first one / stage / step is’ (n=14), ‘The third / fourth / fifth one / stage is … / Stage number three / four is …’ (n=14) and ‘The second (one is)’ (n=11).

Other cohesive devices such as ‘although’, ‘therefore’, ‘in conclusion’ and ‘moreover’ were used in both essays but, with the exception of ‘moreover’, which occurred six times in the IELTS essay and seven times in the EAP essays, (n=6, IELTS; n=7, EAP), these were more prevalent in the IELTS essays.
The results demonstrate that there is a great deal of overlap between the use of cohesive devices in both types of essay but a wider range of 'opinion' phrases were used in the IELTS essay whereas a wider range of sequencing devices was used in the integrated EAP essay. As such, although there is some evidence that the cohesive devices used overlap to some extent, the IELTS Writing task cannot accurately predict performance on an integrated listening-into-writing EAP task in terms of the types of cohesive devices that were used.

The extensive use of cohesive devices in the EAP essay could be due to the fact that the mean word length of the EAP essays was so much higher than the IELTS essays.

The wide use of 'I think' and 'in my opinion' in the IELTS essay is likely to be due to the type of essay it was since the essay prompt specifically asked the test takers: 'To what extent do you agree or disagree with this opinion?'. Similarly, the wide use of signposting phrases in the EAP essay could also be due to the type of task it was. In this case, test takers were required to explain a number of sequences (stages of culture shock) or lists of key points (ways to minimise culture shock) but they were not asked to express their opinion on the topic. This is an important implication for task design when test takers are given the option to choose which essay they want to write about: giving completely different topics can impact on task difficulty (Pollitt et al., 1985 in Hamp-Lyons and Kroll, 1987, p. 24) and lead to very different language being produced and this, in turn, can lead to difficulty with marking as some aspects of the language targeted may be less complex in the one task and somewhat more complex in the other, which has implications for test reliability (Hamp-Lyons and Kroll, 1987, p. 25).

Another reason for the wide use of signposting phrases in the EAP essay could be that the input language was structured using cohesive devices. This may have helped test takers to use more cohesive devices in their essays, either because the topic and the content prompted them to use more to improve the flow of the text or because they picked up some language from the input. This latter point is in line with findings from Plakans and Gebril (2012, p. 30).
who found that the use of input texts served as a 'language repository’, helping their test takers with words they only knew in their L1 and ‘providing technical terminology’.

3.4.3 RQ3: How does the input format of an academic lecture affect performance on an integrated EAP listening-into-writing task?

3.4.3.1 Quantitative data

To investigate how the input format of an academic lecture affects performance on an integrated EAP listening-into-writing task (RQ3), the essays were divided into four sections relating to the information reproduced from each of the four input modes. The input modes were as follows:

   a. **Part one: Kinaesthetic and aural** – test takers saw the lecturer, thus enabling them to make use of the lecturer’s body language, gestures and facial expressions;

   b. **Part two: Aural only** – the test takers heard only the audio track;

   c. **Part three: Visual (diagram) and aural** – the test takers heard the lecturer and saw only a PowerPoint slide with a diagram of the aspects being spoken about;

   d. **Part four: Visual (textual) and aural** – test takers heard the lecturer but saw a PowerPoint slide with textual information in the form of key words relating to the aural input.

The essays were then analysed quantitatively and qualitatively. For the quantitative analysis, the number of words written in each section relating to each of the four input modes was calculated. Similarly, the number and percentage of propositions reproduced in each of the four sections was calculated. For the qualitative analysis, the types of propositions reproduced in each section were investigated.

It should be noted here that the word count for this RQ was calculated in Microsoft Word rather than in Coh-Metrix. For this reason, there are some slight variations in the total word counts for the essays as compared with the results of the Coh-Metrix analysis. However, this is because the Coh-Metrix analyses included the identification codes for the students (usually
3 words) in the word count and is also due to differences in the way Word calculates the word count.

Table 3.11 below shows the word counts for each section of each test taker’s essay with the total words. It also shows the mean word count, standard deviation, minimum and maximum for each section and overall, as well as the number of words from the original lecture transcript.

<table>
<thead>
<tr>
<th></th>
<th>Part one Kinaesthetic and aural</th>
<th>Part two Aural only</th>
<th>Part three Visual (diagram) and aural</th>
<th>Part four Visual (textual) and aural</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>B01</td>
<td>75</td>
<td>59</td>
<td>113</td>
<td>14</td>
<td>261</td>
</tr>
<tr>
<td>B02</td>
<td>92</td>
<td>82</td>
<td>162</td>
<td>92</td>
<td>428</td>
</tr>
<tr>
<td>B04</td>
<td>137</td>
<td>121</td>
<td>239</td>
<td>82</td>
<td>579</td>
</tr>
<tr>
<td>B05</td>
<td>104</td>
<td>156</td>
<td>168</td>
<td>119</td>
<td>547</td>
</tr>
<tr>
<td>B06</td>
<td>73</td>
<td>63</td>
<td>204</td>
<td>80</td>
<td>420</td>
</tr>
<tr>
<td>B07</td>
<td>92</td>
<td>96</td>
<td>121</td>
<td>136</td>
<td>445</td>
</tr>
<tr>
<td>B08</td>
<td>82</td>
<td>91</td>
<td>133</td>
<td>55</td>
<td>361</td>
</tr>
<tr>
<td>B09</td>
<td>81</td>
<td>70</td>
<td>153</td>
<td>87</td>
<td>391</td>
</tr>
<tr>
<td>B10</td>
<td>67</td>
<td>131</td>
<td>164</td>
<td>128</td>
<td>490</td>
</tr>
<tr>
<td>B11</td>
<td>105</td>
<td>132</td>
<td>142</td>
<td>105</td>
<td>484</td>
</tr>
<tr>
<td>R01</td>
<td>107</td>
<td>122</td>
<td>188</td>
<td>125</td>
<td>542</td>
</tr>
<tr>
<td>R02</td>
<td>37</td>
<td>96</td>
<td>134</td>
<td>125</td>
<td>392</td>
</tr>
<tr>
<td>R03</td>
<td>90</td>
<td>163</td>
<td>154</td>
<td>100</td>
<td>507</td>
</tr>
<tr>
<td>R04</td>
<td>73</td>
<td>97</td>
<td>85</td>
<td>0</td>
<td>255</td>
</tr>
<tr>
<td>R05</td>
<td>65</td>
<td>113</td>
<td>116</td>
<td>172</td>
<td>466</td>
</tr>
<tr>
<td>R06</td>
<td>57</td>
<td>134</td>
<td>192</td>
<td>93</td>
<td>476</td>
</tr>
<tr>
<td>R07</td>
<td>115</td>
<td>58</td>
<td>158</td>
<td>0</td>
<td>331</td>
</tr>
<tr>
<td>R08</td>
<td>118</td>
<td>156</td>
<td>190</td>
<td>174</td>
<td>638</td>
</tr>
<tr>
<td>R12</td>
<td>67</td>
<td>115</td>
<td>156</td>
<td>129</td>
<td>467</td>
</tr>
<tr>
<td>R13</td>
<td>70</td>
<td>109</td>
<td>111</td>
<td>149</td>
<td>439</td>
</tr>
<tr>
<td>MEAN</td>
<td>85.35</td>
<td>108.2</td>
<td>154.15</td>
<td>98.25</td>
<td>445.95</td>
</tr>
<tr>
<td>STDEVS</td>
<td>23.877</td>
<td>32.116</td>
<td>36.857</td>
<td>50.254</td>
<td>97.592</td>
</tr>
<tr>
<td>MIN</td>
<td>37</td>
<td>58</td>
<td>85</td>
<td>0</td>
<td>255</td>
</tr>
<tr>
<td>MAX</td>
<td>137</td>
<td>163</td>
<td>239</td>
<td>174</td>
<td>638</td>
</tr>
<tr>
<td>WORDS (Lecture)</td>
<td>406</td>
<td>430</td>
<td>562</td>
<td>555</td>
<td>1953</td>
</tr>
<tr>
<td>Percentage (mean / lecture)</td>
<td>21.02</td>
<td>25.16</td>
<td>27.43</td>
<td>17.70</td>
<td>22.83</td>
</tr>
</tbody>
</table>
for each section and the mean proportion of words reproduced by the test takers as a percentage of the original lecture. These figures were calculated using Excel.

The results show that, for part one, the mean word count was 85.35 words with a standard deviation of 23.877. The minimum word count for this section was 37 and the maximum was 137. The original lecture included 406 words in this section so the mean reproduction was 21.02 per cent of the original length. For part two, the mean was somewhat higher, at 108.2 words, with a standard deviation of 32.116. The minimum word count was 58 while the maximum was 163. The original lecture was just slightly longer than the first section with 430 words so the data show that 25.16 per cent of the original word length was reproduced on average. Section three had the highest mean word count with 154.15 words. The standard deviation for this section was 36.857. The minimum word count was 85 words and the maximum word count was 239. Although the word count in the lecture was also somewhat higher at 562 words for this section, the percentage reproduced as a proportion of the original length was the highest at 27.43 per cent. In the fourth section, the mean word count was 98.25 words with a standard deviation of 50.254. The minimum word count was 0 (as some test takers ran out of time) while the maximum was 174 words. This section in the lecture was quite long again: there were 555 words in the lecture so the percentage of the original word count reproduced was just 17.70 per cent for this section. Altogether, on average, 22.83 per cent of the original word length of the lecture was reproduced in the essays.

These results demonstrate that test takers performed better on parts two (aural only) and three (visual (diagram) and aural), whereas parts one (kinaesthetic and aural) and four (visual (textual) and aural) seemed to be more difficult. Even if the two students who did not write anything are removed from the calculation, the percentage of the original word count reproduced in part four is still only 20 per cent.
Appendix 8 shows the number and type of propositions produced by each student and the total number of propositions produced. The number of propositions in each section of the lecture was as follows:

- Part one (kinaesthetic and aural): 14
- Part two (aural only): 26
- Part three (visual (diagram) and aural): 39
- Part four (visual (textual) and aural): 36

Appendix 9 shows the percentage of propositions reproduced per student. These figures were calculated by dividing the number of propositions a student produced for a given section by the total number of propositions in that section and multiplying by 100 per cent. This was done because, as can be seen above, the number of propositions in each section varied from 14 to 39 so to just report the count would distort the findings. The results show that the mean percentage of propositions reproduced in part one was 34 per cent. The minimum was 21 per cent and the maximum was 71 per cent. For part two, 54 per cent of the propositions were reproduced on average. This is based on a minimum of 23 per cent and a maximum of 81 per cent. In part three, 49 per cent of propositions were reproduced on average, with a minimum of 15 per cent and a maximum of 72 per cent while in part four, the mean percentage of propositions reproduced was 33 per cent. The minimum was zero as two students ran out of time. Another student only produced three per cent of the propositions for this section but other than these three students, the minimum to maximum range was 19 per cent to 56 per cent. These results show that, again, students performed best in parts two and three. While they wrote more on average for part three, the percentage of propositions reproduced was highest for part two.

In addition to the number of propositions reproduced, the number of points which were copied verbatim from the slides in part four (visual (textual) and aural) was also investigated. The results showed that there was a split between people who wrote down the ideas from the
slides and those who did not. Five of the 20 students did not write down any points from the slides although it should be noted that two of the students who did not write any of the points were two of the students who did not write down anything for part four. The other 15 students did write down verbatim one or more of the points on the slides; however, the number of propositions written down verbatim varied considerably. The mean number of propositions reproduced was 4.65 out of 11 possible reproductions and the median was 5. Of those who wrote down any points, the minimum was 1 proposition (n=1) and the maximum was 11 (n=2). These results can be seen in Appendix 10.

The results from the quantitative analysis appear to suggest that students perform best when presented with audio-only input or audio plus diagrammatic visual input. They seem to struggle with kinaesthetic and aural input but audio plus textual visual input appeared to be most difficult.

Although the better performance on the audio-only section concurs with other studies mentioned in the Literature review (King, 1994; McKnight, 1994, both in Flowerdew and Miller, 1997), there could be a number of other reasons for this including differences in the type of discourse and the syntactic complexity of the input language. On reviewing the lecture transcript in an attempt to account for the findings above, the researcher found that, in terms of the content, part two (audio-only) appeared to be linguistically the easiest, as much of the content consisted of adjectives. However, a post-test analysis of the lecture transcript revealed that the second part was actually the second most difficult according to the Flesch-Kincaid Grade Level measure:

- Part one: 14.371
- Part two: 13.001
- Part three: 10.038
- Part four: 5.954
As can be seen, there was a wide variation in the complexity of the four parts according to the Flesch-Kincaid Grade Level measure. What was surprising is that the fourth part, which many students seemed to struggle with, was actually the easiest in terms of linguistic complexity. This highlighted an important aspect for the main study which was that the features of the input needed to be better controlled for. While I had tried to keep the four parts of the lecture quite similar in word length and time, I had failed to control for the complexity of the language in each of the four parts. As the study was not a counterbalanced-design or fully-crossed study, it was therefore difficult to know whether students performed best on this section because of the input format or because of the linguistic complexity. On the other hand, there was no visual support, which could be expected to help them, so it was not clear from the results why students performed best. This also highlighted the need to make the main study a counter-balanced design so that differences could be better accounted for. This issue was therefore addressed in the main study.

Another reason why the students may have performed better on the audio section, however, could be cognitive overload (Piola, Olive and Kellogg, 2005). If students have to watch, read and make notes at the same point, there are a number of risks involved: they may focus on watching/listening and forget to write down the notes. They may focus on copying what they see on the slides thereby missing other important aspects or they may simply be distracted by the video input formats whereas the audio-only input format allows them to focus on picking out the key information and writing this down.

3.4.3.2 Qualitative data
To answer RQ3, the propositions reproduced in the essays were examined qualitatively to investigate which propositions students reproduced most from the input provided in the four different formats. The table showing the types and quantities of propositions can be found in Appendix 8.

For ease of reference, the content of the four parts of the essay were as follows:
• Part one: definitions of culture shock
• Part two: symptoms of culture shock
• Part three: the five stages of culture shock
• Part four: how long culture shock will last and what you can do to minimise the effects of culture shock.

In part one (kinaesthetic and aural), the most frequent propositions reproduced were ‘Dr Hofstede’s definition’ (n=19) of culture shock and the explanation that culture shock is ‘(like) being a child again’ (n=18). The other definition of culture shock, provided by Dr Kalervo Oberg, was the next most widely reproduced proposition (n=13). 50 per cent of the students (n=10) also reproduced the propositions that culture shock means that ‘things are done slightly differently’ (n=10) and when you are suffering from culture shock, you are ‘not happy about how things work’ (n=10). Nobody reproduced the proposition that when you got to a different country, ‘people talk in a different way’.

In part two, the most widely reproduced propositions were the words ‘symptoms’ (n=19), ‘lonely’ (n=19), ‘depression / depressed’ (n=18) and ‘(feel) sad’ (n=18), followed by ‘speak[ing] a different language (can be a strain / hard work)’ (n=16) and ‘[you may have] (no) confidence’ (n=16). The propositions which were reproduced least in this section were ‘how might you feel?’, ‘withdrawn’ and ‘make you appreciate aspects of your own culture / value aspects of your own culture’.

Moving on to part three (visual (diagram) and aural), this section dealt with the five stages of culture shock presented as a diagram showing a ‘W’ curve – the screenshot of the input can be found in Appendix 11 – the first ‘U’ of which is the time in the new culture and the second ‘U’ is the time when you return to your home country. The most frequent proposition from this part was ‘5 stages of culture shock’, which was reproduced by all 20 students, followed by ‘Honeymoon Stage’ (n=19) and ‘Disintegration Stage’ (n=19), which were both reproduced by 19 of the test takers. 18 of the participants included ‘feel happy’ (n=18), ‘Reintegration Stage’
(n=18) and 'Autonomy Stage' (n=18) respectively. At the other end of the scale, nobody mentioned that the "'U' on the right is when you return to the old culture', nor did anyone pick up on the fact that 'you’re going back to a culture that you’re no longer used to'.

The last part (visual (textual) and aural), seemed to be the most difficult for the test takers even though, as mentioned above, it was linguistically the easiest according to the Flesch-Kincaid Grade Level measure. In this part, three-quarters of the test takers (n=15) in each case reproduced the following three propositions: ‘how long culture shock will last’, ‘[it] (depends on the) situation (you find yourself in)’ and ‘prepare yourself before you come to the new country’. In this part, there were 6 propositions that nobody reproduced: ‘knowing about it will help (to reduce the effects)’, ‘it won’t last forever’, ‘enjoy sports’, ‘(fancy) learning something new’, ‘go to an evening class (at a local college)’ and ‘you’re not the only one to experience it’. This latter proposition was right at the end of the lecture and, along with the clause that came just before it: ‘remember that culture shock is perfectly normal and natural’ (n=2), was the ‘take-away’ message of the lecture yet very few people reproduced it.

On the whole, the points that students reproduced were the key points which were related to the headings in the guided notes pages provided as part of the test booklet and the most frequently reproduced propositions were in parts two and three, which is in line with the findings from the quantitative analysis. On the basis of the results of the qualitative analysis, we can conclude that it appears that the audio-only input (part two) and audio (diagram) plus aural (part three) are indeed the formats on which students perform best while part four (visual (textual) plus aural) is the section that test takers perform worst on. However, as discussed in the quantitative analysis above, there may be other reasons for this apart from the input format.

From the results for part one, we can see that students felt that the definitions were the most important points to note down. This is not surprising since the guided notes pages for part one gives the headings: ‘What is culture shock?’ and ‘Definitions of culture shock’. 
In part two, it is to be expected that the most frequently reproduced propositions related to the symptoms of culture shock as this was the heading that was provided. On the other hand, those which were reproduced least frequently may not have been reproduced for a number of reasons. First of all, the question ‘how might you feel?’ came right at the beginning of the section so test takers may have been ‘acclimatising’ to the change of format at this point. It is perhaps wise, therefore, to not have any questions for the first few seconds of the text to allow people to tune in to the accent. Although this was already the second part they had heard and, therefore, they should have already been attuned to the speaker, the change of format may have had a similar impact on them. The other two propositions may not have been frequently reproduced because, in the case of the first one, they may not have known the word so would have found it more difficult to reproduce and, in the case of the longer phrase, it may have been too long for them to write down so they omitted it.

In part three, everyone reproduced the key aspects about the stages of culture shock; however, no-one mentioned the second half of culture shock, which is when you return home. This could be because they had not realised that everything would happen again when they return home, so they had disregarded the importance of that part or it could be that they were not focused on that as they were listening to find out about what they would likely experience, given that the test takers were all international students studying at UK universities.

In part four, the most frequently reproduced propositions have fewer occurrences than the propositions in the other sections. This could be caused by several factors: the obvious one would appear to be that this was the most difficult section. However, as discussed in the quantitative analysis for RQ3 above, the post-test analysis of the linguistic complexity of the lecture transcript revealed that this was actually considerably easier than the other three parts. As such, the reasoning that the lack of reproductions was due to the difficulty is not plausible. Another, possibly more likely, reason could be a fatigue effect. By the time they were getting to the end of the second playback of the lecture, they had already spent a considerable amount
of time doing test tasks and had 2 playbacks of a lecture of over 12 mins each. It could simply be then, that they were getting tired or that their interest was slipping. Another possibility is that they ran out of time to write down everything they had noted down. This was certainly the case for at least two students who did not write anything at all in the last section because they ran out of time. This was therefore something to be considered for the changes to the main study.

3.4.4 RQ4: How do learners’ perceptions of their performance compare with observed performance on an integrated EAP listening-into-writing task?

To answer this question, the parts of the EAP essay in which each student reproduced the largest percentage of propositions and the smallest percentage of propositions were identified. The questionnaire results were also analysed to identify the parts that test takers felt they did best and worst on. As can be seen from Table 3.12 below, 13 test takers reproduced the largest proportion of propositions in part two, yet only three test takers felt that they performed best on this part while only one test taker actually reproduced the highest proportion of propositions from part four, yet seven test takers felt that this was the part they performed best on.

<table>
<thead>
<tr>
<th></th>
<th>Strongest EAP</th>
<th>Strongest Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part 1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Part 2</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Part 3</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Part 4</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Weakest EAP</th>
<th>Weakest Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part 1</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>Part 2</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Part 3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Part 4</td>
<td>9</td>
<td>3</td>
</tr>
</tbody>
</table>

Looking at the results for the fewest propositions reproduced, 11 students performed worst on part one yet only five thought this was the case whereas part two was not the weakest
part for anyone but eight test takers perceived this as their worst performance, when actually, as mentioned above, this was the section that 65 per cent (n=13) performed best on.

The parts with the highest percentage of propositions reproduced were correlated with the sections that test takers felt they performed best on. Similarly, the parts that test takers felt they did worst on were correlated with the parts in which they reproduced the lowest percentage of propositions.

Table 3.13 shows which parts of the essay students reproduced the highest and lowest percentage of propositions in and test takers’ perceptions of their best and worst performance according to the feedback questionnaire. Only two students’ (B08 and B11) perceptions of their performance was in line with observed performance on both highest and lowest percentages.

Table 3.13: Strongest and weakest performance comparison

<table>
<thead>
<tr>
<th></th>
<th>Strongest</th>
<th></th>
<th>Weakest</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EAP</td>
<td>Questionnaire</td>
<td>EAP</td>
<td>Questionnaire</td>
</tr>
<tr>
<td>B01</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>B02</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>B04</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>B05</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>B06</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>B07</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>B08</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B09</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B10</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>B11</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>R01</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>R02</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>R03</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>R04</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>R05</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>R06</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>R07</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>R08</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>R12</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>R13</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
of propositions reproduced. In addition, one student’s (R03) perception of their best performance matched the observed performance while two students (B09 and R04) were able to identify their worst performances correctly. The three students whose perceptions matched observed performance were the three students in Table 3.12 above who correctly identified part two as being their best performance.

Before calculating the correlations, tests of normality were carried out. As would be expected from this kind of data, the assumptions for normality were not met so parametric tests were carried out. Kendall’s tau was again calculated due to the small sample and the large number of tied ranks (Field, 2009, p. 181). The resulting correlation for the highest percentage of reproductions compared to perceived best performance was non-significant: $\tau_b = .104$, $p>0.05$ (2-tailed), while Kendall’s tau for the lowest percentage of reproductions compared to perceived worst performance was also non-significant: $\tau_b = .274$, $p>0.05$ (2-tailed).

The fact that test takers’ perceptions were so different to the observed results in the EAP essay as measured by the percentage of propositions reproduced in each part raises questions about self-assessment: what influenced their decision to choose which part they felt they performed best or worst on? Is this linked to how much they wrote or what percentage of propositions they understood and reproduced? Did they know what ‘success’ meant and how to measure it? These are all questions which must be taken into account when training students to self-assess, especially since ‘assessment as learning’, in which students learn to self- and peer-assess, is often promoted as a way of helping students to ‘take responsibility for their own learning’ and ‘understand the next steps in learning’ (NSW Education Standards Authority, 2019), both of which are extremely important skills for higher education. While the topic of self-assessment is outside of the scope of this project, it is important to bear in mind for the main study that students’ perceptions may not tell the ‘whole’ story if they are not used to measuring their own performance and success.
3.4.5 Questionnaire results

As mentioned above, some of the questionnaire results were used to provide personal information, which is outlined in Section 3.3.1 above while other questions served to answer RQ4 (see Section 3.4.4). The remaining information gleaned from the questionnaire served as validation evidence for the tools used in the exploratory study to provide information which would feed into the study design for the main study. For this reason, the other main results of the questionnaire are presented here. (For a copy of the questionnaire, see Appendix 12.)

The first question asked about test takers’ perceptions of their performance on the IELTS listening test. Four respondents felt they performed better than they normally would in a listening test while six felt that they did as well as they normally would. On the other hand, eight felt they performed worse and two felt they performed much worse than they normally would on a listening test. Nobody felt they did much better than they normally would. The mean score for this question was 3.40 with a standard deviation of .940. The median was 3.5 and the mode was four (‘I think I did worse than I normally would on a listening test’).

Question two investigated perceptions about the difficulty of the IELTS writing test. Three respondents felt it was ‘difficult’ while 14 found it ‘quite difficult’. Three of the respondents felt it was either ‘quite easy’ (n=2) or ‘very easy’ (n=1). The mean score for this question was 2.05 with a standard deviation of .686. The median was 2.00 and the mode was also 2 (‘quite difficult but I could cope with it’).

Questions three and four were designed to investigate the test takers’ background knowledge and the impact of the topic. For question three, two people said they already knew a lot about the topic whereas 11 said they had known a little before the lecture. Seven participants said they had not known anything about the topic beforehand. These responses represented a mean of 2.25 and a standard deviation of .639. The median was 2.00, as was the mode (‘I knew a little about the topic before’).
For question four, 13 respondents said the topic had had a positive impact on their performance. Five felt it had had no effect: One person felt it had had a negative impact and one person responded that they were not sure if it had had any impact on their performance. The mean for this question was 1.50 with a standard deviation of .827. The median and the mode were both 1 (‘positive impact’).

Questions five, six, seven and eight investigated participants’ perceptions about the sections of the essay that they performed best and worst on and possible reasons for these perceptions. The responses to questions five and seven have been dealt with in Section 3.4.4 above but, for ease of reference, these will be summarised again here.

On question five, two people felt they performed best on part 1 and the same for part 2. In parts three and four, eight people in each case felt they performed best on that part. When asked about their reasons for their best performance (question six), eight test takers stated that they were helped by the diagram:

‘Diagram is easily (sic) and clearly to understand. Most important is when speaker is saying (sic), you can connect with diagram’ (R04, male, Taiwanese, aged 22-25)

while four said they were helped by the text:

‘[I can] copy words while listening’ (B04, female, Spanish, aged 22-25)

and:

‘I can guest (sic) the detail by using the words that I have seen’ (R05, female, Thai, 31-40).

Two people said they were able to concentrate on the audio but one person said they felt they performed best on part one as they could see the speaker’s body language and one person felt they performed best on part four because they:
listen (sic) it carefully but didn’t write because of the time [in the essay writing phase]

(B09, female, Taiwanese, aged 18-21)

Four people did not provide any response to question six.

For question seven, five respondents felt that part one was their worst performance and nine people felt this was part two. Four and two people respectively felt they performed worst on parts three and four respectively.

With regard to the qualitative responses to question eight, four people felt they performed worst on part one (n=2) and part two (n=2) because they could not understand enough:

'[It was] very difficult to understand’ (B01, male, Omani, aged 26-30)

'[I] could not understand what is most important’ (B07, female, Japanese, aged 31-40).

Four people said they performed worse on part two because they could not see the speaker:

'[I] can’t (sic) see the speaker that (sic) I felt not comfortable’ (B05, male, Chinese, 22-25)

‘no interaction seem (sic) negative feeling’ (R12, female, Thai, aged 31-40).

Two people felt they performed worse on sections 1 and 4 respectively due to cognitive overload:

'[I] have to read text and listen at same time – difficult for me’ (R04, male, Taiwanese, aged 22-25)

‘might you (sic) miss some words while you are looking’ (R07, female, Iraqi, aged over 40).

Three people said they could not stay concentrated. One person said that:

'(h)earing is not enough to attract your attention’ (B10, male, Omani, aged 26-30)
while others just said they could not focus on the content. Other reasons for the poor performances include being distracted by the reading (n=1), being unclear about the task (n=1) and external factors such as not being able to hear very well (n=1), not being able to see the content (n=1) and being tired (n=1). Two test takers provided no response to question eight.

Question nine was designed to investigate construct-irrelevant variance (Messick, 1989) and any other factors which may have negatively affected test takers’ performances. It is important to note that, for this question, participants could select as many answers as they wished. In total, nine people said that the sound quality of the video negatively affected their performance and seven said the number of times they saw the video (although there is no information about whether they thought watching the video twice was too often or not enough). Thirteen people said that the amount of information they had to write down was an issue. There were also several qualitative comments on this question: two people commented on the speaker’s speech rate, while two people stated that they were either tired (n=1) or stressed (n=1).

Question ten asked about test takers’ perceptions of the difficulty of the EAP essay based on the lecture topic. Eight said they found it ‘quite difficult’ whereas eleven and one respectively felt that it was either ‘quite easy’ or ‘very easy’. Nobody claimed to have found the EAP essay ‘very difficult’. The mean score for this question was 2.65 with a standard deviation of .587. The median and the mode were both 3 (‘quite easy’).

Finally, question eleven asked if there were any other factors which affected test takers’ performances either positively or negatively. The big issue here was the timing. Nine people said they did not have enough time to complete the essay and two people felt they did not have the knowledge of the genre they needed. On the other hand, two people felt that the future benefits of the task had a positive effect on their performance:
‘[It] give me experience to do any test in future and improves my listening’ (B01, male, Omani, 26-30)

and two people commented on the benefit of topic knowledge. One wrote:

‘Positively: if you heard or know something about the topic even little but it make (sic) you feel relax and encourage to know more (sic)’ (R12, female, Thai, 31-40).

One person commented on the thorough preparation through the input:

‘I think the lecture give (sic) full information in order to answer’ (R13, female, Thai, 26-30)

while one person mentioned the ease of comprehension:

‘Lecturer’s talk is very clear so I can listen to understand easily’ (B08, female, Japanese, 31-40).

Three people did not respond to this question.

The results of the first question about test takers’ perceptions of their performance on the IELTS test were very mixed with 50 per cent saying they did as well as usual or better and 50 per cent saying they felt they did worse. The reason for this could be that some students would certainly have done an IELTS (or similar) test to enter the university. However, others may have done a different kind of university entrance test so they may not have been familiar with the IELTS listening task types or may have done a test in which they may have been allowed to listen more than once. Another possible reason is that they may simply not have felt as confident due to a lack of knowledge of the topics in the test. Anxiety could also have been a reason for their perception of a poor performance. ‘Concern about one’s level of performance, negative task expectations and negative self-evaluation’ are cognitive aspects of anxiety (Eysenck, 1979, p. 364 in Arnold, 2000, p. 779) which can have a negative impact
on performance (Elkhaifaifi, 2013, p. 207), with the perception that listening is a difficult skill (Hasan, 2000, in Vandergrift, 2013, p. 191) which may also lead to anxiety.

17 out of the 20 participants in the exploratory study also found the IELTS writing task ‘difficult’ or ‘quite difficult’. As mentioned above, one reason for this could be that they may not have all taken an IELTS test to enter the universities; however, one of the difficulties with the IELTS tests and other independent writing tasks such as the IELTS ones is that test takers may lack ideas and topic knowledge so they may struggle to perform as well as they might otherwise. By contrast, only eight of the participants rated the EAP essay as ‘quite difficult’ while the other 12 found it either ‘quite easy’ or ‘very easy’. This was perhaps due to a combination of having some background information (n=13) before the test and the participants having been given the information they needed to write the essay so they did not need to concern themselves with generating ideas, which is an important aspect of source use in integrated tasks (Plakans and Gebril, 2012, p. 27).

Timing and the amount that students had to write down was clearly an issue for several students so this needed to be considered in the changes to the main study.

With regard to construct-irrelevant variance, the sound quality of the video was an oft-cited factor which participants felt affected their performance. This was due to the acoustics in which the original lecture was filmed so this would need to be considered when refilming the lecture for the main study.

The number of times that participants could watch the video was another issue. Unfortunately, it was not clear whether the participants who selected this option would have preferred to see the video three times instead of two, or whether they would have preferred just once. However, one person commented that they needed to listen again:

‘[It was] difficult to involve (sic) all information ... I should listen more than two times ...’ (B10, male, Omani, 26-30).
However, that would have made the EAP test over 1.5 hours long, and it is likely that at least some of the students would then have become bored by the input and distracted (Ur, 1984, cited in Hamouda, 2013, p. 125) thereby losing the motivation to write the essay.

3.4.6 Conclusions from the exploratory study findings

The exploratory study was intentionally complex in comparing discrete tasks versus integrated tasks and in using four input formats in the integrated task. The objective was to investigate several areas which had been highlighted as gaps in the literature with the intention of revealing aspects which it would be particularly worthwhile to focus on in the main study. In this section, I summarise the results of each of the four RQs and consider which were most promising to investigate in the main study.

RQ1 looked at whether there is a relationship between the IELTS Listening test and the IELTS Writing test, on the one hand, and the EAP task, on the other. The results of the exploratory study demonstrate that there are weak positive, but non-significant correlations between the IELTS listening test, on the one hand, and both the IELTS writing test and the integrated EAP test, on the other hand (see Section 3.4.1). The IELTS tests were used as a way of gauging test takers’ language proficiency before doing the main EAP task. However, requiring test takers to take the IELTS Listening and the IELTS Writing test before the EAP test proved to be rather cumbersome and time-consuming as well as quite tiring for the test takers. As a result, I felt that, while a larger study may reveal significant correlations between these tests, continuing with this RQ for the main study was not the most practical area of research.

RQ2 addressed the relationship between performance on the IELTS Writing test and the integrated EAP test. In this case, a non-significant correlation was found between the IELTS Writing test and the integrated EAP test (see Section 3.4.2.1). However, the results of the analysis of cohesive devices produced in the two types of written outputs revealed that while the EAP essays contained on average more than twice as many cohesive devices as the IELTS
essays, there was a significant, moderately, strong positive correlation between the two types of essays (see Section 3.4.2.1).

Similarly, the Coh-Metrix analyses (see Section 3.4.2.1) revealed statistically significant, moderately strong positive correlations between the two essays for seven of the 13 metrics (sentence count, word count, number of syllables per word, word length, type-token ratio, Flesch Reading Ease and Flesch Kincaid Grade Level). The other six (paragraph count, sentence length, noun phrase density, paragraph length, agentless passive voice and verb phrase density) were all non-significant. Overall, the EAP essays were nearly twice as long as the IELTS essays on average but the IELTS essays were slightly more complex according to the Flesch Reading Ease and Flesch-Kincaid Grade Level indices.

With regard to the cohesive devices used in the two writing tasks, co-ordinating and subordinating conjunctions (and, or, because, so, but, however) and example phrases were the most widely used in both types of writing. However, opinion phrases were used widely in the IELTS essays whereas the EAP essays contained more sequencing phrases (see Section 3.4.2.2).

Although the correlations between the scores for these two tests were not significant, the Coh-Metrix analyses did reveal some significant correlations and a larger study may have shown more significant correlations, including between the scores. The Coh-Metrix results indicated that, to some extent, the two writing tests were testing the same constructs, which was a potentially useful area for the main study. However, since a good deal of research had already been done on discrete writing tasks versus integrated writing tasks, albeit usually listening- and reading-into-writing or just reading-into-writing tasks, I decided that this was not the most promising one in terms of the contribution to the field.

RQ3 investigated the effect of the input format on performance on the integrated EAP task (see Section 3.4.3). The results show that the mean percentage of propositions reproduced
varied across the formats. Test takers performed best in parts two (aural only) and three (visual (diagram) plus aural), reproducing 54 per cent and 49 per cent of the propositions respectively but performed worst on part four (visual (textual) and aural), reproducing on average only 33 per cent of the possible propositions. In terms of the reproductions from the PowerPoint slides in part four, 15 out of the 20 students reproduced verbatim one or more of the propositions on the slides, with the mean being 4.65 of a total of 11 possible reproductions.

In part one, the reproductions focussed on the definitions of culture shock and the two people who provided these definitions while in part two, the focus was on the symptoms of culture shock. In part three, students reproduced the five stages of culture shock most frequently but they did not mention anything about the second ‘U’ curve, that is, when one returns to their home culture. Finally, the reproductions in part four focused on the duration of culture shock and how to minimise it (see Section 3.4.3.2).

There were several aspects that came to light as a result of RQ3 which made this the most promising question to focus on in the main study. These were:

a) the potential differences in performance on an integrated listening-into-writing test when the input is presented as a podcast or as a vodcast. This is particularly relevant as more and more teaching is becoming blended or done completely at a distance, particularly in the current Covid-19 pandemic;

b) the potential impact of different types of visual input media (text versus diagram) on the students’ written output including how many propositions they reproduce from slides with textual or diagrammatic input and in what form (verbatim or paraphrased);

c) the implications for construct that come with a more authentic, video-based lecture input using PowerPoint as well as verbal input.

I felt that this research question had the greatest potential impact for the main study and for the field of language testing because, although a number of studies had already looked at
different input media, these had largely focussed only on listening tests rather than integrated listening-into-writing tests (see Section 2.3.4). How test takers used the audio and visual sources of information from the input in their written output was also a promising area as investigating this would allow an insight into whether test takers reproduced more from the audio input or wrote down information from the visual input. This would potentially highlight the need to reconsider the construct of authentic academic listening. All of the above points meant that, while RQs 1 and 2 may have provided additional insights for the research field, RQ3 seemed to have the most potential for impact as there was a clear gap in the language testing research in relation to source text use from listening texts.

In order to evaluate test takers’ perceptions of their performance, RQ4 looked at the results of the feedback questionnaire (see Section 3.4.4). Although 13 of the test takers performed best in part two, only three test takers perceived this to be their strongest performance while eight considered it the section they performed worst on. In part four, only one test taker performed best on this part yet seven considered this to be the case. In contrast, five test takers considered their worst performance to have been in part one yet the reality revealed that this was the worst section for 11 of the 20 test takers and the other nine performed worse on part four but only three thought this to be the case. In the end, only two test takers were able to evaluate both their best and worst performances while a further three students managed to evaluate one of the two correctly. These results resulted in only very weak, positive but non-significant correlations between the percentage of reproductions in each case and the perceived best and worst performances.

RQ4 was also promising because there is a definite move towards more self- and peer-assessment these days; however, these findings revealed that students may not, in fact, be very good at self-assessing their performance. With only 20 participants, the exploratory study was very small; therefore, a larger sample - as in the main study – could potentially shed more
light on the ability of students to self-assess. For this reason, I decided to keep this RQ for the main study.

Finally, the results of the questionnaire were presented (see Section 3.4.5). These served as validation evidence of the instruments. Overall, test takers felt the topic had a positive impact on their performance and also the future benefits of the task. Two people also commented on the benefit of topic knowledge while one person highlighted the thorough preparation for the writing task from the input. However, a number of areas for improvement were highlighted including the lack of time available, sound quality of the video, the number of times they saw the video (although it is not clear if twice was too much or too little), the amount of information to write down, and the speaker’s speech rate.

These results provided valuable information regarding the quality of the EAP test but, more importantly, highlighted the areas which needed to be addressed before using this assessment tool in the main study.

For the main study then, I decided to compare the two formats which had produced the best and worst performances in the exploratory study, namely, audio only and visual (textual) plus aural respectively. As outlined in Section 3.4.3.2 above, this would allow me to investigate which input format led to the most reproductions. By using textual input, too, I could also investigate what form the reproductions take, that is, whether the propositions reproduced are paraphrased or quoted verbatim from the audio or the textual input. This would provide information about the construct that is being tested when students watch and listen to a typical lecture which involves PowerPoint input compared to audio only input. In turn, this could have potential implications for both teaching and testing.

Before I could start the main study, however, it was clear that there were a number of areas of improvement for both the study design and task design so several changes needed to be made. These are discussed in Section 3.5 below.
3.5 Changes for the main study

The exploratory study revealed areas which were worthy of further investigation but at the same time, it highlighted several issues to be resolved or changed for the main study. An overview of the changes made to the main study in terms of study design is in Table 3.14 in Section 3.5.1 below while Section 3.5.2 highlights the key features of the task design in the integrated EAP test that were retained or refined.

3.5.1 Implications for study design

Table 3.14: Features of the study design to be changed

<table>
<thead>
<tr>
<th>Exploratory study</th>
<th>Main study</th>
<th>Reason(s) for the change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Four different test conditions</td>
<td>Two test conditions (audio only and audiovisual plus PowerPoint).</td>
<td>To simplify and improve main study; To focus in on impact of different input media on output;</td>
</tr>
<tr>
<td>Test takers from various academic and linguistic backgrounds</td>
<td>Participants from similar academic and linguistic backgrounds</td>
<td>To avoid impact that different language and academic backgrounds may have on results;</td>
</tr>
<tr>
<td>Same tasks same order</td>
<td>Counterbalanced measures design</td>
<td>To avoid order effect in the study;</td>
</tr>
<tr>
<td>Controlled for length and number of words in EAP test but not linguistic complexity</td>
<td>Control for length, word count, number of propositions and linguistic complexity in input formats</td>
<td>To remove the issue that differences in performance could have been down to linguistic complexity rather than input format;</td>
</tr>
<tr>
<td>The EAP essay received only an overall mark</td>
<td>EAP summaries to be divided into two halves according to input formats</td>
<td>To focus on differences in performance between two input formats rather than independent versus integrated writing tasks;</td>
</tr>
<tr>
<td>Two trained raters marking essays</td>
<td>Up to three trained raters mark each summary</td>
<td>To improve scoring validity (Weir, 2005, p. 22);</td>
</tr>
<tr>
<td>Calculate rater reliability using Classical Test Theory</td>
<td>Calculate rater reliability using Rasch Analysis and Fair Average score</td>
<td>To improve scoring validity by taking account of rater harshness / leniency in awarding scores;</td>
</tr>
<tr>
<td>Use of IELTS Public Version descriptors for marking</td>
<td>Use of CEFR Overall Writing Production scale for rating summaries</td>
<td>To use a scale familiar to all three raters; Not logical to use IELTS rating scales as IELTS tests not being used;</td>
</tr>
<tr>
<td>Use of Coh-Metrix as text analysis tool</td>
<td>Use of Text Inspector for text analysis</td>
<td>To investigate indices in exploratory study plus EAP vocabulary and CEFR levels of the vocabulary used;</td>
</tr>
<tr>
<td>Analysis of percentage of propositions reproduced from input</td>
<td>Additional analysis of paraphrased versus verbatim reproductions</td>
<td>To refine coding scheme; To analyse how propositions were reproduced.</td>
</tr>
</tbody>
</table>
3.5.1.1 Focus of the study
Further to the study results, I decided to focus on the differences in performance between two input format (audio only versus audiovisual plus PowerPoint) for the main study as the exploratory study results highlighted audio only (part two) as being the part that the majority of participants (n=13) performed best in while the use of audiovisual input in parts one, three and four had different results (see Section 3.4.3). Most participants (n=11) performed worst on part one (kinaesthetic and aural). The other nine participants performed worst on part four (visual (textual) plus aural) while part three (visual (diagram) plus aural) was the part that everyone else except one person performed best on. It therefore made sense to focus on audio vs audiovisual input as the bigger picture but to include textual and diagrammatic input as part of the audiovisual input to see to what extent test takers were influenced by the written input text.

Comparing audio vs audiovisual also fitted in better with the current research at the time, which seemed to be focusing on the use of audio vs video and context vs content visuals (Chang, Lei and Tseng, 2011; Ockey, 2007).

Investigating the impact of input provided via podcast (audio) or vodcast (video) would also potentially have a wider reach as it would have implications for both testing and teaching; the results would not only inform integrated listening-into-writing testing practices but would also contribute to an understanding of how to provide input for flexible teaching approaches such as the flipped classroom, blended learning or online learning, which have been gaining in popularity in recent years (Muldrow, 2013; Webb and Doman, 2016) and which has now come into its own in these times of coronavirus, which is happening as I write.

3.5.1.2 Overall study design
The first major change here was in regard to the participants. In the exploratory study, students were selected from UK universities as it was envisaged that the main study would also involve international students studying at UK universities. However, this was changed for
two reasons. One was purely pragmatic: the difficulty of recruiting international students. The foundation programme at the researcher’s own university had dramatically reduced in size such that there would not have been enough participants there and, despite the fact that there are large numbers of international students at other UK universities, there was a reluctance on the part of the universities approached to allow me to collect data there. The second reason, and more important reason, was that I decided that having a large number of different educational backgrounds may have had an influence on the results as students had different academic and educational backgrounds. This could have impacted on their comprehension of a native speaker accent, their study skills and / or their ability to take notes. Beasley, (1990, p. 11) highlights the fact that East Asian students have ‘study skills problems related to lack of experience in lecture note-taking …’ while Cohen (1979, p. 160 in Beasley 1990, p. 11) found that foreign students sometimes had difficulty identifying ‘conjunctive words and signalling cohesion’ including ‘basic ones like however and thus’. Consequently, I decided to recruit students from similar linguistic and educational backgrounds. Due to the current emphasis on internationalisation in Russia and Ukraine as outlined in Section 4.1 below, I selected one university from each of these two countries.

The most important change to the study, however, was the use of a counterbalanced measures design. In the original study, everyone did the same tasks in the same order. However, this led to an issue with interpreting the results as it was not entirely clear if the test takers’ success or otherwise was due to the input format or the content of the different parts. Although the four parts were approximately equal in length and number of words, the linguistic features of the input had not been controlled for. Consequently, there was considerable variation in the linguistic complexity of the input, which could have impacted on the participants’ ability to understand the input, and there was considerable variation in the number of propositions in each part. Both of these points may have led to participants being able to reproduce more propositions from certain parts of the lecture.
As the study design was changed to a counterbalanced measures design in which the test takers were divided into two groups, this meant a change to the input formats. For the main study, all the participants would hear the same content across the whole lecture but, in terms of the input mode, the lecture was divided into two parts: one part was aural only and the other part showed the speaker and the PowerPoint slides as well. One group would hear the audio part first then watch the video part while the second group would see the video part first followed by the audio only part.

The change of focus from overall performance on an independent EAP essay versus an integrated EAP essay to the differences in performance on an integrated EAP summary based on two different input modes, each of which contributed to one half of the EAP summary, meant that the different sections of the summary now needed to be marked. This had not been the case in the exploratory. Instead, the focus in the exploratory had been solely on the number and type of reproductions from each of the four input modes. Dividing up the summaries would allow each half to still be analysed for reproductions but by giving each half of the summary a score, this could then be used to investigate any statistically significant differences in performance based on the two input formats.

This led to a change in the rating of the summaries for the main study. For the quantitative analysis of the exploratory study, two raters had rated the papers and then agreed on a composite score for each summary. As there were going to be three raters in the main study and considerably more test takers, I decided to use Multi-Faceted Rasch Analysis (Linacre, 1989) to calculate the Fair Average score, which would then be used for the analysis to investigate differences between the two groups.

I also decided that the rating scale should be changed. In the exploratory study, the rating scale used was the IELTS Public Descriptors. However, since the independent EAP (IELTS) essay was no longer the focus, it did not make much sense to use these descriptors. The three raters involved in the main study had all worked together at the same university. In this
institution, the departmental rating scales were based on the scales in the *Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR)* (Council of Europe, 2001). However, the CEFR-based rating scales that the raters had previously worked with were analytic scales. As the use of analytic scales in the exploratory study had been quite time-consuming without providing any useful additional data, I decided to use holistic scales for the main study. As all three raters were familiar with the CEFR and this offers a comprehensive set of writing descriptors, this was the obvious source for rating scales for the main study. Initially, I identified the three scales that were most relevant to the integrated EAP task: the ‘Overall Written Production’ (Council of Europe, 2011, p. 23) scale, the ‘Reports and Essays’ scale (Council of Europe, 2011, p. 24) and the ‘Processing Text’ scale (Council of Europe, 2011, p. 26). The three raters started rating the summary halves using all three scales. However, this proved to be very time-consuming and, consequently, very expensive. While I was fortunate enough to have received a British Council Assessment Research Awards and Grants (ARAGs) Award to help towards the costs involved in data analysis (rating and coding), the time and costs involved for each rater to provide six ratings per test taker (one for each scale for each half of each summary) were prohibitive. The raters discussed what would be the best solution and as a team, we decided that the Overall Written Production scale would be most useful because what we were looking to rate was, in fact, the test takers’ overall written production and this was also the scale that we felt was most likely to produce the highest level of inter-rater reliability. This scale distinguishes well between the levels targeted in the main study, showing a clear development from ‘simple isolated phrases and sentences’ (CEFR A1), followed by ‘simple phrases linked with simple connectors’ (A2) on to the B1 level, which demonstrates ‘straightforward connected texts ... linking a series of shorter discrete elements into a linear sequence’ and then ‘clear, detailed texts ... synthesising and evaluating information and arguments ...’ (B2 level) (Council of Europe, 2001, p. 61) and as we were all familiar with this scale, we felt it would lead to a good level of inter- and intra-rater reliability.
The next change involved the text analysis software. Coh-Metrix (Graesser et al., 2004) had been used for the exploratory study. This tool provides a large number of analyses; however, in the main study, the text analysis tool was changed to Text Inspector as this software produces measures for the descriptive analyses as in Coh-Metrix but also automatically analyses EAP vocabulary and provides an analysis of the CEFR levels of the vocabulary in the text. This was considered a useful measure since the new rating scale was based around the CEFR.

For the qualitative analysis, the coding was also refined to allow a more in-depth look at the propositions which were reproduced in the summaries. In addition to looking at the percentage of propositions reproduced from the audio content and from the slides as in the analysis in the exploratory study, another layer of coding was carried out to distinguish between propositions reproduced in the writer’s own words (i.e. by paraphrasing the input) versus verbatim reproductions from the input text (either form the visual or aural input). However, rather than adopting Keck’s (2006) approach of coding for the extent of the paraphrase (‘near copy’, ‘minimal revision’, ‘moderate revision’ or ‘substantial revision’) or Campbell’s (1990) distinction between ‘near copy’ and ‘paraphrase’, both of which would have resulted in a far more complex coding system, and given that there was no expectation that students had any understanding of citation practices, I decided that it was only necessary to see if the test takers had attempted a paraphrase or copied verbatim what they had heard or seen and to code for these accordingly. The new coding scheme therefore involved coding by word, clause or sentence-length reproductions to investigate the length of the reproductions.

3.5.2 Implications for task design
In this Section, there are two tables (Table 3.15 and Table 3.16). Table 3.15 highlights key features of the task design in the integrated EAP test that were retained unchanged and Table 3.16 highlights the refinements made to the task design. In both cases, reasons for retention or refinements are given.
Table 3.15: Features of the task design to be retained

<table>
<thead>
<tr>
<th>Retained features</th>
<th>Reason for retention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic of culture shock</td>
<td>Most people knew only a little about the topic so background knowledge would not play a big role; To avoid need to pilot test task again as the input would be almost identical and the output, albeit shorter would also be the same as previously;</td>
</tr>
<tr>
<td>Double playback on input</td>
<td>To allow plenty of opportunity for note-taking and checking; To reduce anxiety in accordance with literature; (see Section 2.3.1.2)</td>
</tr>
</tbody>
</table>

Although some refinements were made to the test, the EAP test task was fundamentally the same as was used in the exploratory study in so far as students would take notes on the lecture, which was almost identical in content, and write a summary with the same title as previously in the same time frame allowed, albeit with a shorter word count (see Table 3.16 below). Consequently, no further piloting of this tool was undertaken.

Table 3.16: Refinements to the task design

<table>
<thead>
<tr>
<th>Exploratory study</th>
<th>Main study</th>
<th>Reason(s) for the refinements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use of independent IELTS listening and writing tests</td>
<td>Removal of IELTS tests</td>
<td>Change of focus from relationship between IELTS and EAP test to impact of different input media on performance on EAP test; Overall test length and perceived difficulty possibly contributed to fatigue effect;</td>
</tr>
<tr>
<td>No initial ‘screening’ test for language proficiency</td>
<td>Use of Oxford Quick Placement Test (QPT) used as an initial measure of proficiency</td>
<td>Provide initial assessment of test takers’ proficiency for allocation to groups of approximately equal ability; Quick to administer and mark; Many years’ experience using the test;</td>
</tr>
<tr>
<td>Four different input media</td>
<td>Two input media</td>
<td>Focus on impact of different input media on output; (see Section 3.4.3)</td>
</tr>
<tr>
<td>Poor quality audio</td>
<td>Re-recorded lecture</td>
<td>Improve sound quality</td>
</tr>
<tr>
<td>Varying speech rates across four parts of input</td>
<td>Control for speech rate across both parts of input</td>
<td>To ensure speech rate was within acceptable ranges for a B1/B2 exam and in line with research findings</td>
</tr>
<tr>
<td>Linguistic complexity varied across four parts of input</td>
<td>Control for linguistic complexity across both parts of input</td>
<td>To remove linguistic complexity variable (see Section 3.5.1.2)</td>
</tr>
<tr>
<td>Timing: 45 mins for 500 words</td>
<td>Timing: 45 mins for 350 words</td>
<td>To reduce risk of skewing results due to test takers running out of time</td>
</tr>
<tr>
<td>Detailed information for note-taking and essay plan provided</td>
<td>Note-taking scaffolding and essay plan removed</td>
<td>To increase differentiation across different CEFR levels</td>
</tr>
</tbody>
</table>
3.5.2.1 IELTS Listening test

The feedback from the questionnaire suggested that test takers felt that this task did not accurately reflect their ability. While these findings should be treated with caution as there could be a number of reasons for this, not least of which is a possible inability to judge their performance well as evidenced by the findings of RQ4, I decided to no longer use this task in the main study. The main reason for this was that the focus of the study had changed from investigating the differences between independent and integrated tasks to differences between input formats on an integrated task. It was no longer necessary to measure participants’ listening comprehension as a separate task. However, I still felt it necessary to have some basis on which to group test takers for the integrated task. Developing and validating a new tool was beyond the scope of this PhD so I opted for an off-the-shelf test, namely the Oxford University Quick Placement Test (QPT) (Oxford University Press, 2001) (see Section 4.4.1 below). This is a short lexico-grammatical test which is quick and easy to administer and mark so it would be ideal to use for participants in the main study as they could take the QPT as a measure of their general language proficiency. This was deemed appropriate as the test is considered ‘ideal for placement testing and examination screening’ (Geranpayeh, 2003, p. 8). In addition, since the test was developed in conjunction with Cambridge ESOL, it had already been validated by them and, as part of this validation process, had undergone their quality control procedures for item writing and reliability so there was no need to validate the test beforehand.

Furthermore, I had been using this test as a placement test at my university for approximately ten years before embarking on the main study. In conjunction with an integrated EAP test that I had previously designed (see Section 1.3) and students’ performance on the foundation course that I led, I felt this was sufficient evidence that the QPT was a good measure of general language proficiency and would therefore be a useful tool to enable me to quickly assess the test takers’ proficiency level to allocate them to one of two groups.
3.5.2.2 IELTS writing test
Many of the participants had found the IELTS writing test quite difficult and the focus of the study was changed, so I decided to leave out the independent IELTS writing test. With the combination of all the different test tasks, the perceived difficulty of this test may have contributed to a fatigue effect which could have impacted on the scores of the integrated EAP essay. Therefore, to reduce the risk of fatigue, I reduced the overall length of the test by changing the IELTS listening test and the IELTS writing test for the QPT. The IELTS test had also proven quite cumbersome to mark and marking could not be turned around quickly in the exploratory study so all the participants at each of the two universities involved were in the same group for the exploratory study (irrespective of language proficiency). However, as I wanted the test takers in the main study to be split into two groups of approximately equal proficiency, using the QPT instead would also allow for a quick turnaround on the marking so the participants could be put into groups within a very short space of time, thereby reducing the overall time commitment on the part of the participants and the lecturers helping with data collection.

3.5.2.3 Integrated EAP test

3.5.2.3.1 Input (EAP lecture)
The exploratory study served as a pilot for the EAP test and, as such, revealed both strengths and weaknesses of the design of the integrated EAP test. Therefore, the task design was retained for the most part but a number of problematic aspects were addressed and improvements made.

The topic of the lecture (Culture shock) was retained as the exploratory study revealed that most people knew a little about the topic. This is not surprising since the students were all international students studying in the UK so may have heard a bit about it as part of their university induction programmes. The participants in the main study were likely to be studying in their home country so were unlikely to know much about the topic but may know a little
but not too much. However, if they were preparing to study abroad or study with international students, they were likely to find it of interest and it would hopefully have a positive impact on their performance.

Similarly, to reduce anxiety (Field, 2015, p. 33; Holzknecht, 2019), and to allow participants the opportunity to understand and note down as much as they could or wanted to, the double playback was retained.

Nonetheless, the following changes were made to the input:

a. The input format was changed from four different input formats to just two. To cover approximately the same information, this meant that each part was going to be around twice as long as the three minutes recommended by James (1986, p. 41 in Long, 1990, p. 68-69), who suggested that ‘cognitive overload’ may occur if topics are not highly of interest to the listener. Consequently, it was important that the test takers would find the topic interesting. Some studies have found text length to be a determiner of text difficulty (Brindley and Slayter, 2002; Chen, 2005) yet Bloomfield et al. (2011, p. 2321) argue that there is no strong evidence to suggest that text length affects text difficulty. They argue that text length is ‘confounded with other factors’ and they suggest that the density of information in a text might be a factor which increases the difficulty of a text more than its length (p. 2321). Nonetheless, these factors must be borne in mind.

b. The results from the questionnaire in the exploratory study revealed that the quality of the audio and video was an issue (see Section 3.4.5). The lecture therefore needed to be re-designed and re-recorded for the main study so care was taken to ensure that the acoustics were better.

c. Another issue that came to light in the exploratory study through the questionnaire (See Section 3.4.5) was that, although the four parts of the lecture were designed to
all be of a similar duration, the input text had not been controlled for speech rate. In combination with poor quality audio, an unfamiliar accent, a challenging input text and an unfamiliar task type, speech rate may be perceived as faster than it really is and will add to the strain on the listener (Bloomfield et al., 2010, p. 63).

The speech rate was analysed after the exploratory study and was found to vary across the four sections of the test as follows:

Table 3.17 Speech rates by section

<table>
<thead>
<tr>
<th>Part no.</th>
<th>No. of words</th>
<th>Length</th>
<th>Words per minute (wpm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>406</td>
<td>3:14</td>
<td>125.6</td>
</tr>
<tr>
<td>2</td>
<td>430</td>
<td>3:45</td>
<td>114.7</td>
</tr>
<tr>
<td>3</td>
<td>562</td>
<td>4:02</td>
<td>138.2</td>
</tr>
<tr>
<td>4</td>
<td>555</td>
<td>3:54</td>
<td>142.3</td>
</tr>
</tbody>
</table>

As can be seen from the table above, these speeds were all within the ‘slow’ (below 130 words per minute) or ‘moderately slow’ (130-160 wpm) ranges as set out by Pimsleur et al., 1977 cited in Taylor, 2014, p. 22). When compared to the speech rate in international exams, the four parts of the EAP lecture were found to be slower than the average speed for the Cambridge KET exam, which is an A2 level exam (150.6 wpm / 2.51 wps) (Field, 2013, p. 119). Consequently, taking into account that the main target population for the EAP test was B1 / B2 level students, in the main study, the speed of delivery was adjusted so that it fell within the average speech rate speeds for the Cambridge PET (167.4 wpm / 2.79 wps) and FCE exams (207.6 / 3.46 wps), which are B1 and B2 level exams respectively (Field, 2013, p. 119). The new speech rates were 169.09 wpm for the introduction to allow listeners to get attuned to the accent, 177.75 wpm for part one and 182.26 wpm for part 2. Although part two was slightly faster on the speech rate, the word length for part one (1108 words) was marginally longer than for part two (1048 words). Similarly, the overall duration for part one was 6’14” and 5’45” for part two so both parts were very similar. These speeds were closer to
Brindley and Slayter’s (2002) ‘normal’ speed texts, which were 180 wpm, than the ‘average’ speech rate of 188 wpm proposed by Griffiths (1992) and also more or less in line with the typical speed of lecture delivery, which is between 100-180 wpm (Robinson et al., 1997, p. 260). Although Ladas (1980, in Robinson et al., 1997, p. 260) argues that 135 wpm is the fastest speed at which students can take notes, the fact that the test takers would hear the lecture twice was intended to mediate this.

d. The exploratory study results also highlighted another aspect which had not been adequately addressed in the exploratory study, namely the linguistic features of the input (see Section 3.5.1.2). When these were analysed after the study, using the same 13 analyses as had been used for the test takers’ results, there was some considerable variation between the different parts. Table 3.18 below shows the results.

As can be seen in Table 3.18, there is quite a difference across the four measures. First, for the descriptive statistics, the number of paragraphs ranged from one to seven. However, this may simply have been down to the way the paragraphing was done; in any case, this measure is less important than the next measure, which shows that the sentence count ranged from 12 in part one to 40 in part four. Similarly, there is a large difference in the word count across the four parts, ranging from 423 words in part one to 587 in part four. The number of sentences in a paragraph varies considerably, too, from 3.14 to 14; however, as with the number of paragraphs, this could be down to the layout of the text. The sentence length is also wide ranging, with the minimum at 14.28 words per sentence, on average, to 35.25 words per sentence, on average. The number of syllables per word and the number of letters per word are both similar across the four parts of the essay, ranging from 1.29 syllables to 1.37, and from 3.98 letters to 4.18 respectively.

Regarding the type-token ratio, there was also some variation, ranging from 0.33 to 0.42. Similarly, both the noun phrase density and the verb phrase density varied considerably, with the former ranging from 316.26 to 371.38 and the latter ranging from 180.58 to 246.94. The
agentless passive voice density revealed that two parts (two and four) containing no agentless passives and two parts (one and three) with results of 4.73 and 5.11 respectively.

Table 3.18: Linguistic features of input

<table>
<thead>
<tr>
<th>Label</th>
<th>Full description</th>
<th>Part one</th>
<th>Part two</th>
<th>Part three</th>
<th>Part four</th>
</tr>
</thead>
<tbody>
<tr>
<td>DESPC</td>
<td>Paragraph count, no. of paragraphs</td>
<td>2</td>
<td>1*</td>
<td>7**</td>
<td>7**</td>
</tr>
<tr>
<td>DESSC</td>
<td>Sentence count, no. of sentences</td>
<td>12*</td>
<td>14</td>
<td>22</td>
<td>40**</td>
</tr>
<tr>
<td>DESWC</td>
<td>Word count, no. of words</td>
<td>423*</td>
<td>449</td>
<td>587**</td>
<td>571</td>
</tr>
<tr>
<td>DESPL</td>
<td>Paragraph length, no. of sentences in a paragraph, mean</td>
<td>6</td>
<td>14**</td>
<td>3.14*</td>
<td>5.71</td>
</tr>
<tr>
<td>DESSL</td>
<td>Sentence length, mean no. of words</td>
<td>35.25**</td>
<td>32.07</td>
<td>26.68</td>
<td>14.28*</td>
</tr>
<tr>
<td>DESWLsy</td>
<td>Word length, no. of syllables, mean</td>
<td>1.37**</td>
<td>1.36</td>
<td>1.29*</td>
<td>1.35</td>
</tr>
<tr>
<td>DESWLlt</td>
<td>Word length, no. of letters, mean</td>
<td>4.14</td>
<td>4.16</td>
<td>3.98*</td>
<td>4.18**</td>
</tr>
<tr>
<td>LDTPRa</td>
<td>Lexical diversity, type-token ratio, all words</td>
<td>0.42**</td>
<td>0.38</td>
<td>0.33*</td>
<td>0.36</td>
</tr>
<tr>
<td>DRNP</td>
<td>Noun phrase density, incidence</td>
<td>371.16</td>
<td>316.26*</td>
<td>371.38**</td>
<td>332.75</td>
</tr>
<tr>
<td>DRVP</td>
<td>Verb phrase density, incidence</td>
<td>198.58</td>
<td>229.40</td>
<td>180.58*</td>
<td>246.94**</td>
</tr>
<tr>
<td>DRPVAL</td>
<td>Agentless passive voice density, incidence</td>
<td>4.73</td>
<td>0.00*</td>
<td>5.11**</td>
<td>0.00*</td>
</tr>
<tr>
<td>RDFRE</td>
<td>Flesch Reading Ease</td>
<td>54.82***</td>
<td>58.97</td>
<td>70.62</td>
<td>77.80***</td>
</tr>
<tr>
<td>RDFKGL</td>
<td>Flesch-Kincaid Grade level</td>
<td>14.37**</td>
<td>13.00</td>
<td>10.04</td>
<td>5.95*</td>
</tr>
</tbody>
</table>

* denotes the lowest value for the measure

** denotes the highest value for the measure

*** Note that the lower the Flesch Reading Ease number, the more difficult the text; consequently, the easier part has the higher number and vice versa

As a result of the above analyses, there was a wide variation in the Flesch Reading Ease and the Flesch Kincaid Grade Level. The former ranged from a relatively simple 77.80 to a moderately complex 54.82, while the latter ranged from 5.95 to 14.37. However, both the linguistically easiest (part four) and the linguistically most complex (part one) parts, were the two which test takers found most difficult.
Having said that, as shown in Table 3.18 above, parts one and four have the highest number of highest values with five and four respectively. This could potentially account for why these sections were more difficult for the test takers. On the other hand, part three, which was one of the easier parts also has four of the highest values and five of the lowest. Part two, which was also one of the easier parts has only one of the highest values and three of the lowest. These differences may have skewed the results of the exploratory study. This was addressed in the main study by amending the content to two halves to have the same type of input in each: one slide with one sentence, one slide with a diagram, and one slide with five bullet points. The presentation order of the different slides was also reversed such that in part one, the order was as above but in part two, the order was one slide with five bullet points, one slide with a diagram and one slide with a sentence. The linguistic complexity of both halves was measured using Text Inspector for the main study. Details are presented in Section 4.4.2.2 below.

3.5.2.3.2 Output (EAP essay)

Again, based on the results of the exploratory study and the feedback questionnaire, a number of changes were made ahead of the main study.

e. The amount of information provided in the test booklet as scaffolding for the note-taking task was changed. In the exploratory study, guided notes pages (see Appendix 4) informed the test takers what to take notes on. This was felt to have been too much scaffolding for participants at the B1 or B2 level and may have made the task unrealistically easy. Removing the 'hints' would make the task more authentic but would also allow test takers to decide for themselves what were the salient points of the input that they wanted to note down. This would possibly ease the pressure of having to write so much in the output task.
a. The one change that I would have liked to make but couldn’t for reasons of practicality was the timing. As a number of participants said they would have liked more time, it would have been good to increase the time available; however, due to time constraints at the universities where the main study data was being collected, this was not possible. Instead, I reduced the word count requirement from 500 words to 350.

b. The exploratory study included notes pages with headings and bullet points to provide scaffolding to test takers regarding what points to make notes on. This was intended to support lower-level test takers; however, this, along with the essay plan provided, seemed to be too much scaffolding as the test did not differentiate very well across levels. Furthermore, giving such support seemed inauthentic as test takers would not get such support in a real-world academic environment. For the main study, notes pages were still provided but no input was provided regarding what test takers should take notes on except for the essay title, which was made known at the beginning of the test and which test takers had access to throughout.

c. Similarly, the exploratory study included a very detailed outline for the essay. The intention had been to assist low-level learners with the content and planning. However, this may have provided too much scaffolding and, given the level that the test is aimed at (B1/B2 on the CEFR), this reduced to a great extent the need for ‘organisation’ as a cognitive process (Field, 2004, p. 329). At B2 level, learners are expected to be able to ‘write clear, detailed texts on a variety of subjects, … synthesising and evaluating information and arguments from a number of sources’. Planning is also a salient feature which comes in at B2 on the CEFR: ‘can plan what is to be said and the means to say it …’ (Council of Europe, 2001, p. 64), so I decided to remove the scaffolding to better distinguish between the different levels of performance. Those at B2 and above should be able to plan what to say and have the language to paraphrase the content of the lecture whereas these skills would be expected to be less well-developed at the
lower levels. This should mean that raters can distinguish between the different levels of performance in line with Scardamalia and Bereiter (1987) (see Section 2.5.1.2). Having an outline would therefore mean that everyone (both mature and immature writers) were given the plan so it would be more difficult to differentiate between the more skilled writers and the less skilled ones. This change should therefore manifest itself in the EAP task through variations in performance from repeated, verbatim-cited chunks from the lecture through to better flowing, more coherent texts in which the writers have paraphrased the content of the lecture and structured the content so as to achieve the goal of responding to the different parts of the prompt without reproducing all of the lecture content verbatim.

Further to the changes outlined above, the materials were redesigned as necessary for the main study.

3.6 Conclusion
To conclude, this chapter discussed the research approach used in the exploratory study. The methodology, results and discussion of the exploratory study were presented, along with an outline of the changes made ahead of the main study.

Chapter four will present the Methodology for the main study, giving details of the participants, the instruments used, the research questions and how the data will be analysed to answer the research questions.
4. Main study: Methodology

4.1 Participants

The participants in this study comprised university students from two universities, one in Russia and one in Ukraine: both countries with a post-Soviet educational culture.

Higher education in Soviet Russia was highly regarded and the Soviet Union was seen as the ‘destination of choice’ for many higher education students (ICEF Monitor, 2013), yet in the post-Soviet era, both countries have had a relatively low profile and have fared poorly in international comparisons. Consequently, there has been a significant push in recent years in post-Soviet countries to internationalise higher education (ICEF Monitor, 2013).

Russia plans to increase its international student numbers from 200,000 in 2017 to 710,000 by 2025. In 2017, Russia was ranked 6th ‘in terms of export of educational services, after the United States, the United Kingdom, France, Germany and Australia’ (Medvedev, 2017, quoted in Vorotnikov, 2017) and plans to increase its revenue for Russian universities by over 400 per cent between 2017 and 2025, from RUB84 bn to RUB373 bn (Vorotnikov, 2017). The intention is that the ‘influx of foreign students may raise the prestige and image of Russian universities in the international arena’ (Serdyakov, 2017, quoted in Vorotnikov, 2017).

In order to achieve these ambitious plans, academic writing and the delivery of teaching through English are high on the agenda as the Ministries of Education in both Russia and Ukraine strive to develop a more international education system. For this reason, both of the institutions selected are taking measures to develop English medium instruction.

The Russian university selected for this study already offers a wide range of exchange programmes and employs a number of visiting lecturers who deliver courses in English for both students and faculty. Consequently, it is important for these students to be able to follow lectures and write follow-up essays in English.
Like the Russian system, the Ukrainian higher education is ‘largely unreformed since Soviet times’ (Tatsenko, 2014). In 2014, the then new education minister, Serhiy Kvit, published a report stressing that his priorities were to reform higher education and to improve the levels of English in the university system (Kvit, 2014, cited in Tatsenko, 2014). Furthermore, a draft plan from the Ukrainian Ministry of Education called for ‘at least one foreign faculty member in every university department by 2020’ (Ukrainian Ministry of Education, 2015, cited in ICEF Monitor, 2015).

Internationalisation is a key topic on the Ukrainian government’s agenda and there has also recently been a changing trend with more and more Ukrainian students wanting to study abroad (ICEF Monitor, 2015). On the other hand, as the Ukrainian higher education system is becoming ‘more open and aligned with international standards and perspectives’, there has been a substantial increase in international student numbers in Ukraine, rising from 56,664 in 2011 to 75,605 students from 154 countries in 2018, with India contributing the highest number of international students at 14958 (Ministry of Education and Science of Ukraine State Enterprise Ukrainian State Center for International Education, 2017).

While the numbers are small compared to the United States, the United Kingdom, Australia and, indeed, Russia, there is a clear upward trend. Moreover, while the number of international students from post-Soviet countries has decreased from 29.47 per cent (n=19,539) in 2017-18 to 24.67 per cent (n=18,648) in 2018-19, there has been an increase in students from other countries, rising from 70.53 per cent (n=46,771) in 2017-18 to 75.33 per cent (n=56,957) in 2018-19 (Ministry of Education and Science of Ukraine State Enterprise Ukrainian State Center for International Education, 2017). Similarly, the number of Ukrainian universities accepting international students has risen from 185 in 2015-16 to 230 in 2017-18. This figure almost doubled in the 2018-19 academic year to 443 (Ministry of Education and Science of Ukraine State Enterprise Ukrainian State Center for International Education, 2017) and there is likely to be further demand for incoming and outgoing student and staff mobilities.
in the future (ICEF Monitor, 2015). All of these factors contribute to the need for Ukrainian students to be able to function in English in the academic environment.

Like the Russian Federation university, the Ukrainian university selected also has a number of exchange programmes with universities abroad and often invites foreign lecturers to give classes in English so, again, it is imperative that these students are able to follow lectures and produce work based on these lectures.

University students were chosen because, as with the exploratory study, the main study aimed to investigate academic language use in an integrated EAP test of listening-into-writing.

The language levels of the participants covered a range of language levels (see Table 4.1). Although the test was, in principle, aimed at CEFR levels B1 / B2, students with a range of levels were sought for the study to enable validation of the test at the target level but also to demonstrate how it differentiates across levels. In other words, if the lower-level students had been able to achieve a high CEFR level on the test, it would demonstrate that, in fact, the test was not aimed at B1/B2 but lower. Table 4.1 shows the demographics. In total, data was collected from 131 participants. Five participants did not sign the consent form while three participants did not write anything in the EAP task so they were all excluded from the analysis. For seven other students, the data collection was carried out incorrectly so they were also excluded from the study.

As a result, data from 116 participants was included in the study: 74 from the Russian university and 42 from the Ukrainian university. 57 participants from the Russian university gave information about their nationality: 56 were Russian and one was Tatar. 17 did not answer this question. All 42 participants from the Ukrainian university reported that they were Ukrainian. In terms of the gender balance, there were 26 males and 88 females. Two people did not respond to this question. 102 of the participants were in the 18-21 age range and five were in the 22-25 age bracket. Nine participants did not respond to this question. The
participants were asked to assess their own language level and responses ranged from A1 to C1. 23 participants did not estimate their level but the two largest groups estimated that they were at B1 or B2 level (29 and 26 participants respectively). 17 participants estimated that they were A1, 16 participants self-assessed their level as A2, and another nine said they were C1.

32 students reported that they were in the first year of their degree programme and 35 in their second year. 49 participants did not give details of their year of study but it is likely that they were also first or second years. Instead of the year of study, some participants mentioned their degree programmes: 12 were doing a Chemistry degree, eight were studying Turkish while seven were studying Chinese. Six participants were doing a degree in Computer Security and four were students of Japanese. Two students were studying Indonesian, another two were doing degrees in Biology and one person was an IT student. 73 participants did not give details of their degree programme. All of the students at both universities had English classes as part of their degree programmes.
### Table 4.1: Background information of the test takers

<table>
<thead>
<tr>
<th>Nationality</th>
<th>Gender</th>
<th>Age</th>
<th>Self-assessed language level</th>
<th>Year of study</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF uni</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RF uni</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UA uni</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. RF uni = participants from Russian Federation university
2. UA uni = participants from Ukrainian university
4.2 Ethical consent

In accordance with the Ethics requirements of the University of Bedfordshire and the universities where the data was being collected, all participants were informed about the purpose of the study via an information sheet which they were asked to read before they participated and, if they were happy for their data to be used, they were then asked to sign the consent form provided. They were informed that they could have access to their results and to the results of the study by contacting the researcher via the e-mail address provided. See Appendix 13 for the Information Sheet and Consent Form.

4.3 Research Questions and hypotheses

4.3.1 Research Questions

As the focus of study had changed from comparing performance on an independent listening and an independent writing task to how participants performed based on the input format, as detailed in Section 3.5 above, I decided that it was no longer necessary to have the independent tasks. The general level of proficiency would still be measured through the Oxford University Press Quick Placement Test (Oxford University Press, 2001); however, this would serve merely to split the participants into two groups of roughly equal ability.

This study aimed to answer three RQs. These are:

1. How does performance on an integrated listening-into-writing EAP task differ when test takers are presented with lecture input in the form of a podcast or a vodcast?
2. To what extent are test takers influenced by the written word in an EAP lecture compared to the spoken word?
3. How do test takers’ perceptions of their performance on an integrated EAP listening-into-writing test presented in two different input formats compare to their observed scores as awarded by trained raters?
4.3.2 Expectations

For RQ 1, I anticipated that, when presented with a lecture in an audio-only format (podcast) and an audiovisual format (vodcast) including the use of PowerPoint, students would perform better, that is, they would be awarded a better score on the half of the summary which relates to the textual and aural information than on the half of the summary where they had only audio input. There are several reasons to assume this: one is that listeners may be unfamiliar with the speaker’s accent so reading the words on the PowerPoint slide may help them to decode a part or all of the spoken content, which would otherwise be incomprehensible (Robin, 2007; Vandergrift, 2004). Another reason why the aural only input may be more difficult is that not having a video to watch may mean a drop in concentration after a while (Goh, 2000). This may happen after around three minutes according to James (1986, p. 41 in Long, 1990, p. 68-69). Other reasons for difficulties in listening comprehension include the limited capacity of short-term memory (Goh, 2000), speech rate (Brindley and Slayter, 2002), a lack of vocabulary (Matthews and Cheng, 2015; Stæhr, 2009), aural decoding (Goh, 2000; Leonard, 2019) or anxiety (Elkhafaifi, 2005; Field, 2015; Zhang, 2013).

With regard to RQ2, I expected that lower-level test takers (as measured by the lexicogrammatical test) would reproduce fewer propositions overall than stronger participants due to their lack of vocabulary (Matthews and Cheng, 2015; Stæhr, 2009) and inability to decode the input (Goh, 2000; Leonard, 2019). However, I expected that the lower level test takers would be more likely to be influenced by the content of the PowerPoint slides than higher level test takers and would therefore have more exact matches with the PowerPoint slides than the higher-level test takers, who would be more likely to paraphrase the information from the input. This would be in line with findings from Keck (2014) who found that lower-level test takers relied more on source texts than their higher-ability classmates. I also hypothesised that intermediate-level test takers would have the most matches to the input texts (both written and spoken) as they would be in a position to understand more than the
lower-level test takers but not be as able as the higher-level test takers to paraphrase the information. This would be in line with the thinking on second language acquisition research and developmental steps (Dörnyei and Skehan, 2003, p. 597) and Field’s model of listening processes (Field, 2008, p. 114) outlined in the Literature Review above.

4.4 Instruments

The test takers carried out the following tasks:

1. Lexico-grammatical placement test
2. Integrated EAP listening-into-writing task
   a. Listening to and taking notes on a lecture on the topic of ‘Culture Shock’ provided in two formats (audio/audiovisual or audiovisual/audio)
   b. Writing a summary on the content of the lecture
3. Post-task feedback questionnaire

Table 4.2 below outlines the tasks that test takers completed, what the task comprised, how this task contributes to the research questions and how the information was analysed to answer the research questions.
Table 4.2: Data collection methods and analyses

<table>
<thead>
<tr>
<th>What did the participants do?</th>
<th>What does the task comprise?</th>
<th>What data came out of the task?</th>
<th>Which research question was being answered?</th>
<th>How was the data analysed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete a lexicogrammatical placement test</td>
<td>Answering 60 multiple-choice questions</td>
<td>Data on students’ general language ability</td>
<td>None – to split students into groups of similar language ability</td>
<td>Quantitatively: Descriptive statistics Independent samples ( t )-test</td>
</tr>
<tr>
<td>Integrated EAP listening-into-writing task</td>
<td>Listening to academic lecture in two formats: audio and audiovisual plus PowerPoint Taking notes on lecture Writing a summary based on lecture</td>
<td>Evidence of ability to understand an academic lecture, select key points and supporting points and organise information into a summary.</td>
<td>RQ1 and RQ2</td>
<td>Quantitatively (RQ1): Inter-rater reliability, rater harshness and Fair Average scores for each test taker (TT) on each part of summary using Rasch analysis Hotelling’s ( T^2 )-test carried out using Fair Average scores Text Inspector: no. of words (token count); no. of sentences; sentence length; avge no. of syllables per word and sentence; type count; type-token ratio; lexical diversity; Flesch Reading Ease; % of types and tokens from CEFR levels A1-C2 (Cambridge University Press, 2015); % of types and tokens from the Academic Word List (AWL) (Coxhead, 2000) Descriptive statistics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Qualitatively (RQ2): Propositions and key words reproduced coded and analysed for matches with a) transcript and b) presentation slides; More detailed qualitative analysis of 30 papers to investigate percentage of key propositions reproduced from audio input compared to slides</td>
</tr>
</tbody>
</table>
| Post-task questionnaire | Answering 5 Likert-scale, 2 multiple-choice, 2 multiple-response and 3 open-ended questions | Test-takers’ perceptions of test | RQ3 | Quantitively (RQ3):  
To collect  
TTs’ perceptions of which input format they performed better on  
background information on TT’s validation evidence for instruments  
Qualitatively (RQ3):  
To investigate:  
TTs’ reasons for perceptions of their best and worst performances  
construct-irrelevant aspects which could have affected performance |
4.4.1 Placement test

4.4.1.1 Placement test: Rationale

As outlined in Section 3.4.6 above, the study changed from investigating performance on an independent IELTS writing test compared to performance on an integrated EAP test to investigating which input medium led to better results. The IELTS tests took over an hour to complete (see Section 3.3.2) so this may have contributed to a fatigue effect during the exploratory study. Therefore, to reduce the risk of fatigue and given that such extensive tests were not necessary to split the test takers into two groups as well as the change of focus, I decided these tests should not be used in the main study. However, I considered it important to have some measure of the participants’ proficiency before they took the EAP test, so teachers selected test takers at a range of levels and participants also self-assessed their language level. Nonetheless, I still felt it was wise to also get an idea of their language ability through a placement test to ensure that the study comprised a wide range of linguistic abilities and was used to separate the students into two groups of approximately equal ability so that any differences in performance were a result of the EAP test rather than differences in underlying proficiency. As lexis has been shown to play an important role in both listening (Bonk, 2000; Stæhr, 2008; 2009; Van Zeeland and Schmitt, 2013) and writing (Laufer and Nation, 1995; Stæhr, 2008), I decided to use a lexicogrammatical placement test as these can give a quick estimate of test taker proficiency. I had already been using the Oxford University Press Quick Placement Test (Oxford University Press, 2001) as a placement test on the course I ran at the university for many years so I had experience of its benefit as a quick yet reliable way to measure language proficiency. To start with then, the test takers completed the Oxford University Press Quick Placement Test (QPT) to measure their general English proficiency.

4.4.1.2 Placement test: Design

The QPT is a 30-minute, 60-item, lexicogrammatical placement test published by Oxford University Press. An off-the-shelf test was used as it had already been trialled and validated.
whereas designing a new placement test was not the focus of this thesis and would therefore have been beyond the scope of this research project.

**4.4.1.3 Placement test: Procedure**

Although the QPT is quite an old test, published in 2001, this was chosen as it is quick and easy to administer and mark. This was important as the results had to be turned around quickly so that the students could be split into groups to do the main tasks.

Because the classrooms lacked computers, the test was administered in pen and paper mode. Once the students had completed the test, the papers were quickly marked and the test takers were divided into two groups of approximately equal language level for the main task.

**4.4.1.4 Placement test: Scoring**

The placement test was marked by hand by a team of raters using the overlay answer key provided with the test. Each correct answer scored one point and a score out of 60 was allocated to each test taker.

**4.4.1.5 Placement test: Quantitative analysis**

As outlined in Table 4.2, the placement test results were analysed to calculate descriptive statistics (mean, median, mode, standard deviation, minimum, maximum, range, skewness and kurtosis) for the two groups (audio first and video first).

Next, an independent samples $t$-test was run to establish whether there was any significant difference between the means for the two groups that the students were divided into on the basis of the QPT score.

Independent samples $t$-tests are used to establish whether there are any statistically significant differences between the means of two categorical or ordinal independent groups with one continuous, dependent variable (Laerd, 2015). In the present study, the independent variable was the two groups. The first of these is the ‘audio first’ group and the second is the ‘video first’ group. The dependent variable was the QPT score.
4.4.2 Integrated EAP listening-into-writing test

4.4.2.1 Integrated EAP listening-into-writing test: Rationale

The second task was the main instrument for this study. The research questions focus on how test takers perform in an integrated listening-into-writing test.

As discussed in the Literature Review (see Chapter 2), a number of studies have investigated the effect of different types of input formats on discrete listening tests but very few have looked at integrated tests of listening-into-writing. This seems a little strange given that authentic listening – at least in an EAP environment - often involves other skills such as speaking or writing as recipients are required to respond to the audio and/or audiovisual input either orally or in writing, for example, as part of a discussion in a seminar group, to ask questions in a lecture or to use the lecture input to help with a written assessment.

Nowadays, as input is increasingly provided to students to access outside of the classroom, this raises questions of whether it is more helpful for students to have access to audio only input (e.g. in the form of a podcast) or video input (in the form of a vodcast) and which format is most useful for the students when faced with a follow-up task based on the input. By presenting the test-takers with an integrated test task which involved them interacting with both forms of input and producing a written task based on the input, I was able to investigate which of the input types allowed test-takers to perform better (RQ1). The second research question investigates the extent to which test takers are influenced by the written word when presented with a vodcast including PowerPoint slides. This was important as it may have implications for the construct which is being tested in an integrated listening-into-writing test.

4.4.2.2 Integrated EAP listening-into-writing test: Design

In accordance with Bachman and Palmer’s recommendation that tests should replicate the target language use (TLU) domain (Bachman and Palmer, 1996), an integrated listening-into-writing task was used as this replicates the TLU domain and also adheres to Bachman and Palmer’s (1996) Qualities of Test Usefulness in terms of both authenticity and interactiveness.
The extent to which the task demonstrates construct validity was investigated as part of the analysis of the questionnaire data as outlined above.

For this task, a recording was used of a short lecture delivered by the researcher on the topic of Culture Shock. This lecture was recorded both as a podcast (audio only) and as a vodcast (with video support in the form of PowerPoint slides and film of the lecturer). The PowerPoint slides can be found in Appendix 17 while the transcript of the lecture can be found in Appendix 18.

The whole lecture lasted 12’42” minutes per playback. The first 33 seconds comprised the introduction and outline and the final nine seconds was the close of the lecture. The main body of the lecture therefore occupied 12 minutes. This was divided into two halves of approximately equivalent content in terms of linguistic complexity and length. The first part lasted 6’14” while the second part lasted 5’46”.

Apart from the first slide at the beginning of the lecture, which showed the outline, each part comprised three slides: one with one sentence, one with five bullet points, and one with a diagram showing five keywords that were discussed. This was done to make the two halves as comparable as possible while at the same time providing variation in the way the visual input was presented.

I decided to use the different ways of presenting the content on the slides to see how this affected the test takers’ use of the slides and their perceptions of the influence of the slides. By using a slide with one point on, I was able to investigate whether the test takers copied the whole sentence or not; the slide with five bullet points allowed me to see whether the test takers just noted the key points on the slide or wrote down the spoken words (which were different to the key points on the slide) and if they wrote down any other information related to these points or not; finally, the slide with the diagram was to investigate whether test takers
copied the diagram or listened and noted the spoken words. This information was used to answer RQ2.

Before running a textual analysis on the data to establish the degree of similarity between the two halves of the input, I cleaned the data to remove repeated words or sounds (‘we ... we’re probably ...’) and repeated word beginnings (‘but it’s ... it’s not ...’) but hesitations (er; erm) were left in. The introduction (0’33”) and the lecture close (0’9”) were both removed and the main content of the lecture was divided into two texts in accordance with the two halves of the input. The two texts were then analysed using Text Inspector (Text Inspector, 2019). Text Inspector is an online textual analysis tool, which provides analyses based on descriptive statistics such as length and number of sentences, number of words and syllables per sentence, type/token ratio, various corpora (English Vocabulary Profile (Cambridge University Press, 2015), which also allocates a CEFR level to each word, the British National Corpus (BNC) (University of Oxford, 2015) and the Corpus of Contemporary American English (COCA) corpus (Davies, n.d)). It also analyses input and produces measures of the types and quantities of words from the various lists in the Academic Word List (Coxhead, 2000). As outlined in Section 3.5.1 above, the decision to use this tool in the main study was primarily because it analyses texts based on the CEFR levels and also according to the AWL.

The results show that both halves were very similar. The results of the transcript analysis are outlined below and can be seen in tabular form, along with the results of the analysis of the language on the lecture slides (see Appendix 19). A more detailed analysis of the Text Inspector CEFR level analyses for parts one and two can be found in the corresponding Appendices 20 and 21. Table 4.3 below shows the Descriptive Statistics and the Lexical Diversity measures.

As can be seen from the descriptive statistics above, part one contained 1108 words in 34 sentences with an average of 43.97 syllables per sentence and 32.59 words per sentence. In contrast, part two comprised 1048 words in 31 sentences with an average of 45.19 syllables...
per sentence and 33.81 words per sentence. Part one had 1.35 syllables per word on average; this was very similar to part two, which had 1.34 syllables per word on average. Part one had a Flesch Reading Ease level of 59.61 compared to 59.43 for part two. This resulted in a Flesch-Kincaid Grade Levels of 13.04 and 13.37 for parts one and two respectively. There were 290 types in the first part and 277 in the second part. Taking into account the slight difference in length between the two texts, this was a type/token ratio for both part one and part two of 0.26. Lexical diversity\(^6\) as measured by MTLD\(^7\) was very similar across the two texts: part one had an MTLD measure of 39.76 compared to 40.25 in part two.

Table 4.3: Descriptive Statistics and Lexical Diversity measures: transcript parts one and two

<table>
<thead>
<tr>
<th>Operation</th>
<th>Data</th>
<th>Lecture part 1</th>
<th>Lecture part 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics</td>
<td>Sentence count</td>
<td>34</td>
<td>31</td>
</tr>
<tr>
<td>Statistics</td>
<td>Token count</td>
<td>1108</td>
<td>1048</td>
</tr>
<tr>
<td>Statistics</td>
<td>Type count</td>
<td>290</td>
<td>277</td>
</tr>
<tr>
<td>Statistics</td>
<td>Type/token ratio</td>
<td>0.26</td>
<td>0.26</td>
</tr>
<tr>
<td>Statistics</td>
<td>Average syllables per sentence</td>
<td>43.97</td>
<td>45.19</td>
</tr>
<tr>
<td>Statistics</td>
<td>Average syllables per word</td>
<td>1.35</td>
<td>1.34</td>
</tr>
<tr>
<td>Statistics</td>
<td>Flesch Reading Ease</td>
<td>59.61</td>
<td>59.43</td>
</tr>
<tr>
<td>Statistics</td>
<td>Flesch-Kincaid Grade</td>
<td>13.04</td>
<td>13.37</td>
</tr>
<tr>
<td>Statistics</td>
<td>Average Sentence Length</td>
<td>32.59</td>
<td>33.81</td>
</tr>
<tr>
<td>Lexical Diversity</td>
<td>Lexical diversity (MTLD)</td>
<td>39.76</td>
<td>40.25</td>
</tr>
</tbody>
</table>

With regard to the CEFR level of the input text as measured by the ‘Lexis: EVP’ measure in Text Inspector, which is based on English Vocabulary Profile (Cambridge University Press, 2015), the results can be seen in Table 4.4 below. While there were more A1 level types in part two, there were more A2 level types in part one. There was a slightly higher percentage of B1 types in part two but otherwise, the percentages were very similar, except for the

\(^6\) Lexical diversity is a measure of text difficulty which takes into account ‘flexibility’ and ‘vocabulary richness’ (Read, 2000, in Durán et al., 2004: 221) in evaluating the complexity of a text. The greater the range of words in a text, the higher the diversity measure (McCarthy & Jarvis, 2010: 381)

\(^7\) MTLD (Measure of textual lexical diversity) is a lexical diversity measure which is independent of text length and which is calculated as ‘the mean length of word strings that maintain a criterion level of lexical variation’ (McCarthy & Jarvis, 2010: 381)
Unlisted types, which comprised 11.41 per cent in part one compared to only 6.29 per cent in part two. However, this was due to words such as the names of the researchers (‘Dr Hofstede’ and ‘Oberg’) and the names of the stages of culture shock (‘disintegration’, ‘reintegration’, ‘autonomy’). In both parts, some of the Unknown words included contractions (such as ‘we’ve, ‘you’re’ or ‘you’ve’) and fillers (such as ‘erm’). On the other hand, the second part included Unknown words such as ‘flatmates’, ‘embedded’ and ‘minimise’.

Table 4.4: CEFR levels – Type percentages for lecture transcript parts one and two

<table>
<thead>
<tr>
<th>Operation</th>
<th>Data</th>
<th>Lecture part 1</th>
<th>Lecture part 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexis: EVP</td>
<td>A1 type %</td>
<td>45.97</td>
<td>52.45</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>A2 type %</td>
<td>20.47</td>
<td>16.43</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>B1 type %</td>
<td>10.07</td>
<td>13.99</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>B2 type %</td>
<td>8.72</td>
<td>8.74</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>C1 type %</td>
<td>2.01</td>
<td>1.4</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>C2 type %</td>
<td>0.67</td>
<td>0</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>Known Words type %</td>
<td>0.67</td>
<td>0.7</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>Unlisted type %</td>
<td>11.41</td>
<td>6.29</td>
</tr>
</tbody>
</table>

For the token measures, the results were again very similar. Table 4.5 shows the Token measures for the EVP CEFR levels.

Table 4.5: CEFR levels - Token percentages for lecture transcript parts one and two

<table>
<thead>
<tr>
<th>Operation</th>
<th>Data</th>
<th>Lecture part 1</th>
<th>Lecture part 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexis: EVP</td>
<td>A1 token %</td>
<td>64.09</td>
<td>67.71</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>A2 token %</td>
<td>13.76</td>
<td>11.54</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>B1 token %</td>
<td>4.71</td>
<td>5.06</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>B2 token %</td>
<td>3.68</td>
<td>3.44</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>C1 token %</td>
<td>0.85</td>
<td>0.91</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>C2 token %</td>
<td>0.19</td>
<td>0</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>Known Words token %</td>
<td>3.3</td>
<td>6.07</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>Unlisted token %</td>
<td>9.43</td>
<td>5.26</td>
</tr>
</tbody>
</table>

The percentage of A1 tokens was slightly higher in part two whereas the percentage of A2 tokens was slightly higher in part one. Again, the other measures were all very similar, except
for the percentages of known words, which was almost twice as high in part two than part one and the percentage of unlisted tokens, which was almost twice as high in part one than part two.

The final analyses which will be discussed are the percentage of AWL types and tokens. Table 4.6 shows the results of these analyses.

Table 4.6: AWL analyses – Types / token percentages: lecture transcript parts one and two

<table>
<thead>
<tr>
<th>Operation</th>
<th>Data</th>
<th>Lecture part 1</th>
<th>Lecture part 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexis: AWL</td>
<td>AWL All Types %</td>
<td>5.83</td>
<td>7.48</td>
</tr>
<tr>
<td>Lexis: AWL</td>
<td>AWL All Tokens %</td>
<td>4.98</td>
<td>8.2</td>
</tr>
</tbody>
</table>

As shown in Table 4.6, the percentage of AWL types was a little higher in part two at 7.48 per cent (n=22) compared with 5.83 per cent (n=18) in part one. There was also quite a big difference in AWL tokens with 8.2 per cent (n=79) in part two compared to 4.98 per cent (n=51) in part one. However, a closer inspection of the content revealed that this was mainly due to the word ‘culture’, which appeared 25 times in part one but 45 times in part two, and ‘integrate’, which appeared 6 times in part two but not at all in part one.

By making the two halves of the lecture as similar as possible in terms of length, speed of delivery and linguistic complexity, it was hoped that this would remove as far as possible the differences between the input formats which the exploratory study had highlighted.

In line with several studies which have found that clearly signposted discourse is useful for both L1 and L2 students (Chaudron and Richards, 1986; Rickards et al., 1997), the discourse was clearly signposted to enable test takers to follow the content to the best of their ability.

Following recommendations by Buck (2001) that a listening text should be played twice in a testing situation, I again allowed the recording to be played twice with a short pause between
playbacks (see Section 4.4.2.3 below). During the playbacks, participants would take notes on the content to help them with the next part of the task.

The second part of the task required participants to use the information they had heard and / or noted down to write a handwritten summary entitled ‘What is Culture Shock and how can you reduce the effects of culture shock?’. They had 40 minutes to write a 350-word summary of the content of the lecture. Although longer than the 250 words required in an IELTS Academic writing Task 2, this was considered manageable for B2 level learners as participants had been provided with source content and was still somewhat shorter than the 500 words required in the exploratory study. Participants completed the task using pen and paper.

4.4.2.3 Integrated EAP listening-into-writing test: Procedure
As mentioned above, for this task, the test takers were split into two groups. This was to enable the implementation of an ‘AB-BA’ Counterbalanced Measures Design. Mackey and Gass (2005, p. 353) define a counterbalanced design as ‘an experimental design in which the ordering of test items or tasks is different for different participants or groups of participants’. This is to avoid the potential for order effects which can arise from repeated measures designs (University of North Carolina, 2008). As shown in Table 4.7, both groups heard the same audio input but the lecture was divided in half such that each group heard only the podcast for one half and viewed the vodcast for the other half.

<table>
<thead>
<tr>
<th></th>
<th>Lecture part one</th>
<th>Lecture part two</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Audio first’ group</td>
<td>Podcast (audio only)</td>
<td>Vodcast (audio and video)</td>
</tr>
<tr>
<td>‘Video first’ group</td>
<td>Vodcast (audio and video)</td>
<td>Podcast (audio only)</td>
</tr>
</tbody>
</table>

Table 4.7: EAP lecture: order of input
4.4.2.3.1 Integrated EAP listening-into-writing test: Administration

Further to completing the QPT described above, test takers were split into two groups – an ‘audio first’ group and a ‘video first’ group. As described above, this was to enable the research to be carried out as a Counterbalanced Measures Design to avoid any order effect.

Participants were given the task booklet, which comprised the instructions for the task, space to write notes on the two halves of the test, the summary title, instructions for writing the summary and pages on which to write the summary. Participants were asked to read the instructions and to ask questions if anything was unclear.

When everyone had read the instructions and was ready to start, the invigilator played the MP4 of the lecture. For the ‘audio first’ group, the participants received the first half of the lecture content as audio input and the second half of the lecture as audiovisual input while the ‘video first’ group were given the first half of the lecture as audiovisual input and the second half as audio input.

In line with the literature on double playback mentioned in the Literature Review (see Section 2.3.3.1) above, test takers listened to/watched the lecture twice and took handwritten notes on the content so they could use the information to write a summary on the topic of culture shock afterwards. Participants watched the video as a group so the lecturer running the test had control of the recording.

After the first playback, participants had two minutes to go through their notes before the second playback. After the MP4 had been played twice through, participants had 40 minutes to write a 350-word summary of the lecture content.

4.4.2.4 Integrated EAP listening-into-writing test: Scoring

To develop a model answer, an expert panel of three native speaker raters (of whom I was one) watched the lecture and took notes in the same way as the participants were required to do. The three raters then compared the salient points they had noted down and agreed on
a consensus version of the key points. This was used for deciding to what extent the task had been achieved when rating the summaries and also formed the basis for the analysis of the content reproduced (see Section 4.4.2.5 Analysis below).

As discussed in Section 3.5.1.2 above, originally three CEFR scales were selected for rating but this was narrowed down to the ‘Overall Written Production’ (Council of Europe, 2011, p. 23) scale in the end.

Before starting to rate the summaries, the raters underwent training to ensure a common understanding of the criteria and to benchmark levels of performance. The raters first carried out a familiarisation task using the CEFR descriptors. The descriptors were cut up and the raters had to reconstruct the CEFR scale from A1-C2. Next, the raters used the descriptors to rate the sample writing tasks on the Training Area of the CEFTrain website (University of Helsinki, n.d.). The third step of the rater training involved the raters marking three sample papers together. Each summary was divided into two parts, relating to the content of each of the two parts of the input (the vodcast and the podcast) and each of these two parts was rated by the three expert raters. Raters thus awarded two grades for each test taker - one for each part of the summary - and then discussed their ratings, providing a justification for the rating they gave. After that, they rated a further 7 papers individually then came back together again to discuss the ratings. Where there were differences of opinion, the raters justified their ratings with reference to the CEFR descriptors to develop a shared understanding of the criteria in accordance with Trace, Meier and Janssen (2016, p. 41). As mentioned in Section 3.6.1.2 above, the original intention had been to use all three scales. However, the time and costs involved in this were found to be prohibitive so the rater team decided to use just the ‘Overall Written Production’ scale (Council of Europe, 2011, p. 23). This scale focuses on the development of the text from ‘simple isolated phrases’ at A1 to ‘clear, detailed texts ...’ at B1 and ‘clear, well-structured texts ...’ at C1 level (Council of Europe, 2011, p. 23). This was considered an important aspect as it would differentiate between those test takers who had
not understood the input in detail and were thus only able to reproduce isolated words or phrases to those who were able to produce a more flowing text. The disadvantage of this scale was the mention of ‘synthesising and evaluating information and arguments from a number of sources’ (p. 61) at B2, which was not strictly speaking the case in the EAP test as there was only one lecture; however, the fact that there were two input media was considered an interpretation of the need to synthesise across sources while evaluating was interpreted as deciding what information was important (and therefore should be included) and what was not (so could be omitted) from the written text.

Once a consensus had been reached on all ten benchmark papers, the raters then rated the papers individually using the one scale. Rater 1 rated 96 papers; Rater 2 rated 74 papers and Rater 3 rated 96 papers. As there was sufficient overlap between the ratings, a Rasch analysis was carried out to assess the harshness / leniency of the three raters (see Section 4.4.2.5 Analysis below).

The notes were not assessed as part of this research since this was beyond the scope of the thesis. However, the notes would serve as a back-up for the researcher to provide qualitative insights into how the test takers assimilated the information from the lecture, for example, in the case that they did not produce (much of) a written response.

Furthermore, it should be noted that, although some research has focused on the impact of source text use on scores (Gebril and Plakans, 2016), with regard to RQ1, the focus was on the impact of the input medium on the texts they produced as an integrated listening-into-writing task rather than on the use of the language in the source text. As a result, raters were specifically told to rate the texts as a piece of writing irrespective of how much or how little they had reproduced as this would have impacted on the effect of the input medium. What learners reproduced from the source text and how they integrated this into their texts was the focus of RQ2 (see Section 4.3.1 above).
4.4.2.5 Integrated EAP listening-into-writing test: Analysis

4.4.2.5.1 Integrated EAP listening-into-writing test: Quantitative analysis

Quantitative analysis is an experimental form of research with an initial hypothesis which is investigated by quantifying the data and carrying out numerical analyses (Mackey and Gass, 2005, p. 2). These data can then be analysed using statistical methods (Dörnyei, 2007, p. 24). Mackey and Gass (2005, p. 137) distinguish between two types of quantitative research: associational and experimental. Both of these attempt to establish 'a relationship between or within (independent and dependent) variables’. However, the former looks at the (extent of the) relationship between variables while the latter attempts to find a causal relationship by manipulating one or more variables (Mackey and Gass, 2005, p. 137). This thesis is concerned with both of the above. RQ1 and RQ2 investigate the extent of the relationship between the two forms of lecture input as demonstrated by the summaries produced in the integrated listening-into-writing task, while the use of the feedback questionnaire tries to uncover some of the causes for any (perceived) differences in performance between the two forms of input (RQ3).

Phakiti (2015, p. 27-28) outlines two types of statistics that can be used in quantitative analysis. These are: descriptive statistics, which provide information, for example, regarding the average score or the distribution of grades, and inferential statistics, which allow the researcher to investigate the existence of a causal or linear relationship between variables. In the current study, both of these types of statistics will be used.

In order to get reliable data which are generalisable, it is necessary to ensure that the sample is large enough. This depends on the type of study to be carried out. A qualitative research project may be carried out with very small numbers while for quantitative studies, the numbers are usually higher. Fraenkel and Wallen (2003, cited in Mackey and Gass, 2005, p. 124) suggest that 100 participants is acceptable for a descriptive study while correlational studies
can be carried out with around 50 participants. As explained above, therefore, this study has 116 participants.

4.4.2.5.1.1 RQ1 Quantitative analysis - FACETS
To answer RQ1: ‘How does performance on an integrated listening-into-writing EAP task differ when test takers are presented with lecture input in the form of a podcast or a vodcast?’, before starting to analyse the results of the study, it was necessary to investigate rater reliability since ‘a study cannot be valid if its instruments are not valid’ (Phakiti, 2015, p. 34). Therefore, before analysing the data to answer the research questions, a multi-faceted Rasch analysis was carried out to assess the variations in harshness and leniency between the raters and to determine the Fair Average scores for each summary and input type, i.e. the audio input and the audiovisual input, so that these adjusted scores could be used in the further analysis.

4.4.2.5.1.2 RQ1 Quantitative analysis - SPSS
The Fair Average scores were then used to run Hotelling’s $T^2$ using SPSS to investigate whether there was any significant difference in performance between the two groups with regard to input medium (audio only / audiovisual) and order.

In contrast to a $t$-test, which has only one dependent variable and tests for differences between groups on the independent variable, Hotelling’s $T^2$ is used with multiple dependent variables. However, it is a slight variation on the one-way MANOVA (Multivariate Analysis of Variance), which usually has three or more groups in that it can be used when there are only two groups in the independent variable (Laerd, 2015). In the present study there are two groups, the audio first group and the video first group, and there are two dependent variables, namely, the score for the half of the summary relating to the audio input (podcast) format and the score relating to the video (vodcast) input. It was not wise to run independent $t$-tests due to the risk of a Type 1 error (Field, 2009, p. 348).
4.4.2.5.1.3 RQ1 Quantitative analysis – Text Inspector

The summaries were also analysed quantitatively using Text Inspector (Text Inspector, 2019). This program was used to analyse the summaries using the same analyses as for the lecture transcript above. The analyses used are shown in Table 4.8 below.

While several of these analyses are the same as were used for the Coh-Metrix analysis in the exploratory study, the advantage of this software over Coh-Metrix (Graesser et al., 2004) is that it produces CEFR level analyses based on English Vocabulary Profile (Cambridge University Press, 2015) and AWL (Coxhead, 2000) analyses.

Before running the texts through Text Inspector, I cleaned the data by removing spelling mistakes (e.g. ‘lonelyness’ was corrected to ‘loneliness’ (SA23) or ‘dipressed’ was changed to ‘depressed’ (KA5). Misplaced punctuation such as random or incorrectly placed full stops were also corrected. Also, where participants had abbreviated culture shock to ‘C.S.’, this was replaced with the full form. These changes were made so that metrics such as type count and type-token ratio were not artificially inflated and to reduce the number of words which were classified as ‘Unknown words’ in the CEFR analysis.

The analyses were then carried out on each of the two halves of the summaries (first part and second part) for every test taker. For each group (input mode), the mean, standard deviation, median, mode, minimum and maximum were calculated for each of the analyses in Table 4.8 below. However, to allow a more in-depth analysis of performance by group and input format, the two groups were further sub-divided into:

- Group 1: ‘Audio first group - part one – audio’ texts
- Group 2: ‘Audio first group - part two – video’ texts
- Group 3: ‘Video first group - part one - video’ texts and
- Group 4: ‘Video first group - part two – audio’ texts
It was easier to break the groups down like this rather than to leave them as two groups with two batches of data as this facilitated the overview of the performance of the four sub-groups to compare to the Hotelling’s $T^2$ results from the previous analysis.

Table 4.8: Text Inspector analyses

<table>
<thead>
<tr>
<th>Operation</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics</td>
<td>Sentence count</td>
</tr>
<tr>
<td>Statistics</td>
<td>Token count</td>
</tr>
<tr>
<td>Statistics</td>
<td>Type count</td>
</tr>
<tr>
<td>Statistics</td>
<td>Type/token ratio</td>
</tr>
<tr>
<td>Statistics</td>
<td>Flesch Reading Ease</td>
</tr>
<tr>
<td>Statistics</td>
<td>Average Sentence Length</td>
</tr>
<tr>
<td>Lexical Diversity</td>
<td>Lexical diversity (MTLD)</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>A1 type %</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>A2 type %</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>B1 type %</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>B2 type %</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>C1 type %</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>C2 type %</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>Known Words type %</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>Unlisted type %</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>A1 token %</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>A2 token %</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>B1 token %</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>B2 token %</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>C1 token %</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>C2 token %</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>Known Words token %</td>
</tr>
<tr>
<td>Lexis: AWL</td>
<td>AWL All Types %</td>
</tr>
<tr>
<td>Lexis: AWL</td>
<td>AWL All Tokens %</td>
</tr>
</tbody>
</table>

4.4.2.5.2 Integrated EAP listening-into-writing test: Qualitative analysis

In contrast to quantitative analysis, which is concerned with quantifying results, qualitative analysis is primarily concerned with an in-depth analysis of ‘words’ (Phakiti, 2015, p. 27). Qualitative analysis ‘involves data collection procedures which result primarily in open-ended, non-numerical data which is then analysed primarily by non-statistical methods’ (Dörnyei,
2007, p. 24). It is an approach which is ‘characterized by diversity’ (Dörnyei, 2007, p. 242) because it is ‘associated with a range of different methods, perspectives and approaches’ (Mackey and Gass, 2005, p. 162). Two key characteristics of qualitative research are ‘rich description’ and an iterative, cyclical approach which leads to the researcher reading and re-reading the data to discover the categories that emerge (Mackey and Gass, 2005, p. 162-163). Rather than delving into the data with specific hypotheses that need to be tested, as is the case with quantitative research, the qualitative researcher observes the data and then draws out themes based on what is evident in the data.

Common approaches to qualitative data analysis involve ethnographies, interviews, diaries or journals, case studies and observations (Mackey and Gass, 2005, p. 167). However, qualitative analysis can also be carried out on other types of data, such as essays. For example, as part of a mixed methods approach, Sullivan and Pratt (1996) carried out a qualitative analysis of their students’ essays to investigate the patterns and types of discourse produced under different conditions.

4.4.2.5.2.1 RQ2 Qualitative analysis – content analysis: types of reproductions

To answer RQ2, a qualitative content analysis was carried out. This type of analysis has its origins in quantitative analysis because ‘words, phrases or grammatical structures that fall into specific categories’ are counted (Dörnyei, 2007, p. 245). However, in the case of content analysis, ‘qualitative data may in fact be counted and quantified’ (McDonough and McDonough, 1997, p. 223). As the data usually arise from analysing textual data, content analysis has now become associated with qualitative analysis (Flick, 2014, p. 378) and is used when ‘occurrence is assumed to indicate important trends’ (O’Leary, 2014, p. 315).

To carry out the analysis, the data needed to be coded for occurrences of matches (reproductions) between the spoken text and the PowerPoint slides. By coding the matches, it was possible to investigate how information from the input (lecture and PowerPoint) was reproduced to see whether test takers reproduced verbatim what had been said by the lecturer.
or paraphrased the information. Similarly, verbatim reproductions from the slides could also be analysed.

In accordance with Mackey and Gass (2005, p. 241), I developed a coding system by looking over the data for emergent patterns (such as word matches to the audio, word matches to the PowerPoint slides, paraphrases from the spoken input, etc.) and developing these into a number of headings as follows:

**Table 4.9: Coding system for Qualitative Content Analysis**

<table>
<thead>
<tr>
<th>Code</th>
<th>Type</th>
<th>Level</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWLP</td>
<td>Audio</td>
<td>Word level</td>
<td>Paraphrase</td>
</tr>
<tr>
<td>AWLE</td>
<td>Audio</td>
<td>Word level</td>
<td>Exact</td>
</tr>
<tr>
<td>ACLP</td>
<td>Audio</td>
<td>Clause level</td>
<td>Paraphrase</td>
</tr>
<tr>
<td>ACLE</td>
<td>Audio</td>
<td>Clause level</td>
<td>Exact</td>
</tr>
<tr>
<td>ASLP</td>
<td>Audio</td>
<td>Sentence level</td>
<td>Paraphrase</td>
</tr>
<tr>
<td>ASLE</td>
<td>Audio</td>
<td>Sentence level</td>
<td>Exact</td>
</tr>
<tr>
<td>SWLP</td>
<td>Slide</td>
<td>Word level</td>
<td>Paraphrase</td>
</tr>
<tr>
<td>SWLE</td>
<td>Slide</td>
<td>Word level</td>
<td>Exact</td>
</tr>
<tr>
<td>SCLP</td>
<td>Slide</td>
<td>Clause level</td>
<td>Paraphrase</td>
</tr>
<tr>
<td>SCLE</td>
<td>Slide</td>
<td>Clause level</td>
<td>Exact</td>
</tr>
<tr>
<td>SSLP</td>
<td>Slide</td>
<td>Sentence level</td>
<td>Paraphrase</td>
</tr>
<tr>
<td>SSLE</td>
<td>Slide</td>
<td>Sentence level</td>
<td>Exact</td>
</tr>
</tbody>
</table>

I then used this system to train the coder. After a detailed explanation of the coding system and showing examples of how the codes may look in the data, both the coder and I coded one paper using the coding system and then compared results. The propositions and key words reproduced from the input were coded according to the length of the strings and whether they were an exact match or a paraphrase attempt. A string was based on Weigle and Parker (2012, p. 121-122), whereby a string was considered a 'combination of two or more content words from the same clause' unless a key word such as a sub-heading was two words long, in which case this was not considered a string (clause or sentence level) but was coded at word level.
After a short discussion regarding the areas which we disagreed on and some clarification of the coding system, we coded a small subset of three more papers and compared answers. When we were satisfied that we both understood the coding system, the coder coded all of the other papers individually. Mackey and Gass (2005, p. 245) outline the need to check for inter-coder reliability. To do this, I randomly selected 25 per cent (n=29) of the papers to code from the different levels of performance and the results were compared. In accordance with Mackey and Gass (2005, p. 243), interrater reliability was calculated using the Simple Percentage Agreement method. To calculate this, the number of exact matches is divided by the total number of matches. Where there were discrepancies, the main coder and I then compared results and agreed upon a consensus version. The coder then checked the other papers to ensure they matched the consensus versions.

Further to coding the data, the number and type of reproductions for each participant were counted to investigate the extent to which the participants were influenced by the spoken words of the lecturer and the words on the slide when taking notes and subsequently writing their summary. The number of reproductions from the audio part of the input and the audiovisual part of the input were compared. These data were used to answer Research Question 2.

4.4.2.5.2.2 RQ2 Qualitative analysis – Content Analysis: propositions

As well as investigating the number and types of reproductions from all the test takers, I also qualitatively analysed in depth a small sample of the papers to see which propositions they reproduced from the audio or the video. This was the same kind of analysis as had been done in the exploratory study (see Section 3.4.3 above). To select the samples, a ‘maximum variation sampling’ (Dörnyel, 2007, p. 128) approach was taken as this allows the researcher to explore variations across the different ability levels and input formats and to highlight any commonalities across the different sub-groups. Six samples from each of the five score bands were selected as representative of that ability group, resulting in a qualitative analysis of 30
papers. This analysis aimed to show which words and propositions were reproduced from which types of input. This information was used to assess the extent to which test takers are influenced by the PowerPoint slides when presented with a vodcast compared to the influence of the spoken word in the podcast (RQ2) and to investigate whether there were any differences in the influence of the two types of input across the different levels of proficiency as measured by the EAP summary scores.

The consensus version of the key propositions agreed on by the three native speakers who took the test was used for the coding in this part. These were listed in an Excel spreadsheet as well as the propositions which were on the slides. There were 51 possible propositions for the first part after removing the propositions that formed part of the introduction and 34 in the second part. For the written input from the video, there were 16 propositions in part one (once the introduction had been removed) and 15 in part two.

Due to the numbers of propositions being different across the two parts, a ratio was calculated by dividing the number of propositions reproduced by the total number of propositions in that part. This was done for each person and each part (spoken input part one, spoken input part two, written input part one and written input part two).

The second part of this qualitative analysis involved working out a ratio of the total number of words reproduced divided by the total number of words in the text. For this part, the number of words in each proposition was counted and allocated to the types of propositions (e.g. Audio Word Level Exact or Video Clause Level Paraphrase) discussed in the content analysis in the previous section.

4.4.3 Feedback questionnaire

4.4.3.1 Feedback questionnaire: Rationale

The final task that participants had to complete was a two-page feedback questionnaire. This was developed to serve three purposes:

a) To gather background information on the participants;
b) To gather validation evidence to support the construct validity of the test;
c) To answer RQ3 by investigating participants’ perceptions about their performance according to the different input formats.

Background information about the participants including age, nationality, gender, course of study and English proficiency level. These findings were reported in Section 4.1 above.

To gather validation evidence and information regarding any construct-irrelevant variance (Messick, 1989), the participants responded to questions about how well they thought the tool reflected their ability and whether there were any aspects which may have impacted on their performance (either positively or negatively). A quantitative and qualitative analysis of this data would then provide additional evidence to support (or otherwise) the validity of the tool. For example, if the majority felt that in the listening part of the test task, they did as well as they would in an independent listening test, then this lends support to the validity of the tool. If, on the other hand, they felt that they performed much worse, maybe because of the speed of speech or the quality of the video, then this would highlight areas which would need to be addressed before considering implementing a similar test task.

The third purpose was to answer RQ3. To answer RQ3, participants answered four questions about which of the formats they thought they did better on and why. In addition, they were asked for reasons why they thought they performed less well on the other section. Data from these questions will also be investigated quantitively and qualitatively.

Dörnyei and Taguchi, 2010, p. 109) point out that a well-designed questionnaire can be a useful tool to complement other types of data collection and can be used for the triangulation of results. They argue that questionnaires have several advantages over other types of data collection such as interviews because they are efficient, versatile and cost-effective since the data can be collected and processed easily (p. 6).
4.4.3.2 Feedback questionnaire: Design

Dörnyei and Taguchi (2010, p. 6) warn against the dangers of poor questionnaire design, which can lead to ‘unreliable and invalid data’ and states that a poorly designed questionnaire can also produce superficial and simplistic answers’ (p. 7). The questionnaire was piloted as part of the exploratory study and then, due to the addition of one question and slight amendments to two of the other questions, the native speakers involved in the rating and coding also completed the questionnaire. Another danger is that unmotivated respondents may be careless about responding, so they may miss out or misread questions, thereby leading to unreliable results. They also make the important point that respondents may struggle to understand the questions if their language level is below that of the questions (p. 7). To avoid this issue, questions were kept as simple as possible but, in addition, the administrators collecting the data were all Russian speakers who could translate the questions and answers if any of the participants needed any help. The risk of fatigue when answering a questionnaire is also an important point that Dörnyei and Taguchi (2010) raises. He recommends that a questionnaire should be no longer than 4-6 pages long and take no more than 30 minutes to complete (Dörnyei and Taguchi, 2010). For these reasons, the questionnaire was very brief – only 2 pages long – and took about 5-10 minutes to complete. It was particularly important to keep the questionnaire short as participants were answering it immediately after the testing session, which was already quite long.

With any questionnaire, there is the risk of ‘social desirability (or prestige) bias’ and ‘self-deception’ (Dörnyei and Taguchi, 2010)). With the former, respondents provide the answers that they think the researcher wants to hear or they give answers which present themselves in a good light. The latter is similar but is a more subconscious deceit of oneself (Dörnyei and Taguchi, 2010). Both cases, however, pose a risk to the validity of the data. While one option to guard against this in longer questionnaires is to ask the same questions in different ways to ensure the reliability of the questionnaire, this would have made the questionnaire longer.
and, given that the questionnaire was being answered at the end of a long testing session, a conscious decision to keep the questionnaire as short as possible was taken. However, there was a note on the questionnaire telling participants that the questionnaire had no impact on their test results so they should feel free to give honest answers. While these risks must nevertheless be borne in mind in the analysis of the results, it is hoped that they would evaluate their performance honestly and reasonably well since the participants did not know me personally and the test was very much a low-stakes test for them.

The questionnaire comprised five background questions and 12 content-related questions. The background questions related to their gender, age group, nationality, a self-assessment of their CEF R level, and their course of study (see Section 4.1). As discussed in the previous section, the content-related questions investigated test takers’ perceptions of their performance and the reasons relating to why they performed better or worse on each of the two parts of the lecture task and their feelings about the test more generally. There were five Likert-scale questions, two multiple choice questions, two questions to which participants could provide multiple responses and three open questions. The Likert-scale questions investigated their opinions about the following:

1. The extent to which the results reflect their listening ability
2. Their perception of the difficulty of the writing test overall
3. Their previous knowledge of the lecture topic
4. Their perception of the impact of the topic on their performance
5. Their perception of the difficulty of the summary writing task.

The two multiple-choice questions (Q5 and Q7) asked on which of the two sections (audio or video) the test takers performed better while two of the three open questions (Q6 and Q8) asked them to suggest reasons for their better or worse performance. The two multiple-response questions (Q9 and Q10) asked participants to select the aspects that they felt had a positive effect on their performance and those which had a negative effect respectively and
the third open question (Q12) asked them to suggest any other factors which they felt had an impact on the amount they could write in the summary task.

The questionnaire can be found in Appendix 22.

**4.4.3.3 Feedback questionnaire: Procedure**

The questionnaire was administered in pen-and-paper mode immediately after the participants had completed the main task.

**4.4.3.4 Feedback questionnaire: Analysis**

**4.4.3.4.1 Feedback questionnaire: Quantitative analysis**

The data from the questionnaire were analysed both quantitatively and qualitatively. For the quantitative analysis, descriptive statistics (mean, standard deviation, median, mode) were calculated for the Likert-scale questions and the multiple-choice questions. The results of the background questions fed into Section 4.1 above while the results of the questions relating to test takers’ perceptions of their performance on the two input modes fed into answering Research Question 3 (see Section 5.5). The results of the other questions were used to evaluate the validity of the testing instrument. These results are presented in Section 5.6.

**4.4.3.4.2 Feedback questionnaire: Qualitative analysis**

The open questions on the questionnaire were analysed qualitatively following the ‘classic method for analysing qualitative data’ (Holliday, 2015), which involves the following steps:

1. Coding
2. Determining current themes
3. Constructing an argument
4. Going back to the data (Holliday, 2015, p. 53-54)

In the case of this project, the method was as follows:

1. Coding – the questionnaires were coded for the different aspects which contributed either positively or negatively to their performances
2. Determining the recurrent themes – I analysed the types of effects and grouped these into themes such as ‘timing’, ‘external factors’ (e.g. tiredness) and ‘speech rate’.

3. Constructing an argument – I used the themes outlined above as headings and wrote down examples from the data for the different themes to be used as evidence.

4. Going back to the data – as outlined by Holliday (2015, p. 54), part of the process of qualitative analysis involves going back to the data to reassess the coding system and the themes selected and amending these as necessary. Therefore, I also did this as part of the qualitative analysis in this case.

This information served to answer RQ3 and as a method for collecting validation evidence to investigate to what extent there were issues with the instruments and what helped or hindered the test takers’ performances so that these issues could be taken into account in the recommendations, conclusions and suggestions for future research.

4.5 Conclusion
To summarise, the current research study involved a mixed methods (QUAN + qual) (Dörnyei, 2007, p. 173) approach. Quantitative analyses were used to answer RQ1 and RQ3 while qualitative analyses were used to answer RQ2 and RQ3. Both quantitative and qualitative analyses were used to gather validation evidence from the post-task questionnaires.

For RQ1, Rasch analysis was used to establish rater reliability and to calculate Fair Average scores for both halves of the summaries, which related to the audio and the video input. Next, Hotelling’s $T^2$ was carried out to investigate differences between groups and input formats. To get a deeper analysis of test takers’ performances at the different proficiency levels, the summaries were also analysed using Text Inspector (Text Inspector, 2019) for descriptive statistics, lexical diversity, the percentage of types and tokens from the different CEFR levels according to the EVP (Cambridge University Press, 2015) and to analyse the percentage of types and tokens from the AWL (Coxhead, 2000).
For RQ3, quantitative analyses were used to investigate the relationship between test takers’ perceptions of their performance and, on the one hand, the scores awarded by trained raters, and the length of their texts, on the other.

Quantitative analyses also provided validation evidence for the instrument and served to highlight any issues which need to be addressed.

On the other hand, qualitative analyses were carried out to answer RQ2. The qualitative analysis approach taken was a qualitative content analysis of the types and numbers of reproductions of language (both verbatim or paraphrased) from the audio and video input produced in the summaries.

Qualitative analyses were also used for further information about RQ3 and to gain better insights into test takers’ perceptions of the test.

To conclude, this chapter has introduced the research approach used in the main study. On the basis of the findings from the exploratory study and the changes implemented for the main study, the methodology for the main study - including the design, administration, scoring and analysis of the findings - has been presented here. The next chapter will present the results of the main study.
5. Results

5.1 Introduction
This chapter will present the results from the data collection, which will be analysed both quantitatively and qualitatively. As described in Section 4.4 above, the participants in this study completed four tasks:

1. 30-minute, 60-item, pen and paper, lexico-grammatical placement test
2. Integrated EAP listening-into-writing task
   a. Listening to and taking notes on a lecture on the topic of ‘Culture Shock’ provided in two formats (audio/audiovisual or audiovisual/audio)
   b. Writing a summary on the content of the lecture
3. Post-task feedback questionnaire

First of all, the results of the Quick Placement Test will be presented. After that, the quantitative and qualitative results as they relate to each of the three RQs will be presented in turn. For ease of reference, the RQ will first be restated before presenting the results of these analyses in detail. In the final section, the results of the questions from the feedback questionnaire will be presented. This data serves as validation evidence and will highlight any potential issues with the test that may need to be addressed in a future study or before considering operationalising such a test.

5.2 Quick Placement Test (OUP, 2001)

5.2.1 Introduction
This test aimed to provide a rough estimate of the test takers’ ability so they could be split into two groups of approximately equal ability ahead of the main task (see Section 4.4.1.1).

For the study, data was collected from 116 test-takers. However, some of the test-takers had not been available in the first session when the QPT was completed so, for that reason, QPT data was only collected from 50 test-takers in the audio first group (group 1) and 42 test-takers in the video first group (group 2).
5.2.2 QPT: Quantitative results

5.2.2.1 Descriptive statistics

As outlined in Section 4.4 above, the data were analysed using SPSS to calculate descriptive statistics (mean, standard deviation, median, mode, minimum, maximum, range, skewness and kurtosis). After that, an independent samples t-test was run to establish if any differences in QPT scores between the two groups were significant.

The results revealed that participants in the audio first group scored slightly higher on the QPT ($M = 36.38$, $SD = 10.349$) than participants in the video first group ($M = 34.40$, $SD = 10.719$). The median for the audio first group was 37 compared to 32.50 for the video first group. There were three modes in both groups. In the audio first group, these were 27, 37 and 38 whereas in the video first group, the modes were 20, 27 and 31. Scores in the audio first group ranged from 18 to 58 out of 60 while scores in the video first group ranged from 16 to 55.

5.2.2.2 Inferential statistics: independent samples t-test

Before running the t-test, tests of normality were carried out and the distributions for both groups were found to be normal, so a standard independent samples t-test was run to determine whether there were any statistically significant differences between the QPT scores in the audio first group and the video first group. The mean QPT score for the audio first group was 1.975, 95 per cent CI [-2.398 to 6.349], higher than the mean QPT score for the video first group. There was no statistically significant difference in mean QPT scores between the audio first group and the video first group, $t(90) = .897$, $p = .372$. Therefore, we can accept the null hypothesis for this analysis.
5.3 RQ1: How does performance on an integrated listening-into-writing EAP task differ when test takers are presented with lecture input in the form of a podcast or a vodcast?

5.3.1 Introduction
As described in Table 4.2 and Section 4.4.2 above, the main task for this study was an integrated listening-into-writing EAP test in which test takers listened to and watched a short lecture presented as two halves: one as a podcast and one as a vodcast (or vice versa) then wrote a summary of the input. The summaries were split into two mini-summaries whereby one mini-summary related to the content of the first half of the lecture and the second mini-summary related to the content of the second half of the lecture. Each mini-summary was then marked by between one and three trained raters (see Section 4.4.2.4).

5.3.2 Quantitative analyses
For RQ1, three types of quantitative analyses were carried out (see Section 4.4.2.5.1 above). The data were analysed using FACETS to calculate the Fair Average score for each mini-summary, the Hotelling’s $T^2$ was calculated to establish how the two groups of test takers performed according to input type and order and whether there were any significant differences between the two groups. Finally, Text Inspector (Text Inspector, 2019) was used to analyse the text according to several metrics. These included descriptive measures such as sentence count and token count, lexical diversity measures, percentages of types and tokens at different CEFR levels and the percentages of AWL types and tokens. These will be presented in more detail below.

5.3.2.1 FACETS analysis
The FACETS analysis revealed that the three raters were behaving as independent raters but were in very close agreement: there were 147 exact agreements (50 per cent) compared to expected agreements of 151.2 (51.4 per cent). The Infit MnSq for all three raters was within the 0.5 to 1.5 range that would be considered ‘productive for measurement’ (Linacre, 2012, p. 11). Rater 1 had an Infit MnSq of .85, rater 2 had an Infit MnSq of 1.08 and rater 3’s Infit
MnSq was 1.05. Similarly, the Outfit MnSq values were also within the good range: rater 1 had an Outfit MnSq of .86, rater 2’s Outfit MnSq was .96 and rater 3 had an Outfit MnSq value of .78. This indicates that inter-rater reliability between the three raters was good.

5.3.2.2 Hotelling’s $T^2$

Hotelling’s $T^2$ was run to determine the effect of input format on output performance on the integrated listening-into-writing EAP test. Test takers were assessed on their performance in an essay on the topic of culture shock, which was divided into two halves based on the input from the EAP lecture. Each half of the essay was allocated a band score according to the Overall Written Production scales of the CEFR. Preliminary assumption checking revealed that data were not normally distributed for each group, as assessed by the Shapiro-Wilk’s test ($p<0.05$). However, with MANOVA, ‘violating the assumptions does not necessarily invalidate the results’ since this is ‘relatively robust to violations of the assumptions in many circumstances’ (Bray and Maxwell, 1985, p. 33); there were four univariate outliers in the audio scores for the audio first group but there were no univariate outliers in the data for the video first group, as assessed by inspection of a boxplot; there were no multivariate outliers in the data, as assessed by Mahalanobis distance ($p>.001$); there were linear relationships, as assessed by scatterplot; no multicollinearity ($|r|<.9$); and there was homogeneity of variance-covariance matrices, as assessed by Box’s test of equality of covariance matrices ($p=0.055$); For Hotelling’s $T^2$, the results show that the audio first group scored higher for both the audio input format ($M = 4.006$, $SD = 1.246$) and the video input format ($M = 3.571$, $SD = 1.340$) than the video first group who, in turn, performed better in the video input format ($M = 3.452$, $SD = 1.212$) than in the audio input format ($M = 2.886$, $SD = 1.444$). Figure 5.1 shows these results.
There was a statistically significant difference between the groups on the combined dependent variables, $F(2, 113) = 17.832, p<.0005$; Wilks' $\Lambda = .760$; partial $\eta^2 = .240$. A Bonferroni adjusted $\alpha$ level of .025 with a simultaneous 95 per cent confidence level was used. Since there was a statistically significant result, post-hoc independent samples $t$-tests were run. These results revealed that mean audio scores for the audio first group were 1.12 bands (95 per cent CI, 0.552 to 1.689) higher than mean audio scores for the video first group. There was also a statistically significant difference in audio scores between test takers in the audio first group and test takers in the video first group, $p<.0005$, but this was not the case for the video scores, $p=.618$, ns. The combined group means were statistically significantly different ($p<.05$). Therefore, the null hypothesis can be rejected and we can accept the alternative hypothesis.

Since the independent samples $t$-test assumes a normal distribution and the scores were not normally distributed, a Mann-Whitney U test (CI 97.5 per cent) was also run to check whether the findings stated above would be confirmed by a non-parametric test.

**Figure 5.1: Performance by group**

There was a statistically significant difference between the groups on the combined dependent variables, $F(2, 113) = 17.832, p<.0005$; Wilks' $\Lambda = .760$; partial $\eta^2 = .240$. A Bonferroni adjusted $\alpha$ level of .025 with a simultaneous 95 per cent confidence level was used. Since there was a statistically significant result, post-hoc independent samples $t$-tests were run. These results revealed that mean audio scores for the audio first group were 1.12 bands (95 per cent CI, 0.552 to 1.689) higher than mean audio scores for the video first group. There was also a statistically significant difference in audio scores between test takers in the audio first group and test takers in the video first group, $p<.0005$, but this was not the case for the video scores, $p=.618$, ns. The combined group means were statistically significantly different ($p<.05$). Therefore, the null hypothesis can be rejected and we can accept the alternative hypothesis.

Since the independent samples $t$-test assumes a normal distribution and the scores were not normally distributed, a Mann-Whitney U test (CI 97.5 per cent) was also run to check whether the findings stated above would be confirmed by a non-parametric test.
For the audio scores, the distributions were not similar, as assessed by visual inspection. The audio scores for the audio first group (mean rank = 72.17) were statistically significantly higher than for the video first group (mean rank = 44.83), $U = 889$, $z = -4.389$, $p < .0005$.

For the video scores, the distributions were also not similar, as assessed by visual inspection. The video scores for the audio first group (mean rank = 60.77) was not statistically significantly higher than for the video first group (mean rank = 56.23), $U = 1,550.500$, $z = - .727$, $p = .467$.

Both parametric and non-parametric independent samples $t$-tests showed that the audio scores for the audio first group were statistically significantly higher than for the video first group, whereas there was no statistically significant difference between video scores for either group using both parametric and non-parametric tests. Figure 5.2 shows these results.

![Performance by input mode](image-url)

**Figure 5.2: Performance by input mode**
Looking at the performances across the two input orders (part one and part two), a large proportion of test takers performed worse on the second part (n=57), while slightly more test takers in total (n=59) across the two groups performed equally well on both parts (n=46) or scored higher on the second part (n=13) irrespective of input medium. Breaking this down further into the two groups, there were 25 test takers in the audio first group who scored lower on the second part (video input) than on the first part (audio input) while 25 scored the same. Eight test takers received a higher score on the second part (video input) than on the first part (audio input). For the video group, more test takers (n=32) scored lower on the second part (audio input) than on the first part (video input). 21 test takers were awarded the same score for both halves but only five test takers performed better on the second part (audio input) than on the first part (video input).

5.3.2.3 Text Inspector

- As discussed in Section 4.4.2.5.1.3 above, the two groups were sub-divided into four groups in total to allow a more in-depth insight into performance by group and input format:
  - ‘Audio first group - part one – audio’ texts
  - ‘Audio first group - part two – video’
  - ‘Video first group - part one - video’ and
  - ‘Video first group - part two – audio’ texts

For ease of reference, the measures selected (see Table 4.8 in Section 4.4.2.5.1.3 above) were:

- sentence count
- sentence length
- token count
- type count
Descriptive statistics (mean, standard deviation, median, minimum and maximum) were calculated for the analyses for each of the four sub-groups using SPSS.

Table 5.1 below shows the descriptive statistics and lexical diversity measures for the four groups.

As might be expected from the results for Hotelling’s $T^2$ in Section 5.3.2.2 the sentence count for the audio input in the audio first group (Group 1) ranged from 1 to 29 ($M = 12.737$, $SD = 7.247$, $median = 13$) and was the highest of the four groups. This range was very similar to the range in the audio input texts for the video group (group 4), which ranged from 1 to 28 ($M = 5.966$, $SD = 5.718$, $median = 4$); however, the mean and median for group 4 are much lower. For group 3, the mean sentence count for the video input summaries ranged from 1 to 24 ($M = 8.586$, $SD = 6.297$, $median = 6$), which was higher than in the video input texts for the audio first group (group 2), which ranged from 1 to 23 ($M = 7.386$, $SD = 5.525$, $median = 6$). However, although the median was the same across both groups, the spread of scores was wider for group 3 than for group 2. For groups 2 and 3 then, in contrast to the results of Hotelling’s $T^2$, for the video input, the video first group (group 3) had a higher mean number of sentences according to the Text Inspector analysis than the audio first group (group 2) for the video input, despite the latter having a higher overall mean on the scores awarded for the mini-summaries.
The audio first group also produced longer sentences on average than the video first group.

For the video input, the audio first group (group 2) produced the longest sentences on average \((M = 15.071, SD = 8.770, median = 13.740)\) while for the audio input (group 1), the average length of their sentences was just slightly shorter \((M = 14.692, SD = 6.022, median = 14.530)\). The video first group (group 3) generally produced slightly shorter sentences \((M = 13.537,\)
\(SD = 5.427, \text{median} = 12.400\) for the video input, dropping to \(M = 12.606, SD = 11.383, \text{median} = 11.380\) for the audio input part (group 4).

The results for the token count followed a similar pattern, with the group 1 results being the highest \((M = 195.175, SD = 110.198, \text{median} = 226)\) and group 4 being the lowest \((M = 82.569, SD = 91.473, \text{median} = 49.000)\) despite the fact that group 4 had the widest range on the token count (1 to 353) after group 1 (range: 1 to 414). Groups 2 and 3 were again very similar with the number of tokens in group 3 ranging from 1 to 272, which was slightly higher \((M = 117.203, SD = 86.858, \text{median} = 89.000)\) on average than for group 2, which ranged from 1 to 272 \((M = 110.351, SD = 76.408, \text{median} = 106)\). In this case, group 3 again had a wider spread of scores than group 2 but the median in group 2 was higher.

With regard to the type count, there was once again a big difference between groups 1 and 4. While the range was very similar across the two groups with group 1 ranging from 1 to 180 \((M = 102.754, SD = 49.950, \text{median} = 117.000)\) and group 4 ranging from 1 to 170, for group 4, the mean was considerably lower \((M = 49.966, SD = 45.696, \text{median} = 37.500)\). Again, groups 2 and 3 were similar to each other but group 3 had a marginally wider range of tokens (ranging from 5 to 145) \((M = 71.576, SD = 43.420, \text{median} = 63.000)\) compared to a range of 1 to 136 \((M = 66.561, SD = 37.791, \text{median} = 69)\) in group 2. As with the other results, the spread was wider in group 3 while the median was higher in group 2.

The combination of the type and token counts resulted in differences between the type-token ratios across the four groups ranging from the lowest mean in group 1 \((M = 0.605, SD = 0.162, \text{median} = 0.540)\) and the highest mean in group 4 \((M = 0.780, SD = 0.179, \text{median} = 0.765)\). The mean for group 2 \((M = 0.702, SD = 0.159, \text{median} = 0.670)\) was slightly higher than for group 3 \((M = 0.699, SD = 0.136, \text{median} = 0.680)\). Given that the mean type-token ratio and the median are both higher for group 4 than for group 1, this looks like it contradicts the other data; however, it should be borne in mind that a type-token ratio of 1.00 is very easy to achieve on short texts which have no repeated words.
There was also quite a wide variation in the Flesch Reading Ease in the four groups. Group 3 produced the most difficult texts (i.e. those with the lowest Flesch Reading Ease) on average ($M = 54.467$, $SD = 16.740$, $median = 57.040$) while group 4 produced the easiest ones ($M = 74.869$, $SD = 25.064$, $median = 70.225$). The mean difficulty of the texts as measured by Flesch Reading Ease for groups 1 and 2 were very similar ($M = 65.352$, $SD = 18.349$, $median = 65.230$ and $M = 62.992$, $SD = 24.582$, $median = 61.670$) although the second part appears to be slightly more complex than the first part.

Finally, looking at the lexical diversity measure, MTLD, we can see that the audio first group produced the most complex language in the second part of their summaries (group 2) ($M = 69.685$, $SD = 18.404$, $median = 67.620$) but the texts are less complex in the first part (group 1) ($M = 67.692$, $SD = 17.243$, $median = 66.385$). Conversely, the video first group produced more complex summaries in the first part (group 3) ($M = 68.342$, $SD = 20.393$, $median = 68.690$) while the complexity drops off in the second part (group 4) ($M = 67.420$, $SD = 21.284$, $median = 64.040$).

The next table (Table 5.2) shows the results of the Text Inspector analyses for the percentage of types that test takers produced in their texts at different CEFR levels.

In all four groups, as with the original lecture (see Section 4.4.2.2 above), the majority of tokens were at A1 level, with the audio first group producing the highest percentage of A1 level types in the summaries pertaining to the video input (group two) ($M = 49.455$, $SD = 18.405$, $median = 54.840$) and then the audio input (group 1) ($M = 48.581$, $SD = 13.937$, with a $median$ of 52.030. These were slightly higher than the video first group with group 3 (video first group – part one: video input) producing slightly fewer types at A1 level ($M = 46.629$, $SD = 12.522$, $median = 49.020$) while in the second part (group 4) produced slightly fewer again ($M = 47.211$, $SD = 24.337$, $median = 56.730$). However, for group 4, the median is much higher than the mean compared to the other groups. Similarly, the maximum percentage of A1 types produced by group 4 is much higher than the other groups, at 87.5
per cent, compared with 70.59 per cent, 70.59 per cent and 68.29 per cent for groups 1, 2 and 3 respectively. It is also evident from the standard deviation for group 4 that there was a much wider spread in the percentage of A1 types produced by this group than the other groups.

Table 5.2: Percentage of CEFR types reproduced

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th>Group</th>
<th>Percentage CEFR types</th>
<th>N</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>Audio first group - part one: audio</td>
<td>A1 type %</td>
<td>58</td>
<td>0.000</td>
<td>70.590</td>
<td>48.581</td>
<td>13.937</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A2 type %</td>
<td>58</td>
<td>0.000</td>
<td>22.220</td>
<td>14.202</td>
<td>5.166</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B1 type %</td>
<td>58</td>
<td>0.000</td>
<td>38.460</td>
<td>16.238</td>
<td>5.976</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B2 type %</td>
<td>58</td>
<td>0.000</td>
<td>22.390</td>
<td>6.538</td>
<td>4.171</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C1 type %</td>
<td>58</td>
<td>0.000</td>
<td>6.250</td>
<td>2.067</td>
<td>1.613</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C2 type %</td>
<td>58</td>
<td>0.000</td>
<td>6.000</td>
<td>0.622</td>
<td>1.018</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Known words type %</td>
<td>58</td>
<td>0.000</td>
<td>1.720</td>
<td>0.415</td>
<td>0.497</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unlisted words type %</td>
<td>58</td>
<td>2.440</td>
<td>100.000</td>
<td>11.336</td>
<td>21.235</td>
<td>6.030</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Audio first group - part two: video</td>
<td>A1 type %</td>
<td>58</td>
<td>0.000</td>
<td>70.590</td>
<td>49.455</td>
<td>18.405</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A2 type %</td>
<td>58</td>
<td>0.000</td>
<td>22.890</td>
<td>10.415</td>
<td>5.645</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B1 type %</td>
<td>58</td>
<td>0.000</td>
<td>33.330</td>
<td>16.200</td>
<td>8.415</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B2 type %</td>
<td>58</td>
<td>0.000</td>
<td>12.730</td>
<td>4.577</td>
<td>3.155</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C1 type %</td>
<td>58</td>
<td>0.000</td>
<td>12.120</td>
<td>4.071</td>
<td>3.320</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C2 type %</td>
<td>58</td>
<td>0.000</td>
<td>4.000</td>
<td>0.192</td>
<td>0.720</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Known words type %</td>
<td>58</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unlisted words type %</td>
<td>58</td>
<td>0.830</td>
<td>100.000</td>
<td>15.090</td>
<td>29.532</td>
<td>4.420</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Video first group - part one: video</td>
<td>A1 type %</td>
<td>58</td>
<td>0.000</td>
<td>68.290</td>
<td>46.629</td>
<td>12.522</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A2 type %</td>
<td>58</td>
<td>0.000</td>
<td>21.310</td>
<td>11.949</td>
<td>4.942</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B1 type %</td>
<td>58</td>
<td>7.140</td>
<td>40.000</td>
<td>18.038</td>
<td>6.205</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B2 type %</td>
<td>58</td>
<td>2.440</td>
<td>40.000</td>
<td>10.628</td>
<td>5.906</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C1 type %</td>
<td>58</td>
<td>0.000</td>
<td>12.500</td>
<td>3.277</td>
<td>2.865</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C2 type %</td>
<td>58</td>
<td>0.000</td>
<td>5.000</td>
<td>1.260</td>
<td>1.318</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Known words type %</td>
<td>58</td>
<td>0.000</td>
<td>3.120</td>
<td>0.331</td>
<td>0.676</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unlisted words type %</td>
<td>58</td>
<td>1.180</td>
<td>40.000</td>
<td>7.888</td>
<td>5.634</td>
<td>6.980</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Video first group - part two: audio</td>
<td>A1 type %</td>
<td>58</td>
<td>0.000</td>
<td>87.500</td>
<td>47.211</td>
<td>24.337</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A2 type %</td>
<td>58</td>
<td>0.000</td>
<td>25.000</td>
<td>10.027</td>
<td>7.079</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B1 type %</td>
<td>58</td>
<td>7.140</td>
<td>40.000</td>
<td>18.038</td>
<td>6.205</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B2 type %</td>
<td>58</td>
<td>2.440</td>
<td>40.000</td>
<td>10.628</td>
<td>5.906</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C1 type %</td>
<td>58</td>
<td>0.000</td>
<td>13.040</td>
<td>2.076</td>
<td>2.650</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C2 type %</td>
<td>58</td>
<td>0.000</td>
<td>2.880</td>
<td>0.163</td>
<td>0.520</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Known words type %</td>
<td>58</td>
<td>0.000</td>
<td>2.820</td>
<td>0.494</td>
<td>0.370</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unlisted words type %</td>
<td>58</td>
<td>1.390</td>
<td>100.000</td>
<td>22.367</td>
<td>36.098</td>
<td>5.260</td>
</tr>
</tbody>
</table>
For A2 level types, the audio first group produced slightly more in both parts one and two (see groups 1 and 2) ($M=14.202$, $SD=5.166$, $median=15.130$ and $M=10.415$, $SD=5.645$, $median=11.110$ respectively) than the video first group did in both parts (see groups 3 and 4) ($M=11.949$, $SD=4.942$, $median=12.200$ and $M=10.027$, $SD=7.079$, $median=10.530$ respectively). However, the results for both groups were very similar for part two (group 2: $M=10.415$; group 4: $M=10.027$) irrespective of input format.

For the B1 types, there seems to be a change because the video first group starts to produce higher level language in the first part. This group produced somewhat more B1 level types on average in the first of their two mini-summaries (group 3 - video input) ($M=18.038$, $SD=6.205$, $median=17.190$) than they did in the second mini-summary (group 4 – audio input) ($M=14.070$, $SD=11.705$, $median=14.040$). On the other hand, the audio first group produced a very similar percentage of types across the two mini-summaries, which was almost exactly in the middle between the video group’s two results. The results for the audio first group’s first part (see group 1) are: $M=16.238$, $SD=5.976$, $median=16.150$ whereas for the second part (group 2), the results are: $M=16.200$, $SD=8.415$, $median=16.440$.

This trend continues for the mini-summaries related to the first part (see group 3) in that the video group consistently produced a higher percentage of types at levels B2 ($M=10.628$, $SD=5.906$, $median=9.090$), C1 ($M=3.277$, $SD=2.865$, $median=2.440$) and C2 ($M=1.260$, $SD=1.318$, $median=0.930$) compared to the audio first group (group 1) B2: $M=6.538$, $SD=4.171$, $median=6.210$; C1: $M=2.067$, $SD=1.613$, $median=1.800$; C2: $0.622$, $SD=1.018$, $median=0.000$. In contrast, the audio first group (see group 2) consistently produced a higher percentage of types on average at the higher CEFR levels for the second part of the mini summaries than the video first group (see group 4). These were: B2: $M=4.577$, $SD=3.155$, $median=4.550$; C1: $M=4.071$, $SD=3.320$, $median=3.880$ and C2: $0.192$, $SD=0.720$, $median=0.000$ for the audio first group (group 3) compared to B2: $M=4.038$, $SD
\[ M = 3.703, \text{ median} = 3.840; \text{ C1: } M = 2.076, SD = 2.650, \text{ median} = 1.550 \text{ and } \text{ C2: } M = 0.163, SD = 0.520, \text{ median} = 0.000 \] respectively for the video first group (group 4).

There is a much higher percentage of unlisted words on average in both part two groups (see groups 2 and 4): \( M = 15.090, SD = 29.532, \text{ median} = 4.420 \) compared to \( M = 22.367, SD = 36.098, \text{ median} = 5.260 \) respectively than in the part one groups (see groups 1 and 3): \( M = 11.336, SD = 21.235, \text{ median} = 6.030 \) and \( M = 7.888, SD = 5.634, \text{ median} = 6.980 \).

The next table (Table 5.3) shows the percentage of tokens at the different CEFR levels. The results of this part of the analysis reveal a similar picture to the previous one in as far as group 1 has a higher percentage of lower level tokens whereas group 3 has a higher percentage of higher-level tokens. Overall, both the audio first and the video first groups performed better on the first half than on the second half, which again corresponds to the findings from the Hotelling’s \( T^2 \) test. There was a gradual decline in the average percentage of A1 level tokens across the four groups ranging from \( M = 58.338, SD = 16.107, \text{ median} = 63.890 \) for group 1 (audio first group, part one – audio input) to \( M = 51.427, SD = 26.290, \text{ median} = 62.420 \) for group 4 (video first group, part two audio input). In accordance with the results of the statistical analysis described in Section 5.3.2.2, group 2 also had a slightly higher percentage of A1 level tokens on average than group 3. These figures were in contrast to the analysis on the lecture transcript as this revealed an increase in A1 level words between the transcript for the first part and the second from 64.09 per cent to 67.71 per cent.

For the A2 tokens, both groups again had a higher percentage of A2 tokens in the first half of their text than in the second half but once more, the audio first group was higher for the first part (see group 1) than the video first group was on the corresponding part (see group 3): \( M = 10.774, SD = 3.862, \text{ median} = 11.840 \) compared to \( M = 9.605, SD = 4.384, \text{ median} = 8.240 \). Conversely, however, for the second part, the video first group (group 4) had a higher average percentage than the audio first group (group 2): \( M = 8.019, SD = 5.686, \text{ median} = 8.505 \) and \( M = 7.936, SD = 4.252, \text{ median} = 8.240 \).
Table 5.3: Percentage of CEFR tokens reproduced

<table>
<thead>
<tr>
<th>Group</th>
<th>Percentage CEFR tokens</th>
<th>N</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>St. Dev</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Audio first group - part one: audio</td>
<td>A1 token %</td>
<td>58</td>
<td>0.00</td>
<td>74.07</td>
<td>58.338</td>
<td>16.107</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2 token %</td>
<td>58</td>
<td>0.00</td>
<td>16.00</td>
<td>10.774</td>
<td>3.862</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B1 token %</td>
<td>58</td>
<td>0.00</td>
<td>40.00</td>
<td>15.319</td>
<td>6.435</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B2 token %</td>
<td>58</td>
<td>0.00</td>
<td>17.09</td>
<td>4.298</td>
<td>3.118</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C1 token %</td>
<td>58</td>
<td>0.00</td>
<td>4.670</td>
<td>1.291</td>
<td>1.071</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C2 token %</td>
<td>58</td>
<td>0.00</td>
<td>4.44</td>
<td>0.369</td>
<td>0.694</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Known words token %</td>
<td>58</td>
<td>0.00</td>
<td>0.97</td>
<td>0.229</td>
<td>0.287</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unlisted words token %</td>
<td>58</td>
<td>1.33</td>
<td>100.00</td>
<td>9.381</td>
<td>3.990</td>
</tr>
<tr>
<td>2</td>
<td>Audio first group - part two: video</td>
<td>A1 token %</td>
<td>58</td>
<td>0.00</td>
<td>74.32</td>
<td>54.896</td>
<td>20.730</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2 token %</td>
<td>58</td>
<td>0.00</td>
<td>17.89</td>
<td>7.936</td>
<td>4.252</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B1 token %</td>
<td>58</td>
<td>0.00</td>
<td>38.46</td>
<td>16.851</td>
<td>9.106</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B2 token %</td>
<td>58</td>
<td>0.00</td>
<td>10.14</td>
<td>4.101</td>
<td>2.442</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C1 token %</td>
<td>58</td>
<td>0.00</td>
<td>10.26</td>
<td>3.022</td>
<td>2.675</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C2 token %</td>
<td>58</td>
<td>0.00</td>
<td>3.70</td>
<td>0.164</td>
<td>0.694</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Known words token %</td>
<td>58</td>
<td>0.00</td>
<td>0.00</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unlisted words token %</td>
<td>58</td>
<td>0.45</td>
<td>100.00</td>
<td>13.921</td>
<td>29.918</td>
</tr>
<tr>
<td>3</td>
<td>Video first group - part one: video</td>
<td>A1 token %</td>
<td>58</td>
<td>0.00</td>
<td>75.88</td>
<td>52.175</td>
<td>15.090</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2 token %</td>
<td>58</td>
<td>0.00</td>
<td>23.14</td>
<td>9.605</td>
<td>4.384</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B1 token %</td>
<td>58</td>
<td>6.90</td>
<td>40.00</td>
<td>19.781</td>
<td>7.585</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B2 token %</td>
<td>58</td>
<td>1.56</td>
<td>40.00</td>
<td>8.452</td>
<td>6.242</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C1 token %</td>
<td>58</td>
<td>0.00</td>
<td>10.71</td>
<td>2.464</td>
<td>2.384</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C2 token %</td>
<td>58</td>
<td>0.00</td>
<td>4.350</td>
<td>1.022</td>
<td>1.167</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Known words token %</td>
<td>58</td>
<td>0.00</td>
<td>8.520</td>
<td>0.342</td>
<td>1.174</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unlisted words token %</td>
<td>58</td>
<td>0.71</td>
<td>40.00</td>
<td>6.160</td>
<td>5.753</td>
</tr>
<tr>
<td>4</td>
<td>Video first group - part two: audio</td>
<td>A1 token %</td>
<td>58</td>
<td>0.00</td>
<td>88.240</td>
<td>51.427</td>
<td>26.290</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2 token %</td>
<td>58</td>
<td>0.00</td>
<td>20.00</td>
<td>8.019</td>
<td>5.686</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B1 token %</td>
<td>58</td>
<td>0.00</td>
<td>75.00</td>
<td>14.283</td>
<td>11.890</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B2 token %</td>
<td>58</td>
<td>56.9</td>
<td>13.33</td>
<td>2.951</td>
<td>2.878</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C1 token %</td>
<td>58</td>
<td>0.00</td>
<td>11.11</td>
<td>1.590</td>
<td>2.220</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C2 token %</td>
<td>58</td>
<td>0.00</td>
<td>2.41</td>
<td>0.131</td>
<td>0.464</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Known words token %</td>
<td>58</td>
<td>0.00</td>
<td>5.21</td>
<td>0.090</td>
<td>0.684</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unlisted words token %</td>
<td>58</td>
<td>0.62</td>
<td>100.00</td>
<td>21.509</td>
<td>36.490</td>
</tr>
</tbody>
</table>

For the B1 tokens and above, the same pattern reveals itself as for the average percentage of types described above: the video first group (group 3) consistently has a higher percentage of the tokens at levels B1-C2 for the first part of the input compared to the audio first group (group 1). However, for the corresponding levels, the audio first group (group 2) consistently
has a higher average percentage of tokens on the second part than the video first group (group 4). The audio first group has a consistently higher average percentage of tokens for the first half than for the second half except for the B1 and C1 levels, where the second part has a higher percentage than the first part: The B1 level percentage for group 2 is \( M = 16.851, SD = 9.106, \text{median} = 15.940 \) compared to \( M = 15.319, SD = 6.435, \text{median} = 14.470 \) for group 1. The C1 level for group 2 is \( M = 3.022, SD = 2.675, \text{median} = 2.430 \) compared to \( M = 1.291, SD = 1.071, \text{median} = 1.090 \) for group 1. For the video first group, part one (group 3) has a higher average percentage of tokens for each of the CEFR bands except the unlisted percentage which is much higher for group 4 \( (M = 21.509, SD = 36.490, \text{median} = 4.170) \) than for any of the other groups.

The final part of the Text Inspector analysis looks at the percentage of AWL types and tokens that test takers produced in their texts. Table 5.4 shows the results of this analysis.

### Table 5.4 Percentage of AWL types and tokens analysis

<table>
<thead>
<tr>
<th>Group</th>
<th>AWL types and tokens (%)</th>
<th>N</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 Audio first group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- part one: audio</td>
<td>AWL all types %</td>
<td>58</td>
<td>0.000</td>
<td>28.570</td>
<td>8.155</td>
<td>4.950</td>
<td>7.270</td>
</tr>
<tr>
<td></td>
<td>AWL all tokens %</td>
<td>58</td>
<td>0.000</td>
<td>28.570</td>
<td>8.584</td>
<td>4.885</td>
<td>7.690</td>
</tr>
<tr>
<td><strong>2 Audio first group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- part two: video</td>
<td>AWL all types %</td>
<td>58</td>
<td>0.000</td>
<td>25.000</td>
<td>11.400</td>
<td>6.026</td>
<td>10.950</td>
</tr>
<tr>
<td></td>
<td>AWL all tokens %</td>
<td>58</td>
<td>0.000</td>
<td>25.000</td>
<td>11.697</td>
<td>6.177</td>
<td>10.810</td>
</tr>
<tr>
<td><strong>3 Video first group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- part one: video</td>
<td>AWL all types %</td>
<td>58</td>
<td>2.500</td>
<td>40.000</td>
<td>11.916</td>
<td>7.780</td>
<td>10.340</td>
</tr>
<tr>
<td></td>
<td>AWL all tokens %</td>
<td>58</td>
<td>3.490</td>
<td>40.000</td>
<td>13.027</td>
<td>7.755</td>
<td>11.360</td>
</tr>
<tr>
<td><strong>4 Video first group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- part two: audio</td>
<td>AWL all types %</td>
<td>58</td>
<td>0.000</td>
<td>20.830</td>
<td>8.578</td>
<td>5.753</td>
<td>9.645</td>
</tr>
<tr>
<td></td>
<td>AWL all tokens %</td>
<td>58</td>
<td>0.000</td>
<td>25.000</td>
<td>8.922</td>
<td>6.042</td>
<td>9.010</td>
</tr>
</tbody>
</table>

This table shows that group 3 produced the highest percentage of AWL types and tokens \( (M = 11.916, SD = 7.780, \text{median} = 10.340 \) and \( M = 13.027, SD = 7.755, \text{median} = 11.360 \) respectively but group 2 had very similar results: \( M = 11.400, SD = 6.026, \text{median} = 10.950 \) and \( M = 11.697, SD = 6.177, \text{median} = 10.810 \). On the other hand, group 1 produced the lowest percentage: \( M = 8.155, SD = 4.950, \text{median} = 7.270 \) and \( M = 8.584, SD = 4.885 \).
median = 7.690 respectively but group 4 had very similar results: \( M = 8.578, SD = 5.753 \), \( \text{median} = 9.645 \) and \( M = 8.922, SD = 6.042, \text{median} = 9.010 \) respectively.

5.3.3 Summary of results for RQ1
To summarise, the QPT was used to divide the participants into two groups of approximately equivalent language proficiency level. The results showed that there was no statistically significant difference between the two groups.

Next, a multifaceted Rasch analysis to establish rater reliability and to calculate Fair Average scores for all test takers for the scores on each of the two input formats. The three raters were found to be behaving like independent raters yet still within a good level of consistency, which fell within the Infit MnSq range of 0.5 to 1.5 that Linacre (2012, p. 11) considers useful for measurement.

The Hotelling’s \( T^2 \) analysis revealed that the audio first group scored better overall while both groups scored more highly in part one than part two, irrespective of input format. There was a statistically significant difference between performance on the audio input between the two groups but there was no significant difference between video input.

With regard to the linguistic analyses carried out using Text Inspector, the results showed the group 1 (audio first - part one: audio input) generally scored highest while group 4 (video first – part two: audio input) generally had the lowest scores. However, the scores for groups 2 (audio first – part two: video input) and 3 (video first – part one: video input) were often similar, although, in contrast to the findings from the statistical analysis above, group 3 performed better than group 2 in some cases. Examples of this include the sentence count, the token count, type count and Flesch Reading Ease in the descriptive statistics analysis (Table 5.1). In the CEFR type analysis (Table 5.2), the results were the same: group 1 had the highest percentages for the lower CEFR level types while group 3 had the higher percentages for the higher CEFR level types. However, this could have been a result of the word counts involved. For the third analysis (CEFR tokens) (Table 5.3), group 3 again had
higher percentages of tokens at all levels from B1 upwards. The last analysis (AWL types and tokens) (Table 5.4) reveals the same pattern: group 3 has consistently higher percentages than groups 1, 2 and 4. Group 2 has the second highest percentages while group 1 and group 4 are very similar. These results would therefore appear to contrast with the results from Hotelling’s $T^2$ but these results could have been affected by the token counts in each case.

5.4 RQ2: To what extent are test takers influenced by the written word in an EAP lecture compared to the spoken word?

5.4.1 Introduction: Qualitative analyses

RQ2 looks at how the test takers use the input in their texts. A qualitative analysis of the number and types of reproductions in all 116 texts was carried out using a Qualitative Content Analysis approach. For ease of reference, the classification system outlined in the Methodology chapter above (see Section 4.4.2.5.2.1) is reproduced as Table 5.5 below.

The second part of the analysis looked at 30 participants’ texts in more detail and analysed what percentage of the propositions they reproduced from the spoken content and which of the propositions from the PowerPoint slides they reproduced.

Table 5.5: Coding system for coding analysis

<table>
<thead>
<tr>
<th></th>
<th>Audio</th>
<th>Word level</th>
<th>Paraphrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWLP</td>
<td>Audio</td>
<td>Word level</td>
<td>Paraphrase</td>
</tr>
<tr>
<td>AWLE</td>
<td>Audio</td>
<td>Word level</td>
<td>Exact</td>
</tr>
<tr>
<td>ACLP</td>
<td>Audio</td>
<td>Clause level</td>
<td>Paraphrase</td>
</tr>
<tr>
<td>ACLE</td>
<td>Audio</td>
<td>Clause level</td>
<td>Exact</td>
</tr>
<tr>
<td>ASLP</td>
<td>Audio</td>
<td>Sentence level</td>
<td>Paraphrase</td>
</tr>
<tr>
<td>ASLE</td>
<td>Audio</td>
<td>Sentence level</td>
<td>Exact</td>
</tr>
<tr>
<td>SWLP</td>
<td>Slide</td>
<td>Word level</td>
<td>Paraphrase</td>
</tr>
<tr>
<td>SWLE</td>
<td>Slide</td>
<td>Word level</td>
<td>Exact</td>
</tr>
<tr>
<td>SCLP</td>
<td>Slide</td>
<td>Clause level</td>
<td>Paraphrase</td>
</tr>
<tr>
<td>SCLE</td>
<td>Slide</td>
<td>Clause level</td>
<td>Exact</td>
</tr>
<tr>
<td>SSLP</td>
<td>Slide</td>
<td>Sentence level</td>
<td>Paraphrase</td>
</tr>
<tr>
<td>SSLE</td>
<td>Slide</td>
<td>Sentence level</td>
<td>Exact</td>
</tr>
</tbody>
</table>
5.4.2 Qualitative Content Analysis: types of reproductions

5.4.2.1 Results: Inter-coder reliability for Content Analysis: types of reproductions

To answer RQ2, the number and types of propositions in each of the mini-summaries were counted along with the total number of matches in each mini-summary for each participant. The Simple Percentage Agreement analysis revealed an inter-coder reliability of 71.43 per cent for all exact matches. This represented 490 exact matches from a possible sample total (n=29) of 686. There were also 88 similar matches. A ‘similar match’ occurred when one coder had identified an exact word match and coded that but the other coder saw the word as part of a paraphrased clause, for example, the phrase ‘culture shock is a natural state of everyone’ (KV10) was coded as an exact word match (for ‘natural’) by one coder while the other coded the whole clause as a paraphrase of the audio. If these close matches were added in, on the basis that a match of some description had been identified by both coders, the inter-coder reliability was 84.26 per cent.

The outstanding matches arose when one coder interpreted something as a match but the other coder did not. These amounted to 74 for Coder A and 34 for Coder B. An example of a coded paper is shown in Appendix 23.

5.4.2.2 Results: Content Analysis: types of reproductions

The results of the coding revealed that the audio first group produced over 40 per cent more matches of all types in part one (n=1081) compared to the video first group (n=765), resulting in a mean and a standard deviation of $M = 18.638$ and $SD = 11.195$ respectively. The mean was almost 50 per cent higher than that of the video first group, which had a mean and standard deviation of $M = 13.190$ and $SD = 9.745$ respectively. The median number of matches in the audio first group was 21 while this was only 10 in the video first group. However, the range of matches for the audio first group was slightly lower than for the video first group, ranging from 0.0 to 36 as compared to 0.0 to 38 respectively. For part two, the picture was similar. However, in this case, the audio first group reproduced more than 57 per
cent more matches (n=572) than the video first group (n=363) resulting in a mean that was also over 50 per cent higher in the audio first group ($M = 9.862, SD = 6.464$) compared to the video first group ($M = 6.259, SD = 5.674$). The median was also more than twice as high for the audio first group than the video first group while the range of matches per person varied from 0 to 24 in the audio first group compared to 21 in the video first group. We can also see that the total number of matches for both groups for part one was almost twice as high as for the second part (1846 and 935 matches respectively). Together, these figures make a total of 2781 matches of different types reproduced by the 116 test takers.

As can be seen from Table 5.6, in line with the findings from RQ1 above, both groups reproduced considerably more matches in part one than in part two. The audio first group produced 88.98 per cent more matches in part one than part two while the video first group produced 107 per cent more matches in part one than part two.

Table 5.6: Summary of coding results – number of matches

<table>
<thead>
<tr>
<th></th>
<th>Part one matches</th>
<th></th>
<th>Part two matches</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Audio first group</td>
<td>Video first group</td>
<td>Audio first group</td>
</tr>
<tr>
<td>Total</td>
<td>1081</td>
<td>765</td>
<td>572</td>
</tr>
<tr>
<td>Mean</td>
<td>18.638</td>
<td>13.190</td>
<td>9.862</td>
</tr>
<tr>
<td>Median</td>
<td>21</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Min</td>
<td>0.0</td>
<td>1</td>
<td>0.0</td>
</tr>
<tr>
<td>Max.</td>
<td>36</td>
<td>38</td>
<td>24</td>
</tr>
<tr>
<td>GRAND TOTAL</td>
<td>1846</td>
<td>935</td>
<td></td>
</tr>
</tbody>
</table>

We will now consider the breakdown of these matches according to type and length of match.

Tables 5.7 and 5.8 in Section 5.4.2.2.1 and 5.4.2.2.2 show the group matches for parts one and two respectively.
5.4.2.2.1 Part one

For part one, what is immediately clear from Table 5.7 (below) is that the audio group reproduced far more matches from the spoken input (codes beginning with A…) than the video first group, with the exception of ASLE (Audio Sentence Level Exact). These differences amounted to around one-third more for AWLP (Audio Word Level Paraphrase) and for ACLP (Audio Clause Level Paraphrase). These were 99 AWLPs ($M = 1.707$, $SD = 1.664$) and 62 AWLPs ($M = 1.069$, $SD = 1.282$) for the two groups respectively. The audio first group produced 378 ($M = 6.517$, $SD = 4.139$) ACLPs in total compared to 204 ($M = 3.517$, $SD = 3.743$) for the video first group. Similarly, the audio first group reproduced around one-quarter more ASLP (Audio Sentence Level Paraphrase) matches than the video first group with 39 ($M = 0.672$, $SD = 0.929$) and 30 ($M = 0.517$, $SD = 0.941$) respectively. For AWLE (Audio Word Level Exact) and ACLE (Audio Clause Level Exact), the audio first group produced around three times as many matches as the video first group. There were 496 AWLEs ($M = 8.552$, $SD = 5.800$) in the audio first group while there were only 151 AWLEs ($M = 2.603$, $SD = 3.549$) in the video first group. The audio first group reproduced 63 ACLEs ($M = 1.086$, $SD = 1.291$) compared to just 24 ($M = 0.414$, $SD = 1.093$) in the video first group. Only the ASLE (Audio Sentence Level Exact) matches were reproduced more frequently by the video first group than the audio first group where the former produced 50 per cent more matches, that is nine ($M = 0.155$, $SD = 0.523$) compared to six ($M = 0.103$, $SD = 0.673$) for the latter. The second half of the first table shows the matches from the written input (codes starting with S…). The audio first group did not produce any matches for these codes as they did not have access to the video in part one.

In part one, when they had access to the video, the video first group produced 175 ($M = 3.034$, $SD = 2.669$) SWLE (Slide Word Level Exact) matches on average from the video input and a mere 9 SWLP (Slide Word Level Paraphrase) ($M=0.155$, $SD = 0.489$) matches from the video. Although the number of ACLPS (204 in total) was high for the audio input, the SCLPs
(Slide Clause Level Paraphrases) matches were considerably lower at 13 ($M = 0.224, SD = 0.531$). On the other hand, the video first group reproduced 69 SCLE (Slide Clause Level Exact) matches ($M = 1.190, SD = 1.162$). The SSLPs (Slide Sentence Level Paraphrases) from the video were much lower than from the audio with a total of only two reproductions across the whole group ($M = 0.034, SD = 0.184$) while the SSLE (Slide Sentence Level Exact) matches were higher than the corresponding matches from the audio with 16 reproductions ($M = 0.276, SD = 0.451$). Given that there was only one sentence that they could reproduce exactly from the slides, this meant that only 16 out of the 58 test takers in the video first group wrote down and reproduced the one full sentence from the slides, which was the definition of culture shock according to Oberg (1954).
Table 5.7: Group results for part one matches

<table>
<thead>
<tr>
<th>Part one</th>
<th>AWLP</th>
<th>ACLP</th>
<th>ASLP</th>
<th>AWLE</th>
<th>ACLE</th>
<th>ASLE</th>
<th>SWLP</th>
<th>SCLP</th>
<th>SSLP</th>
<th>SWLE</th>
<th>SCLE</th>
<th>SSLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio first group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTALS</td>
<td>99</td>
<td>378</td>
<td>39</td>
<td>496</td>
<td>63</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MEAN</td>
<td>1.707</td>
<td>6.517</td>
<td>0.672</td>
<td>8.552</td>
<td>1.086</td>
<td>0.103</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Stdev</td>
<td>1.664</td>
<td>4.139</td>
<td>0.929</td>
<td>5.800</td>
<td>1.291</td>
<td>0.673</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>median</td>
<td>2.0</td>
<td>7.0</td>
<td>0.0</td>
<td>11.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>min</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Max</td>
<td>6.0</td>
<td>15.0</td>
<td>4.0</td>
<td>19.0</td>
<td>5.0</td>
<td>5.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Video first group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>62</td>
<td>204</td>
<td>30</td>
<td>151</td>
<td>24</td>
<td>9</td>
<td>9</td>
<td>13</td>
<td>2</td>
<td>176</td>
<td>69</td>
<td>16</td>
</tr>
<tr>
<td>MEAN</td>
<td>1.069</td>
<td>3.517</td>
<td>0.517</td>
<td>2.603</td>
<td>0.414</td>
<td>0.155</td>
<td>0.155</td>
<td>0.224</td>
<td>0.034</td>
<td>3.034</td>
<td>1.190</td>
<td>0.276</td>
</tr>
<tr>
<td>Stdev</td>
<td>1.282</td>
<td>3.743</td>
<td>0.941</td>
<td>3.549</td>
<td>1.093</td>
<td>0.523</td>
<td>0.489</td>
<td>0.531</td>
<td>0.184</td>
<td>2.669</td>
<td>1.162</td>
<td>0.451</td>
</tr>
<tr>
<td>median</td>
<td>1.0</td>
<td>2.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>min</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Max</td>
<td>6</td>
<td>12</td>
<td>4</td>
<td>17</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>9</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>GRAND TOTAL</td>
<td>161</td>
<td>578</td>
<td>68</td>
<td>640</td>
<td>86</td>
<td>15</td>
<td>9</td>
<td>16</td>
<td>2</td>
<td>265</td>
<td>71</td>
<td>16</td>
</tr>
</tbody>
</table>
Table 5.8: Group results for part two matches

<table>
<thead>
<tr>
<th></th>
<th>Audio first group</th>
<th>Video first group</th>
<th>GRAND TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AWLP</td>
<td>ACLP</td>
<td>ASLP</td>
</tr>
<tr>
<td><strong>TOTALS</strong></td>
<td>31</td>
<td>175</td>
<td>31</td>
</tr>
<tr>
<td><strong>MEAN</strong></td>
<td>0.534</td>
<td>3.017</td>
<td>0.534</td>
</tr>
<tr>
<td><strong>Stdev</strong></td>
<td>0.825</td>
<td>2.859</td>
<td>1.001</td>
</tr>
<tr>
<td><strong>median</strong></td>
<td>0.0</td>
<td>2.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>min</strong></td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Max</strong></td>
<td>3.0</td>
<td>11.0</td>
<td>4.0</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>13</td>
<td>177</td>
<td>28</td>
</tr>
<tr>
<td><strong>MEAN</strong></td>
<td>0.224</td>
<td>3.052</td>
<td>0.483</td>
</tr>
<tr>
<td><strong>Stdev</strong></td>
<td>0.531</td>
<td>2.800</td>
<td>0.843</td>
</tr>
<tr>
<td><strong>median</strong></td>
<td>0.0</td>
<td>3.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>min</strong></td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Max</strong></td>
<td>2.0</td>
<td>11.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>
5.4.2.2.2 Part two

Moving on to the results of the second part of the lecture input (Table 5.8 above), we see immediately that this was the part in which the audio first group had access to the video whereas the video first group had only audio input in this part. With the exception of the AWLP matches, the number of audio-based reproductions is quite similar.

The audio first group still produced more than twice as many AWLP matches with 31 ($M = 0.534$, $SD = 0.825$) compared to 13 ($M = 0.224$, $SD = 0.531$) in the video first group. For the other measures though, the reproductions were much more similar. The audio first group again produced more ASLP and AWLE matches with 31 ($M = 0.534$, $SD = 1.001$) and 70 ($M = 1.207$, $SD = 1.376$) compared to 28 ($M = 0.483$, $SD = 0.843$) and 65 ($M = 1.121$, $SD = 1.609$) for the video first group. On the other hand, the video group produced more ACLPs with 177 ($M = 3.052$, $SD = 2.800$) compared to 175 ($M = 3.017$, $SD = 2.859$) and more ACLEs with 70 ($M = 1.207$, $SD = 1.609$) compared to 59 ($M = 1.017$, $SD = 0.132$). For ASLEs, there were eight reproductions ($M = 0.138$, $SD = 0.576$) in the video first group compared to only one ($M = 0.017$, $SD = 0.132$) from the audio first group.

Regarding the reproductions that the audio first group produced from the video input, the highest frequency of reproductions was in the SWLE and SCLE bands with 107 ($M = 1.845$, $SD = 2.172$) and 71 ($M = 1.224$, $SD = 1.584$) reproductions respectively. While the number of clause level reproductions from the video input was similar for both groups, the audio first group produced much fewer SWLEs than the video first group had done. The audio first group also produced far fewer exact word matches overall in the second part than in the first, even when both the audio and slide exact word matches were added together. As for the sentence level exact matches, there was again only one sentence in the whole of the video input which could be reproduced verbatim from the slides but this time only six students out of the 58 reproduced it. The sentence in this part was a summary style sentence, stating that culture
shock is perfectly natural and that an individual suffering from culture shock is not likely to be the only person suffering from it.

These results show that the participants produced more clause level paraphrases and exact word matches from the audio input than other reproductions. Similarly, the exact word level reproductions from the video were also substantial for both groups when they had access to the video input. However, very few students were able to reproduce sentence level paraphrases or exact matches.

5.4.3 Qualitative Content Analysis: propositions

In order to get a deeper insight into exactly which propositions students were reproducing from the audio and video input, a qualitative analysis of 30 test papers (which comprise two mini-summaries each, one representing the audio input and one representing the video input) was carried out. To ensure a balanced sample, 15 of the papers were from the audio first group while the other 15 were from the video first group. 5 papers were selected from each of the five CEFR bands. The 30 papers were analysed for matches to the spoken and written propositions identified by the three native speakers (see Appendix 24) who took the test (see Section 4.4.2.4).

5.4.3.1 Results: Qualitative Content Analysis: propositions

5.4.3.1.1 Part one

With part one, there was a total of 238 occurrences of propositions from part one in the 15 texts from the audio group while the video first group reproduced 169 propositions. Table 5.9 shows the most frequently reproduced propositions across both groups and as totals while Table 5.10 shows the least frequently reproduced propositions across both groups and as totals. The full table showing all the reproductions across both groups and as totals can be found in Appendix 25.

As can be seen from Table 5.9 below, the most frequently reproduced proposition for both groups was ‘feeling depressed’ with 11 and 14 students respectively out of the total sample
of 30 reproducing this proposition. The first and last stages of culture shock, ‘Honeymoon Stage’ (n=11) and ‘Independence Stage’ (n=10) were the next two most frequently produced propositions in the audio first group but the other stages of culture shock also featured heavily in the audio first group’s texts. On the other hand, in the video first group, ‘lonely’ (n=12) was the second most widely cited followed by ‘feelings of insecurity caused by new culture’ (n=9) and ‘difficulty solving problems’.

Table 5.9: Most frequently reproduced propositions in part one across both groups

<table>
<thead>
<tr>
<th>Culture shock lecture - Expert Panel propositions</th>
<th>AUD FIRST</th>
<th>VID FIRST</th>
<th>AUD + VID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Propositions - part one</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>feelings of anxiety and insecurity caused by new culture</td>
<td>5</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>feeling depressed;</td>
<td>11</td>
<td>14</td>
<td>25</td>
</tr>
<tr>
<td>lonely</td>
<td>6</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>insecurity;</td>
<td>4</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>difficulty solving problems;</td>
<td>7</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>Five stages:</td>
<td>9</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>1. Honeymoon stage:</td>
<td>11</td>
<td>7</td>
<td>18</td>
</tr>
<tr>
<td>excited;</td>
<td>9</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>2. Disintegration (or Distress) stage;</td>
<td>9</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>(you realise) things done differently in new culture / (notice cultural differences);</td>
<td>8</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>3. Re-integration stage:</td>
<td>8</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>4. Autonomy stage:</td>
<td>9</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>get to grips with new culture / coming to terms with culture;</td>
<td>8</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>5. Independence stage:</td>
<td>10</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>TOTAL REPRODUCTIONS PART ONE</td>
<td>238</td>
<td>169</td>
<td>407</td>
</tr>
</tbody>
</table>

At the other end of the scale (see Table 5.10), there is more agreement about what students did not write with seven of the bottom results applying to both groups. Three of the propositions were not reproduced by anyone in either group: ‘unsure’ (n=0), ‘bottom of the U-curve’ (n=0) and ‘halfway up the right hand side’ (n=0). The other four least frequent reproductions had just one reproduction in one group and zero in the other group. Three of these: ‘applies to everyone’, ‘top left of U curve’ and ‘W curve: return to old country’ had only one reproduction by the audio first group while nobody in the video first group reproduced
these at all; conversely, there was one occurrence of ‘might still not like it’ in the video first group but nobody in the audio first sample reproduced this. The other least frequently reproduced propositions included only between two and three reproductions in total across the 30 papers whereby many of these were reproduced by the audio first group and not at all by the video first group (see Appendix 25).

Table 5.10: Least frequently reproduced propositions in part one across both groups

<table>
<thead>
<tr>
<th>Culture shock lecture - Expert Panel propositions</th>
<th>AUD FIRST</th>
<th>VID FIRST</th>
<th>AUD + VID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Propositions - part one</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unsure;</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>applies to everyone;</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>top left of U curve;</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>bottom of the U curve;</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>half way up right hand side;</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>might still not like it;</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>W curve: return to old country;</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>TOTAL REPRODUCTIONS PART ONE</strong></td>
<td><strong>238</strong></td>
<td><strong>169</strong></td>
<td><strong>407</strong></td>
</tr>
</tbody>
</table>

In terms of differences between the propositions reproduced (see Appendix 25 for full details), the audio first group reproduced the third and fourth stages of culture shock more frequently than the video first group with nine and five reproductions respectively for the ‘Autonomy Stage’ and ten and six respectively for the ‘Independence Stage’. In addition, ‘excited’ (n=9), ‘happy’ (n=8), ‘get to grips with the new culture / coming to terms with culture’ (n=8), ‘Hofstede’ (n=7), ‘happy with / accept new culture’ (n=7), ‘sad’ (n=6), ‘shy / withdrawn in new culture’ (n=6), ‘don’t know how things work’ (n=6) and ‘go through everything / stages of culture shock again when you re-adapt to your old culture’ (n=6) were all reproduced somewhat more frequently by the audio group than by the video group. On the other hand, the video group reproduced ‘lonely’ (n=12), ‘feelings of anxiety and insecurity caused by new culture’ (n=9), ‘insecurity’ (n=9) and ‘difficulty solving problems’ more frequently than the audio first group.
When looking at the individual results, what becomes clear is that students at the higher end of the marking scale seem to generally have produced most propositions overall. In the audio first group, with the exception of SA7, who produced only 23.53 per cent (n=12) of the propositions, the top six ranged from a reproduction percentage of 35.29 per cent (n=18) to 68.63 per cent (n=35). In the video first group, the top six students reproduced between 35.29 per cent (n=18) and 49.02 per cent (n=25) of the propositions that the expert panel identified. Conversely, at the bottom end of the scale, with the exception of 2KA20, who reproduced 54.90 per cent (n=28), the other three students at the bottom end of the scale in the audio first group reproduced between 0 per cent and 9.80 per cent (n=5) of the propositions. The students at the bottom end of the scale in the video first group first group reproduced between 1.69 per cent (n=1) and 15.69 per cent (n=8) of the propositions.

Before moving on to look at the propositions reproduced from the PowerPoint slides, it should be noted here that the propositions highlighted above may have been reproduced verbatim or in a paraphrased format whereas only exact matches from the video have been coded as such in Table 5.11. For example, we can see that the video first group sample included five instances of 'Autonomy stage' and six instances of 'Independence stage' in Table 5.9 above; however, in Table 5.11 below, there are only two instances of 'Autonomy Stage' and one instance of 'Independence stage'. The qualitative analysis reveals that student SV10, who produced both of these matches, has written the exact words; however, student 4SV8 who had also produced these in the table above, has written 'Autoneney' and 'Independent', which suggests that they wrote these from the audio rather than from the written form on the PowerPoint slides. For this reason, the figures vary between the two tables.

Looking at what the video first group actually produced from the Power Point slides then, we can see that 'loneliness' was the most popular reproduction, with ten out of the 15 sampled reproducing this word in this exact form. The two next most frequent reproductions were 'depression' (n=9) and 'insecurity' (n=9) followed by '[t]he anxiety that results from losing
Table 5.11: Part one PowerPoint propositions reproduced by the video first group

<table>
<thead>
<tr>
<th>Slides part one</th>
<th>TOTAL VID FIRST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition:</td>
<td></td>
</tr>
<tr>
<td>The anxiety that results from losing familiar signs and symbols of social intercourse</td>
<td>8</td>
</tr>
<tr>
<td>Oberg, (1954)</td>
<td>1</td>
</tr>
<tr>
<td>Symptoms of culture shock</td>
<td>1</td>
</tr>
<tr>
<td>If you suffer from culture shock, you may experience:</td>
<td>1</td>
</tr>
<tr>
<td>depression</td>
<td>9</td>
</tr>
<tr>
<td>loneliness</td>
<td>10</td>
</tr>
<tr>
<td>insecurity</td>
<td>9</td>
</tr>
<tr>
<td>a lack of confidence</td>
<td>6</td>
</tr>
<tr>
<td>an inability to solve easy problems</td>
<td>8</td>
</tr>
<tr>
<td>Stages of culture shock</td>
<td>2</td>
</tr>
<tr>
<td>Honeymoon Stage</td>
<td>2</td>
</tr>
<tr>
<td>Disintegration Stage</td>
<td>1</td>
</tr>
<tr>
<td>Reintegration Stage</td>
<td>2</td>
</tr>
<tr>
<td>Autonomy Stage</td>
<td>2</td>
</tr>
<tr>
<td>Independence Stage</td>
<td>1</td>
</tr>
<tr>
<td><strong>TOTAL REPRODUCTIONS FROM SLIDES - PART ONE</strong></td>
<td><strong>57</strong></td>
</tr>
</tbody>
</table>

familiar signs and symbols of social intercourse’ (n=8) and ‘an inability to solve easy problems’ (n=8). There were also six occurrences of ‘a lack of confidence’ but the remaining propositions were only produced by one or two test takers in each case. The majority of these were produced by the same person (5SV20) who reproduced 75 per cent (n=12) – the highest number in the video first group - of the propositions from the PowerPoint slide but only 27.45 per cent (n=14) of the propositions from the audio version. This student did not reproduce any propositions at all in part two. Similarly, student SV15 reproduced ten matches in part one, eight of which were from the slides, while only two were paraphrased clauses from the audio. This was the second highest percentage of matches at 50 per cent (n=8) of the part one matches from the slides. This student did not reproduce any matches in part two. The other student who reproduced 50 per cent (n=8) of the matches from the slides was student SV10, who reproduced 26 matches in total in part one but only ten in part two. Along with
the exact matches from the slides, this student reproduced a variety of paraphrased word (n=1), clause (n=10) and sentence level (n=4) matches as well as some exact matches from the audio at word (n=3) and clause (n=2) level. These three students were all awarded different grades. SV10 was awarded a grade in the Band 4 (equivalent to B1) while the other two were both in the Band 2 (equivalent to A1), with 5SV20 receiving a Fair Average score of 2.55 compared to 2.01 for SV15. At the top end of the scores for the video first group, there were only one, five and five reproductions from the video for the top three students respectively while at the bottom end, there were three, zero and zero respectively for the bottom three.

5.4.3.1.2 Part two

Let us now turn to the results for part two. With regard to the individual participants, the top three students in the audio first group produced 26.47 per cent (n=9), 38.24 per cent (n=13) and 58.82 per cent (n=20) respectively of the propositions whereby 58.82 per cent was the highest out of the group. Since the top scorer did not reproduce the highest number of propositions, this suggests that there were other aspects to the top scorer’s paper which helped them to do well. A closer look at the paper shows the writer’s fluency as they included an introduction to set the scene and additional ‘linking’ phrases such as ‘First of all, culture shock is not a huge problem …’ or ‘What you should pay attention to are aspects which shape one’s culture shock …’.

In the video first group, the top scorer reproduced 55.88 per cent (n=19) propositions but the three next highest scoring participants also produced somewhat fewer propositions with 26.47 per cent (n=9), 26.47 per cent (n=9) and 38.24 per cent (n=13) respectively. KV2, the top scorer, also paraphrased much of the content, rather than using exact words. Indeed, the only exact match that this person had from the slides was the sentence which was the definition of culture shock from Oberg (1954). Similarly, they started off with an introduction
and also used linking phrases to help the text hang together more naturally: ‘First of all, we need to clearly understand what culture shock is’ and ‘Other important factors are ...’.

Table 5.12 shows the most frequently reproduced propositions from part two of the audio input across both groups (see Appendix 26 for the full table of propositions reproduced in part two).

Table 5.12: Most frequently reproduced propositions from part two across both groups

<table>
<thead>
<tr>
<th>Culture shock lecture - Expert Panel propositions</th>
<th>TOTAL</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Propositions - part two</td>
<td>AUD FIRST</td>
<td>VID FIRST</td>
</tr>
<tr>
<td>Duration:</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>· depends on individual;</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>· attitude:</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>· situation:</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>· degree of difference between two cultures:</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>· steps you take to minimise it</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Minimising / reducing culture shock:</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>· Understand what culture is;</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>· Be well prepared;</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>(find out about the new culture / find out how life works before going);</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>· Learn the language;</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>· Be aware of culture shock;</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>· Integration:</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>helps with friends / make new friends;</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>find a hobby;</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>normal;</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>natural;</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td><strong>TOTALS PART TWO</strong></td>
<td><strong>145</strong></td>
<td><strong>86</strong></td>
</tr>
</tbody>
</table>

What is immediately apparent is that far more propositions were produced by the audio first group than by the video first group and this is reflected in the high number of propositions which were reproduced eight times or more, that is, by just over 50 per cent of the participants in the qualitative study. Both groups reproduced the same proposition as the most frequent reproduction. This was ‘learn the language’, which was reproduced by 13 out of the 15
participants in the audio first group and by 11 out of the 15 in the video first group. However, this was the only proposition that both groups reproduced frequently.

If we now consider the differences between the two groups, we can see that the other propositions which were produced frequently by the audio first group were again much less frequently reproduced by the video first group and vice versa. Perhaps the most striking difference is the word 'natural', which was reproduced eight times by the audio first group yet only once by the video first group. Other propositions which were frequently reproduced by the audio first group but not by the video first group are: 'be well prepared' (n=10), 'integration' (n=10), 'understand what culture is' (n=9), 'depends on individual' (n=8), 'situation' (n=8), 'be aware of culture shock' (n=8), 'natural' (n=8), 'duration' (n=7), 'attitude', (n=7), 'helps with friends / make new friends' (n=7), 'find a hobby' (n=7), 'degree of difference between the two cultures' (n=6), 'normal' (n=6), and 'steps you take to minimise it' (n=5). On the other hand, the only two propositions that were frequently reproduced by the video first group and not so much so by the audio first group were 'minimising / reducing culture shock' (n=9) and '(find out about the new culture / find out how life works before going)' (n=7). This latter proposition was in brackets as only two of the three native speakers who took the test considered this an important one yet, clearly, this was considered important by the test takers.

At the other end of the scale (see Table 5.13 below), there was again much more consistency across the groups with neither group producing 'depends on society around us', 'becomes / it is embedded; affect us', nor 'develop (better) understanding of the new culture'. Only one person in the audio first group reproduced the proposition 'takes longer if you don’t try to integrate' whereas nobody in the video first group reproduced this. In the video first group, two people mentioned ‘how we behave as individuals’ but nobody in the audio first group picked up on this. The other propositions reproduced by the video first group but not by the
audio first group were 'lasts longer if things are not good / bad? = longer’ (n=1) and ‘helps with language’ (n=1).

Table 5.13: Least frequently reproduced propositions from part two across both groups

<table>
<thead>
<tr>
<th>Propositions - part two</th>
<th>TOTAL</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUD FIRST</td>
<td>VID FIRST</td>
</tr>
<tr>
<td>takes longer if you don't try to integrate;</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>lasts longer if things are not good / bad? = longer;</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>how we behave as individuals;</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>depends on society around us;</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>becomes / it is embedded; affects us;</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>develop (better) understanding of the new culture;</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>helps with language;</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>TOTALS PART TWO</strong></td>
<td><strong>145</strong></td>
<td><strong>86</strong></td>
</tr>
</tbody>
</table>

At the bottom end of the scale, in both groups, the qualitative analysis of the papers shows that almost all of the content comes from the lecture:

- 'Culture shock causes depression. To overcome (sic) it, need do sports, to learn the language, walk and be with your friends' (3SV3):

- 'When you have culture shock, you are lonely. You feel depression. It lasts long and has stages. It is easier if you can speak the language and understand the culture’ (5SV16);

- Cultural shock is a natural phenomenon is (sic) us. The duration of a cultural shock depends on the situation. The cultural shock get rid help (sic) friends, music, theater (sic)'.

In the second part, the audio first group also had access to the video. Table 5.14 below shows the number of propositions that were reproduced exactly in the participants’ texts.
These results demonstrate that there were far fewer reproductions from the slides for the audio first group than there were for the video first group. The most frequent reproductions were ‘integration’ (n=8) and ‘preparation’ (n=7), followed by ‘linguistic mastery’ (n=5), ‘awareness’ (n=5), ‘culture shock is a natural phenomenon’ (n=5) and ‘your attitude’ (n=4), yet these reproductions were somewhat less frequent than those reproduced in part one by the video first group (which had reproductions by between six and ten people at the higher end) while the rest were only produced once, twice or not at all by the 15 candidates.

In terms of individual reproductions, the two students who produced the highest percentage of reproductions were both in the Band 4 (equivalent to B1) range. These two students reproduced 66.67 per cent (n=10) and 40 per cent (n=6) of the possible total reproductions from the video. On the other hand, the top four students (all in Band 5, which is equivalent to a CEFR B2 score) reproduced between 6.67 per cent (n=1) and 20 per cent (n=3). At the lower end of the scale, however, the results were much more mixed, varying from 0 per cent
to 33.33 per cent (n=5). Three students (in Bands 4, 2 and 2 respectively) did not have any matches to the video but had reproduced varying numbers of propositions which they had paraphrased from the audio and / or video.

5.4.3.1.3 Reproductions index

The final part of the qualitative analysis considered what percentage of words from the students’ texts overall could be attributed to the content. To do this the total number of words which the test taker had reproduced from the lecture content in each mini-summary was divided by the test taker’s word count.

The results here were very varied. In part one, the top scorer produced a ‘reproductions to total word count’ ratio of 0.26 and the lowest level test taker (after the test taker who did not produce any content to be measured) was 0.29. However, in between these two extreme scorers there were values as high as 0.79 (for the 3rd lowest test taker) and 0.61 for the 4th highest test taker and as low as 0.08 for the 5th lowest test taker. The median test taker had a ratio of 0.33. Figure 5.3 below illustrates the variation across the sample of 15 test takers in the audio first group for the audio input.

![Bar chart](image.png)

Figure 5.3: Ratio of reproductions to total word count for audio first group: audio input (part 1)
As can be seen from the graph, based on the sample selected there does not seem to be any pattern emerging. It was therefore decided not to proceed with this analysis for the remaining scripts.

For the video first group, the situation was similar with the top scoring test taker having a ratio of 0.5 while the next highest test taker had a ratio of 0.39. At the other end of the scale, the two lowest scoring test takers had ratios of 0.20 and 0.11 respectively. However, again, in between these two extremes, there was quite a spread from 0.12 for the test taker ranked 9 out of 15 to 0.83 for the 3rd lowest scoring test taker. The median test taker in this case had a ratio of 0.67. Figure 5.4 illustrates these figures and, as with the results for the audio group for part one, there does not seem to be any pattern for the video first group based on the sample of 15 test takers in this analysis.

Turning now to the results for part two, the results here are quite varied, too. These results can be seen in Figure 5.5 below.

The top scoring participant had a ratio of 0.16 while the lowest scoring participant had a ratio of 0.25. In the range in between these two extremes was from 0.00 for the 2nd lowest test taker to 0.70 for the 3rd lowest test taker. The median test taker had a ratio of 0.41.
Finally, the results for the audio input (part 2) for the video first group are also very diverse as can be seen from Figure 5.6:

In this part, the ratios ranged from 0.34 for the top scorer to 0.47 for the bottom scorer. The 2nd highest test taker had a ratio of 0.18 while the 2nd lowest test taker had a ratio of 0.71. The median test taker in this case had a ratio of 0.00 while highest ratio was 0.82 for test taker number 3.
5.4.4 Summary of RQ2

To summarise the findings of RQ2, both groups produced approximately twice as many reproductions in the first half as in the second half and the audio first group produced roughly 50 per cent more matches than the video first group in both cases.

It is clear that, when students had access to the video, they made use of it because there was a large percentage of the word and clause level exact reproductions in both groups. However, the number of sentence level reproductions in each group was relatively low.

In general, students reproduced more clause level paraphrases than any other type of reproduction; however, in part one, the audio first group produced far more word level matches than any other type of match. For the video first group, when they had access to the video in part one, word level matches from the video input were higher than the corresponding word level matches yet the paraphrased clauses from the audio still represented the highest number of reproductions.

In part two, both groups had similar results: the clause level paraphrases represented the highest number of propositions while the word and clause level exact reproductions from the audio were similar across both groups as was the number of exact clauses from the video input for the audio first group, who had access to the video in this second part. However, exact word matches from the video were much higher than exact word matches from the audio in this case.

Looking in more detail at these reproductions, the qualitative analysis of the propositions revealed that there were very few similarities between the two groups as to which propositions were frequently reproduced, although there were far more similarities in terms of which propositions they did not reproduce. Both groups tended to have a higher number of reproductions of the propositions which featured in both the audio and video input. As such, the video first group had high numbers of reproductions for the propositions which pertained to the video content while the audio group reproduced these propositions far less frequently.
The same was true of the audio group when they had the support of the video: the number of reproductions of propositions pertaining to those on the PowerPoint slides were reproduced far more frequently by the audio first group than the video first group.

Finally, in an attempt to find some pattern regarding which test takers produced which number of propositions relative to the word count, a ratio of reproduced words to total words was calculated for the 30 test takers in the qualitative sample. However, while some of the ratios could be explained due to high numbers of propositions and high word counts or low numbers of propositions and low word counts, there was no noticeable pattern across the different levels of test takers.

5.5 RQ3: How do test takers’ perceptions of their performance on an integrated EAP listening-into-writing test presented in two different input formats compare to their observed scores as awarded by trained raters?

5.5.1 Introduction
After the test, all test takers completed a post-test feedback questionnaire (see Section 4.4.3 above). This served to answer this question, on the one hand, and to provide validation evidence for the instruments on the other. Four of the questions related to this RQ. Two of these were quantitative while the other two were qualitative.

5.5.2 Quantitative analysis
In the questionnaire, questions five and seven investigate test takers’ perceptions of how they performed on the different input formats to find out how these compare to their observed performances as measured by their Fair Average score from the Rasch analysis carried out to answer RQ1 above.

Descriptive statistics were calculated using IBM SPSS (IBM Corp., 2013) and then the test takers’ responses regarding their perceptions were compared to the observed scores.

In the questionnaire, question five is: ‘Think of the two sections of the lecture. Which part do you think you performed best on?’ while question seven is: ‘Think of the two sections of the lecture. Which part do you think you performed worst on?’ The responses were compared to
their observed test performances as measured by their Fair Average scores from the Rasch Analysis carried out to answer RQ 1 above. Table 5.15 below shows the results of the comparison for both of these questions for the audio first group.

As can be seen from Table 5.15, 23 of the test takers thought they did better on part one but the actual number was slightly higher at 26. On the other hand, 30 students thought they performed better on part two whereas actually only seven students performed better on this section. However, there were 25 students who scored equally well on both parts. Five students did not respond to the questions in the questionnaire.

Table 5.15 Audio first group - Perceived performance compared to observed test performance

<table>
<thead>
<tr>
<th>Audio first group</th>
<th>Part one</th>
<th>Part two</th>
<th>Equally good</th>
<th>No response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived best (no. of students)</td>
<td>23</td>
<td>30</td>
<td>--</td>
<td>5</td>
</tr>
<tr>
<td>Observed best (no. of students)</td>
<td>26</td>
<td>7</td>
<td>25</td>
<td>--</td>
</tr>
<tr>
<td>Word count best (no. of students)</td>
<td>48</td>
<td>8</td>
<td>2</td>
<td>--</td>
</tr>
<tr>
<td>Perceived worst (no. of students)</td>
<td>28</td>
<td>25</td>
<td>--</td>
<td>7</td>
</tr>
<tr>
<td>Observed worst (no. of students)</td>
<td>7</td>
<td>26</td>
<td>25</td>
<td>--</td>
</tr>
<tr>
<td>Word count worst (no. of students)</td>
<td>8</td>
<td>48</td>
<td>2</td>
<td>--</td>
</tr>
</tbody>
</table>

Since the scores on the test did not provide enough information to answer the question due to the fact that a high number of students had been awarded the same score for both halves, I decided to analyse the word count for each section to investigate whether this might shed some light on the similarity between students’ perceptions and observed performance. As can be seen from Table 5.16 below, when the word count was considered, a large proportion of the 25 students who received the same score on both halves were found to have performed better, i.e. had a higher word count for part one (n=22). One person had scored better on part two and two students had also written exactly the same number of words in both mini-summaries.
However, it was still not entirely clear if the people who said they had performed better on a given section were indeed the people that did perform better so I decided to take the analysis one step further by doing a person-by-person match count. These results were as follows:

Table 5.16: Audio first group - Person by person count of matches between perceived and observed best performances

<table>
<thead>
<tr>
<th>PERCEIVED BEST VERSUS OBSERVED BEST</th>
<th>Score match total</th>
<th>Word match total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio first</td>
<td>17</td>
<td>30</td>
</tr>
<tr>
<td>No match</td>
<td>34</td>
<td>21</td>
</tr>
<tr>
<td>No response</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Same score and word count</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5.16 shows that, across both input formats, only 17 out of a total of 58 people’s perceptions matched with observed best performances as measured by the scores. However, when the word count was taken into consideration, this number increased to 30 matches. This meant that, in the case of the score match, 41 people in the audio first group did not predict their ability correctly based on the scores; however, this fell to 28 people once the word count matches were calculated.

Table 5.17: Audio first group - Person by person count of matches between perceived and observed worst performances

<table>
<thead>
<tr>
<th>PERCEIVED WORST VERSUS OBSERVED WORST</th>
<th>Score match total</th>
<th>Word match total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio first</td>
<td>18</td>
<td>32</td>
</tr>
<tr>
<td>No match</td>
<td>34</td>
<td>19</td>
</tr>
<tr>
<td>No response</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Same score and word count</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Looking at the results for the worst performance (see Table 5.17 above), these results show that slightly more people were able to judge their worst performances with 18 and 34 test takers predicting their performance correctly based on the score and word count respectively.

For the video first group, the results are shown in Table 5.18 below.

Table 5.18: Video first group - Perceived performance compared to observed performance

<table>
<thead>
<tr>
<th>Video first group</th>
<th>Part one</th>
<th>Part two</th>
<th>Equally good</th>
<th>No response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived best (no. of students)</td>
<td>42</td>
<td>11</td>
<td>--</td>
<td>5</td>
</tr>
<tr>
<td>Observed best (no. of students)</td>
<td>31</td>
<td>6</td>
<td>21</td>
<td>--</td>
</tr>
<tr>
<td>Word count best (no. of students)</td>
<td>43</td>
<td>13</td>
<td>2</td>
<td>--</td>
</tr>
<tr>
<td>Perceived worst (no. of students)</td>
<td>12</td>
<td>39</td>
<td>--</td>
<td>7</td>
</tr>
<tr>
<td>Observed worst (no. of students)</td>
<td>6</td>
<td>31</td>
<td>21</td>
<td>--</td>
</tr>
<tr>
<td>Word count worst (no. of students)</td>
<td>13</td>
<td>43</td>
<td>2</td>
<td>--</td>
</tr>
</tbody>
</table>

The results show that 42 of the video first group thought they performed better on part one while 11 thought that they performed better on part two. Five people did not respond. In reality, 31 people scored better in part one and 6 in part two. However, 21 students scored equally well on both mini-summaries. On the other hand, 12 people felt that part one was their worst performance and 39 felt that part two was their worst. Seven people did not respond to this question. The observed results revealed that while only 6 performed worse on part one, 31 received a lower score for part two. However, there were again 21 students who scored equally well on both parts.

When the word count was taken into consideration, an additional 12 people were found to have performed better on part one while an additional seven were found to have written more on part two. Two people in this group had also written the exact same number of words in both mini-summaries. This resulted in a much closer match between students’ perceptions and reality but, as we saw above, this does not tell us whether the people who predicted their best and worst performances as part one or part two respectively were indeed the people who
performed best on these parts so a person-by-person analysis was carried out for this group, too. These results are shown in Table 5.19 below.

Table 5.19: Video first group - Person by person count of matches between perceived and observed best performances

<table>
<thead>
<tr>
<th>PERCEIVED BEST VERSUS OBSERVED BEST</th>
<th>Score match total</th>
<th>Word match total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video first</td>
<td>24</td>
<td>36</td>
</tr>
<tr>
<td>No match</td>
<td>27</td>
<td>15</td>
</tr>
<tr>
<td>No response</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Same score and word count</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

These findings reveal that the video first group were better at predicting their performance, with 24 people correctly predicting their performance when compared with their scores and 36 correct predictions based on the word count. Nonetheless, this meant that 27 and 15 students respectively were unable to judge their performance correctly.

With regard to their worst performance, one person fewer in each case predicted their performance correctly with 23 and 35 people predicting their worst performance based on score and word count respectively as shown in Table below.

Table 5.20 Video first group - Person by person count of matches between perceived and observed worst test performances

<table>
<thead>
<tr>
<th>PERCEIVED WORST VERSUS OBSERVED WORST</th>
<th>Score match total</th>
<th>Word match total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video first</td>
<td>23</td>
<td>35</td>
</tr>
<tr>
<td>No match</td>
<td>26</td>
<td>14</td>
</tr>
<tr>
<td>No response</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Same score and word count</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
In summary, it would appear that students’ perceptions often do not match with their observed scores; however, more agreement was found when the word count for each of the two parts was taken into consideration.

5.5.3 Qualitative analysis

5.5.3.1 Results: feedback questionnaire

Following on from the quantitative questions above which asked which of the two sections test takers felt they performed best and worst on respectively, the two qualitative questions correspondingly asked ‘Why do you think you did best on that part?’ (Q.6) and ‘Why do you think you did worst on that part?’ (Q.8).

Qualitative responses were received from 90 test takers for the first question. The largest group of test takers (n=38) thought that they performed better in the video part because the text and/or the visuals helped them while another 9 test takers felt that the ‘video was better perceived’ (see Figure 5.7). They commented that they felt they performed better because they were able to ‘see slides in the presentation (2SA3), because ‘there is a visual material’ (KA7), because they ‘could both listen and read information’ (SV10) and because they ‘could read something [they] didn’t understand on audio’. One test taker explained their preference for the video very clearly:

‘Because there were slides that improved the understanding and because visual perception helped to focus’ (2SV8).

On the other hand, the next largest group of respondents (n=18) expressed the opinion that listening only was easier. Several participants stated that they were ‘focused only on listening’ and one participant felt that they ‘didn’t waste [their] concentration on the pictures’.

The next highest opinion regarding the best performance was that participants said they felt they understood the content of part two better (n=7) while others said they understood the content of part one better (n=3). Two participants said that they found the topic interesting and one test taker said that the clear presentation structure helped them to perform better.
There were 7 incomprehensible / miscellaneous responses and five people said they did not know why they thought they had performed better on the part in question.

The second qualitative question (see Figure 5.8 below) looked at reasons why the test takers felt they had performed worse on the given section. Again, there were responses from 90 test takers because 26 did not respond. In this case, 24 participants (21 per cent) felt they performed worse on the audio only section as listening only is more difficult. One person said of the audio only section:

‘it don’t concentration (sic) my attention’ (2SA4)

while another stated:

‘it’s difficult for me to understand information from audio’ (4SA4).

One test taker said:

‘the audio information is much more difficult’ (SV9)

and another stated:
‘I badly remember information when I listen’ (SA6).

These differing responses are examples of four different reasons why the test takers found the listening only part more difficult.

Conversely, 16 respondents (14 per cent) said watching and listening was difficult or they were distracted by the video. Several respondents stated that they were distracted by reading the presentation on the video (KA6; KA8; 2KA18; 2KA19; KV7) whereas others seemed to realise that they wasted time:

‘because I spent some time for reading information’ (KA4)

and others stated that:

‘it’s difficult for me to make notes while watching’ (2KA3).

A large number of participants (n=20) admitted that they performed worse on a given part because they did not understand the content while eight said they had been unable to concentrate or had difficulty concentrating. Five participants felt that the speech rate was too
quick and four thought that there was too much information. Two people said that they felt there was a lack of time while another two said the 'video was better perceived'. On the other hand, only one person said that listening only was easier. One person said the acoustics had been bad and one person said they thought they performed the same on both parts. There were four incomprehensible / miscellaneous answers and two people said they did not know why they had performed worse on the part in question.

5.5.4 Summary of RQ3 results
To summarise the results of RQ3, 41 and 28 participants out of 58 in the audio first group were unable to judge their best performance compared to their score and the word count respectively. With regard to the worst performance, the figures were similar in that 40 misjudged their worst performance compared to the score and 26 people misjudged their performance compared to the word count. In the video first group, 34 out of 58 were not able to predict their best performance compared to the score and 22 were unable to predict their performance compared to the word count. Again, the figures for the worst performance were similar with 35 out of 58 unable to predict their performance correctly compared to the score and 23 out of 58 unable to make the correct prediction compared to the word count.

With regard to the qualitative analysis, 41 per cent of the test takers felt that they were helped by the text / visual (33 per cent) and that video was better perceived (8 per cent). Conversely, 15 per cent felt that listening was easier. 21 per cent stated that listening only was more difficult while 14 per cent felt they were distracted by the video.

5.6 Feedback questionnaire
5.6.1 Introduction
In addition to the four questions discussed above, test takers were asked to provide responses to a number of other questions to enable the researcher to collect validation evidence or to highlight aspects of the test which still need addressing. The questionnaire was analysed quantitatively and qualitatively and the results will be presented briefly here.
5.6.2 Quantitative results

5.6.2.1 Descriptive statistics

Descriptive statistics (frequencies) for the questions which are analysed quantitatively are presented in this section.

- Q1. How well did you do on the listening test?

The results in Figure 5.9 show that 21.6 per cent (n=25) felt they understood much more (1) or more (2) than they normally would in a listening test while 44.8 per cent (n=52) felt they did as well as ever (3). However, 25.9 per cent felt they did worse (4) or much worse (5) in the listening test. Nine people did not respond to the question (6).

Figure 5.9: Q1. How well did you do on the listening test?
• Q2. How difficult was the writing test?

Figure 5.10: Q2. How difficult was the writing test?

Figure 5.10 above shows that, cumulatively, 34.5 per cent (n=40) of the test takers found the EAP writing test very easy (1) or quite easy (2) while 43.1 per cent (n=50) found it quite difficult (3) and 15.5 per cent (n=18) found it very difficult (4). Eight people did not respond to the question (5).

• Q3. How much did you know about the topic?

Figure 5.11: Q3. How much did you know about the topic?

Question 3 investigated the test takers’ topic knowledge. As can be seen from Figure 5.11 per cent (n=14) knew a lot about the topic (1) while 45.7 per cent (n=53) knew a little (2). 35.3
per cent (n=41) knew nothing about the topic (3) and there were eight people who did not respond (4).

- Q4. What effect did the topic have on your performance?
The pie chart in Figure 5.12 below shows that 55.2 per cent (n=64) of test takers thought that the topic had a positive impact on their performance while 25 per cent (n=29) thought it had no impact. Only 2.6 per cent (n=3) thought it had a negative impact on their performance while 9.5 per cent (n=11) did not know what effect it had had. Again, there were nine people who did not respond to this question.

![Pie chart showing the distribution of responses to Q4](image)

**Figure 5.12: Q4. What effect did the topic have on your performance?**

Questions 5, 6, 7 and 8 were discussed as part of RQ3 above so they will not be discussed again here.

- Q9. Did anything else have a positive impact on your understanding of the lecture?
This pie chart in Figure 5.13 below reveals that seeing text related to the content had the biggest impact on the test takers’ performance with 44.8 per cent (n=52) of people selecting this option while another 25.9 per cent (n=30) said that seeing the speaker had a positive impact. 20.7 per cent (n=24) said they were helped by the diagram while 8.6 per cent (n=10) did not respond.
Q9. Did anything else have a positive impact on your understanding of the lecture?

- Q10. Did anything have a negative impact on your understanding of the lecture?

The results of Q10, shown in the bar chart in Figure 5.14 below, reveal that the sound quality was noted as having a negative impact on performance for 23.3 per cent (n=27) of the participants while the video quality accounted for another 4.3 per cent (n=5). The number of times participants heard the video and the amount of information they had to write down also had a negative effect on performance for 20.7 per cent (n=24) in each case. Other factors were also mentioned as having a negative impact on performance for 4.3 per cent (n=5) of the respondents. 31 participants did not respond to this question.

The other factors were qualitative answers that the students added to this question. Twelve comments were made. The most frequent of the comments (n=4) was ‘too little information’. While it is not entirely clear what test takers meant by this, we might assume that they mean they were not able to understand enough information to write the text. Three people commented that they wanted to hear the lecture again. Two other comments related to external factors: poor eyesight and noise outside the room while one person each made the following comments: ‘the speech was too fast’, ‘they understand very little English’ and ‘the audio text was difficult’.
Figure 5.14: Q10. Did anything else have a negative impact on your understanding of the lecture?

- Q11. Overall, how difficult was the essay?

Figure 5.15: Q11. Overall, how difficult was the essay?

The results for this question, shown in Figure 5.15 above, reveal that, overall, test takers found the essay quite difficult with 21.6 per cent (n=25) saying it was very difficult (1) and 31 per cent (n=36) who thought it was quite difficult (2). 35.3 per cent (n=41) thought the
essay was quite easy (3) overall and 3.4 per cent (n=4) thought it was very easy (4) while 8.6 per cent (n=10) did not provide any response (5) to this question.

- Q12 was a qualitative question so that will be analysed in Section 5.6.3.1 below.

5.6.2.2 Summary of quantitative results
To summarise the quantitative results of the questionnaire, over 66 per cent of test takers performed as well as they would normally or better on the listening test. For the writing test, over 40 per cent of test takers regarded the writing test as quite difficult while 30 per cent found it quite easy. As the targeted level is the border between B1 and B2 and these two levels are where the majority of the test takers are, this seems to suggest that the test is testing at approximately the correct level. Many test takers knew a little or nothing about the test but the majority of test takers felt the topic had had a positive impact on their performance. Test takers also felt that the visual support had had a positive effect on their performance with nearly 45 per cent stating that the text had been helpful for them. On the other hand, over a quarter of people felt that the sound and video quality might have had a negative impact on their performance along with the amount of information they had to write down and the number of times they heard the video. Finally, over 50 per cent of the test takers found the test challenging, rating it either very difficult or quite difficult.

5.6.3 Qualitative results
5.6.3.1 Results
Apart from the two open-ended questions which were discussed in relation to RQ3, there was one other open-ended question, the results of which are presented here. This question aimed to collect validation evidence for the EAP test by investigating potential sources of construct-irrelevant variance.
Q12. Any other factors that affected your performance?

Table 5.21 below shows the summary of the qualitative comments. There were comments from half of the participants while the other half either gave no response or stated that nothing else had had any effect on their performance.

As can be seen, the largest proportion of positive comments related to the topic with 12.1 per cent (n=15) stating that they were interested in the topic or the topic made the experience positive while others stated that their familiarity with the topic also positively affected their performance:

- ‘I had positively affect (sic) because it was so interesting experinace (sic) for me.’ (4SA5)
- ‘I would write more than 350 words (what (sic) I probably did), because I find this topic interesting’ (KA10)
- ‘Yes, greatly influenced by the factor (sic) that I was familiar with the topic of the lecture’ (2SV8)
- ‘Yes, the topic was interesting’ (KV3)

Other positive comments related to the diagrams (4.3 per cent, n=5), the positive impact of the lecturer’s accent (3.4 per cent, n=4), the structure of the presentation (3.4 per cent, n=4) and comments referring to the fact that the visuals were helpful (n=4):

- ‘Diagrams in text helped me understand text lecure (sic).’ (SV4)
- ‘The accent and the voice of the lecturer and the topic (interesting) affected positively (sic).’ (KA6)
- ‘Well-structured information in video and audio helped me to build well (sic) the structure of my summary.' (2KA10)
- ‘I have written everything I wanted to write down. Seeing the speaker affected positively on my writing.’(2KA4)
Table 5.21: Q12. Any other factors that affected your performance?

<table>
<thead>
<tr>
<th>Valid responses</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>interesting topic / positive experience / familiarity with the topic</td>
<td>15</td>
<td>12.1</td>
<td>12.1</td>
<td>12.1</td>
</tr>
<tr>
<td>out of time</td>
<td>9</td>
<td>7.8</td>
<td>7.8</td>
<td>19.9</td>
</tr>
<tr>
<td>positive impact of lecturer’s accent</td>
<td>4</td>
<td>3.4</td>
<td>3.4</td>
<td>23.3</td>
</tr>
<tr>
<td>incomprehensible / miscellaneous response</td>
<td>6</td>
<td>5.2</td>
<td>5.2</td>
<td>28.5</td>
</tr>
<tr>
<td>poor English</td>
<td>5</td>
<td>4.3</td>
<td>4.3</td>
<td>32.8</td>
</tr>
<tr>
<td>diagrams were positive</td>
<td>5</td>
<td>4.3</td>
<td>4.3</td>
<td>37.1</td>
</tr>
<tr>
<td>presentation structure</td>
<td>4</td>
<td>3.4</td>
<td>3.4</td>
<td>40.5</td>
</tr>
<tr>
<td>speech rate</td>
<td>2</td>
<td>1.7</td>
<td>1.7</td>
<td>42.2</td>
</tr>
<tr>
<td>video helped</td>
<td>2</td>
<td>1.7</td>
<td>1.7</td>
<td>43.9</td>
</tr>
<tr>
<td>external factors (noise)</td>
<td>2</td>
<td>1.7</td>
<td>1.7</td>
<td>45.6</td>
</tr>
<tr>
<td>too much info</td>
<td>2</td>
<td>1.7</td>
<td>1.7</td>
<td>47.3</td>
</tr>
<tr>
<td>seeing the speaker</td>
<td>1</td>
<td>0.9</td>
<td>0.9</td>
<td>48.2</td>
</tr>
<tr>
<td>sound and/or video quality helped</td>
<td>1</td>
<td>0.9</td>
<td>0.9</td>
<td>49.1</td>
</tr>
<tr>
<td>difficult topic</td>
<td>1</td>
<td>0.9</td>
<td>0.9</td>
<td>50</td>
</tr>
<tr>
<td>no effect / no response</td>
<td>58</td>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>116</strong></td>
<td><strong>100</strong></td>
<td><strong>100</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

On the other hand, there were several comments which highlighted aspects which would have had a negative impact on performance. These included 7.8 per cent (n=9) of test takers stating that they ran out of time, test takers’ poor English skills (4.3 per cent, n=5), external factors such as the volume or outside noise (1.7 per cent, n=2) as well as the speech rate, difficulty of the topic and the amount of information (n=6):
• ‘I could write more in the essay. Unfortunately I do not have a lot time for this.’ (4SV6)
• ‘I don (sic) not know English well’ (SA2)
• ‘The sound was quite low, so it may affect negatively on quality of my essay.’ (KV9)
• ‘... lecture was fast, so writing notes was difficult’ (SA25).

5.6.3.2 Summary of qualitative results
In total, there were 58 comments and 32 of these were positive. Almost half of the positive comments (n=15) related to the topic or the fact that test takers found taking the test a positive experience while the other positive comments related to positive features of the lecture itself. On the other hand, a large proportion of the negative comments relate to students having run out of time.

5.6.4 Summary of feedback questionnaire results
All in all, the feedback questionnaire has provided valuable evidence of the concurrent validity of the test and has provided important feedback regarding the appropriateness of the topic and the usefulness of the visuals, which, combined with the findings from RQ3, suggests that the use of video is perhaps more valuable than audio only.

At the same time, the questionnaire had highlighted areas for improvement, specifically quality of the video and sound, and the length of the test, which need to be addressed before such a test can be operationalised.

5.7 Conclusion
RQ1 investigated how performance on an integrated listening-into-writing EAP lecture differs when test takers are presented with lecture input in the form of a podcast or a vodcast. The results demonstrated that there was a statistically significant difference in performance on the audio part (podcast) between the audio first group and the video first group, yet there was no statistically significant difference between the two groups when the input was presented as video input with textual and diagrammatic support (vodcast). The results to this question
also showed that both groups performed better on the first part, which might suggest a fatigue effect on the second part whereby the first group are supported by having access to the video while the lack of visual support in the second part of the lecture may have led to the performance of the video first group suffering somewhat more than that of the audio first group.

The textual analysis for this question, carried out using Text Inspector (Text Inspector, 2019) showed that, in general, the audio first group performed best with the audio input while the video first group performed worst on the audio input. In contrast to the statistical analyses, however, the Text Inspector analysis showed that the video first group sometimes outperformed the audio first group on the analyses relating to the video input. With regard to the CEFR levels, however, the video first group often had more CEFR vocabulary from the higher levels in the mini-summaries pertaining to the video input than the audio first group had in their mini-summaries pertaining to the audio input. This was also the case for the AWL types and tokens. However, both of these could have been affected by the differences in word counts.

RQ2 investigated the extent to which test takers are influenced by the written word in an EAP lecture compared to the spoken word. The results for this question show that both groups produced around twice as many reproductions in the first part compared to the second part and the audio first group produced around twice as many as the first group in both cases. This is in line with findings from RQ1 above.

Although the audio first group produced more word level matches from the audio than any other reproductions across both groups, it was clear that, when students had access to the video, this was being widely used as the number of exact word and clause level reproductions from the video was high in both groups; however, sentence level reproductions were low. When the video first group had access to the video, they produced more word level matches
from the video than they did from the audio, yet the clause level paraphrases from the audio represented the highest number of reproductions overall.

In part two, however, the results were similar across both groups: the highest number of reproductions related to clause level paraphrases from the audio while both groups had a similar number of exact word and clause level reproductions from the audio. However, the exact word reproductions from the video was higher than exact word matches from the audio. In short then, both the audio first group and the video first group reproduced more word level matches from the video but the audio first group produced more exact word level matches from the audio in part one. Clause level paraphrases from the audio were reproduced more widely by both groups and sentence level reproductions were minimal by both groups.

The qualitative analysis revealed that in both cases, the number of people in each group reproducing a given proposition was generally higher in the group that had access to the video if the proposition was on the video; however, the converse was also true: if the proposition was not one of those mentioned in the video, it was most frequently reproduced by the group which had only audio input.

There seemed to be no noticeable pattern regarding the proportion of words reproduced from either input according to the different levels of proficiency.

Finally, RQ3 looked at how test takers’ perceptions of their performance on an integrated EAP listening-into-writing test compared to their observed scores as awarded by trained raters. These findings suggest that there is relatively little agreement between test takers’ perceptions of their performance compared to their scores; however, there was more agreement when test takers’ perceptions were compared to the word counts in the mini-summaries but this still amounted to only around 20-40 per cent of test takers whose perceptions matched the reality of their performance as measured by their scores and/or the performance as measured by the word counts. In terms of their preferences for the input medium (irrespective of their
observed performance), 41 per cent felt they benefitted from having access to the visuals while 21 per cent said that the video distracted them or they found it more difficult to read, write and listen. While this suggests that there is no real definitive answer as to which input format is best, the findings from the post-task questionnaire lend weight to the fact that it is more helpful to have access to the video.

With regard to the post-task questionnaire, this provided some concurrent validity evidence for the test in terms of it appearing to target the correct level while also demonstrating that a good proportion (45 per cent) of the test takers felt they benefitted from the visual support and 15 per cent stating that they benefitted from the choice of topic. On the other hand, around 25 per cent of the test takers commented that the video quality might have negatively impacted on their performance and nearly 8 per cent felt they did not have enough time for the test.
6. Discussion

6.1 Introduction

This chapter will discuss and offer explanations for the results in relation to the three RQs, linking these to other studies in the field.

Section 6.2 comments briefly on the results of the QPT then section 6.3 will consider the results of RQ1, looking first at the three quantitative analyses one by one and then bringing together the results in the final section of 6.3. Section 6.4 will discuss the results for RQ2, which were qualitative analyses. These will also be discussed one by one before bringing everything together at the end of section 6.4. Section 6.5 will analyse both the quantitative and qualitative results of RQ3 before summarising the results in the final part of 6.5. Section 6.6 will discuss the quantitative and qualitative results of the post-task questionnaire and finally section 6.7 will conclude the Discussion chapter.

6.2 Quick Placement Test (OUP, 2001)

6.2.1 Quantitative analyses

The descriptive statistics showed that, although there was a difference between the medians of the two groups, they otherwise appeared to be similar in level in terms of the range of scores, the mean and the standard deviation. Similarly, the independent samples $t$-test revealed that there was no statistically significant difference between the two groups.

It was important to establish whether the differences between the groups was significant or not because if the groups had been very different, this may have skewed the results in the main integrated EAP task. As such, despite the fact that data were not available for all test takers, we can be reasonably comfortable that the differences in the main task are not due to one group being significantly stronger than the other.
6.3 RQ1: How does performance on an integrated listening-into-writing EAP task differ when test takers are presented with lecture input in the form of a podcast or a vodcast?

6.3.1 Quantitative analyses

6.3.1.1 FACETS analysis – inter-rater reliability

The results show that the three raters were all within the 0.5 to 1.5 range that is considered 'productive for measurement' (Linacre, 2012, p. 11) so the rater reliability was good. This is probably due to the fact that the three raters all worked at the same institution and used a rating scale based on the CEFR in their daily work. As such, they were all familiar with the rating scale in general but the in-depth familiarisation that they had as part of the training should have also contributed to a better shared understanding of the rating scale and the levels of performance required for a given band score (Trace, Meier and Janssen, 2016, p. 41). This subsequently resulted in a high level of exact matches between the different raters while at the same time behaving as independent raters.

The texts produced by the test takers were marked using the CEFR Overall Written Production descriptors. Although the inter-rater reliability was good, the raters, did, however, struggle with ‘the restrictive nature of the descriptors’ (Rater 3) as they were not specifically designed for this type of task and this led to the need to interpret the descriptors quite loosely in some cases, for example, the B2 descriptor mentions ‘synthesising and evaluating information and arguments from a number of sources’ (Council of Europe, 2009, p. 138). In this case, there was only one source so there was no opportunity to synthesise across sources; however, this was taken to mean synthesising the information from both input formats, even though there was only one source, namely the lecture. In terms of evaluating the sources, this was interpreted as evaluating which parts of the input were key points and therefore needed to be reproduced. However, the C1 and C2 descriptors proved very difficult to apply due to the fact that they ‘defied interpretation for the task’ (Rater 3). The reason for this was the emphasis, at this level, not only on ‘clear, well-structured texts’, but also on ‘... expanding and
supporting points of view at some length with subsidiary points, reasons and relevant examples ...’.

While all three raters did, in fact, award a total of 7 C1 grades (which were ‘downgraded’ in the Fair Average scores calculated by FACETS) across all the texts, there appeared to be something of a ‘ceiling effect’ caused by the task requirements. Davies et al., (1999, p. 19) define a ‘ceiling effect’ as the situation which arises when most of the candidates ‘get a score near the top of the scale on a particular test’, which results in the test not discriminating sufficiently at the higher levels. They point out that this can be addressed by using ‘more detailed descriptors’ or training the raters better.

In this case, the use of other CEFR scales such as the ‘Processing text’ and the ‘Reports and essays’ scales may have been beneficial and may have provided useful insights. However, this was not possible due to the additional costs of rating that this would have incurred whereas using a single holistic rating scale was a relatively fast process (Hughes, 2003, p. 95). Furthermore, using several scales may have lowered rater reliability. Another possibility would have been to design a specific rating scale for such integrated tasks, but that would have required separate trialling and validation, which was beyond the scope of this research project. Nonetheless, using specifically designed, detailed rating scales has been shown to result in higher rater reliability than using scales which were not designed specifically for the task in hand (Knoch, 2009, p. 275).

Thus, while the rating scale worked well overall, a different choice of rating scale(s) may have been more appropriate. Indeed, since the rating was carried out for this project, the Companion Volume to the CEFR has been published, which specifically includes mediation scales such as ‘Relaying specific information in writing’ and ‘Processing text in writing’. Although these scales were not designed specifically for this task, they were designed for integrated task types so they may have led to even higher rater reliability.
With regard to the use of the source text language and the impact of this on scores, some research has looked into this (Gebril and Plakans, 2016); however, in this case, the impact of source text vocabulary was specifically not taken into account in the scoring as the focus was on the impact of the input medium on the texts they produced as the integrated listening-into-writing task rather than on the use of language from the source text, which was investigated separately in RQ2.

6.3.1.2 Hotelling’s $T^2$ analysis – effect of the different input media on performance

The mean scores for Hotelling’s $T^2$ for both groups and input formats show that the audio first group scored more highly than the video first group irrespective of the input format; however, both groups received higher scores in their first input format than in the second. Furthermore, there was a statistically significant difference in audio scores between the audio first group and the video first group but there was no statistically significant difference between scores on the video input. These findings are at odds with Gruba’s (1993) study of the difference between audio and video in an EAP lecture-based listening comprehension test, which did not reveal any significant difference between the two input media.

There could be several explanations for these findings. One possible reason for the drop in performance in the second half of the essays is that the participants may have found the input in the second half of the lecture more difficult to comprehend despite the efforts made to balance the linguistic complexity and the content. As a result, participants may have needed the video to support their understanding of the concepts.

Another reason for the difference between performance in the first half of the essay and the second could be fatigue since ‘effortful listening’ can result in stress and fatigue (McGarrigle et al., 2017, p. 95). Mental fatigue, in turn, is demonstrated through a slowdown and reduction of cognitive functions following an intensive, extended period of mental effort (DeLuca, 2005, in McGarrigle et al., 2014, p. 5). If there was a fatigue effect, students may have lost concentration as they listened to the second part. If they started to lose enthusiasm or were
getting tired when taking notes on the second part, this would have led to their having less
to write about in the second part of the essay and would consequently have affected their
scores on that part. A fatigue effect has been shown to occur if the items in a test become
more difficult the later they come in a test (Kingston and Dorans, 1984, in Debeer and Janssen,
2013, p. 164). While a good proportion of test takers (n=57) scored lower on the second part
(irrespective of input mode), just over half performed equally well or better (n=59), so, as
Debeer and Janssen, (2013, p. 177) suggest, if not everyone is affected by the order effect,
the effect caused by the position of an item could be ‘a person-specific trait … rather than a
generalized “fatigue” effect’. Nonetheless, as mentioned in Section 5.3.2.2 above, more people
performed worse on the second part when this was the audio input (n=32) than when the
second part was video input (n=25) so the change of format from having a visual stimulus to
then having none may have led to more of a fatigue effect than when the change was from a
non-visual stimulus to a visual one. Alternatively, this could have been a similar effect to the
study by Locke (1977, p. 93) whereby students simply failed to write down the information
which was not written down in front of them.

Similarly, a drop in motivation may have affected the scores. This could be one reason why
the video first group scored considerably lower in the audio input than the video input. While
changing from audio to video may have been enough to keep the audio first group engaged,
the change from video to audio only may have led to a decrease in motivation (either
conscious or unconscious) which resulted in a lack of concentration (Goh, 2000) and thus a
drop in performance.

A further possible explanation is simply that test takers may have run out of time when writing
the second half of the essay, especially if they had written a lot of detail for the first half of
the essay.
6.3.1.3 Text Inspector analyses

The results for the sentence count, average sentence length, token count and type count are all highest for the audio first – audio input group (group 1) and are all lowest for the video first – audio input group (group 4). These findings concur with those of Hotelling’s $T^2$.

On the other hand, for the audio input, the video first group (group 4) had almost as high a range in the sentence count (28) as the audio first group (group 1) (29). However, a closer look at this data shows that the mean was considerably lower, as were the median and mode. In fact, both the median and the mode were lower than the mean, indicating that the data was positively skewed (Green, 2013, p. 45). This highlights the fact that, although there were some very strong students in group 4, the majority were much weaker than the mean. At first glance, then, this looks like the textual analysis does not concur with the results given by the trained raters. However, when the median is taken into account, we can see that group 2 has a higher median, which suggests that, although group 3 had some very strong students, which took the mean higher, there were more stronger students in group 2 than in group 3. These findings are in line with the results from the SPSS analysis.

With regard to the type and token counts, groups 1, 2 and 4 all had minimum token counts of 1 and minimum type counts of 1 while group 3 had a minimum token count of 5 and a minimum type count of 5, so all of these short texts will have had an effect on the type token ratio, as a type-token ratio of 1.000 is easily achieved on a very short text which has no repeated words. Add to that the fact that the median token count for group 4 is only 49 and the median type count is 37.5, compared to a median token count of 226 and a median type count of 117 respectively for group 1, it is not surprising that the type-token ratio for group 4 is higher than for group 1.

In terms of the results for the Flesch Reading Ease, the video group produced the most complex texts as measured by the Flesch Reading Ease for the video input (the first part) but this dropped off considerably in the second part (audio input). Conversely, while the audio
first group maintained very similar levels of complexity across both parts of the test, their second part is marginally more complex than the first part on average. This might suggest that the video input helped test takers in both groups to understand the concepts better and therefore produce more complex output for the video input part while the drop in performance in the second part for the video first group could be due to a fatigue effect that some participants in the video group seem to have experienced (McGarrigle et al., 2017), thereby leading to fewer notes and thus less to write. This may have been confounded by the change from the video stimulus to the audio only input.

In terms of the MTLD measure, it is worth highlighting that, despite the statistically significant difference in scores awarded by the raters, the MTLD lexical diversity measures on the mini-summaries which related to the audio texts (the first part for group 1 and the second part for group 4) are very similar across the two groups with $M = 67.692$ $SD = 17.243$ for the first group and $M = 67.420$ $SD = 21.284$. This is in stark contrast to the Flesch Reading Ease measure, which evaluates the group 4 texts as much easier for the audio part than it does for group 1, whereas the type-token ratio suggests that the group 4 texts are lexically more diverse than the group 1 texts.

The three contrasting findings are due to the different ways that these three measures are calculated (see Section 2.9.1.2). The differences between the three measures is confounded by the fact that Text Inspector only calculates lexical diversity on texts of 100 words or more. As such, the shorter texts in each group will have been discounted by Text Inspector and only those which have over 100 words will have been used in the calculation. Consequently, it is important to note how many texts were included in the calculations in each group. Bearing in mind that the group size (n=58) is the same across all four groups, in group 1, only 48 of the 58 papers were analysed. This steadily dropped over the four groups to n=42 in group 2, n=40 in group 3 and n=29 in group four. As a result of the shorter texts having been removed from the analysis, the remaining texts returned an average measure which is quite similar.
across all groups. Flesch Reading Ease and Type-token ratio, on the other hand, have been calculated for all texts. However, as outlined in Section 2.9, several measures of lexical diversity, among them, type-token ratio, are sensitive to text length so caution is recommended when interpreting this figure (McCarthy and Jarvis, 2010, p. 381). On the other hand, MTLD is independent of text length (provided the text is longer than 100 words) so may be more useful as a measure (McCarthy and Jarvis, 2010, p. 381). It is therefore to be expected that the results across these three measures are quite different and they highlight the danger of taking any one of them as a single measure of lexical diversity. On the other hand, when calculating lexical diversity using MTLD in Text Inspector, it is important to take into consideration the number of scripts which have actually been included in the analysis, rather than assuming that the mean is based on the whole group, which was clearly not the case here.

Bearing in mind that the MTLD did not measure all of the texts, of the ones that were included in the calculation, the level of lexical diversity among those students across the four groups who wrote over 100 words in their texts appears to be quite similar.

For the second part of the Text Inspector analysis, the percentage of types at different CEFR levels were presented (see Table 5.2: Percentage of CEFR types reproduced in Section 5.3.2.3).

These findings highlighted the fact that both groups used more A1 types in the second part of their texts than in the first part. Given that the percentage of A1 types in the original lecture was also higher in part two (52.45 per cent) than part one (45.97 per cent), this is to be expected. Similarly, test takers produced more A2 types in the first part in both cases than they did in the second part. This also concurs with the results from the analysis of the lecture transcript. However, the audio first group produced a higher percentage of CEFR level A2 types in each case compared to the respective parts for the video first group. At B1 level and above, however, the video first group consistently produced a higher percentage of types at
the different levels for the first part of their texts compared to the audio first group but the reverse is the case for the second part of the texts.

These percentages in all cases are higher than the corresponding percentages in the original lecture. However, this may be due to the fact that the test takers are picking up on the key points from the lecture in each case, which consequently make up a larger proportion of their texts overall.

This may also explain why it appears that group 3 performed better: the mean type count for group 1 (audio first group – part one: audio input) was 102.754 compared to 71.576 for group 3 (video first group - part one: video input). Therefore, a count of, say, 10 types in both groups would relate to a higher percentage in group 3 than it would in group 1. As a result, it appears that group 3 performed better on the first part of the text than group 1. While they did perform much better in the first half than in the second, it is not clear if actual performance concurs with the findings of Hotelling’s $T^2$ or not.

Assuming that group 3 did indeed perform better, this could be due to having access to the video input in the first part, which gave a definition of culture shock, which was not only quite complex linguistically but it was the one sentence on the slide for that part of the input. Therefore, it was only reproduced by the video first group and not the audio first group, which could account for some of the higher-level language they produced.

Another aspect of the CEFR type percentage analysis that should be highlighted is the high percentage of unlisted words on average in the texts for both part 2 groups (groups 2 and 4). This is striking as there is a much lower percentage of unlisted words in the original lecture transcript (11.41 per cent in part one compared to 6.29 per cent in part two). Going back to the data to investigate the reason for this, I found that this could be attributed to people who wrote nothing or very little in the second part of their mini-summaries so the code number
which was on each paper for identification purposes was the unknown word, which consequently comprised a large(r) proportion (or all) of the text.

For the token percentage, the pattern is similar in that both groups performed better in the first half than in the second except for the higher level of B1 and C1 tokens in the second part of the video first group (group 3). Again group 3 seem to produce more higher-level language across bands B1-C2 but this may be a result of the differing average token counts across the groups.

B1 percentages are high for all four groups with means ranging from $M = 14.283$ in group 4 to $M = 19.781$ in group 3. These values are much higher than in the original input in the lecture, where only 4.71 per cent of tokens were B1 level in the first half and only 5.06 per cent were B1 in the second half. However, the text length may account for this to some extent as the original lecture had a much higher token count overall (1108 tokens in part one and 1048 tokens in part two), so this amounts to 50 tokens in each case, but based on the average token counts for the four groups, this results in 29.25 tokens for group 1 on average; 18.59 tokens for group 2; 23.17 tokens for group 3 and 11.81 for group 4. As such, it is very likely that the test takers picked up on a good deal of the words at this level and reproduced them; however, due to their texts being considerably shorter, the overall percentage looks much higher. Therefore, differences in text length should be borne in mind when interpreting these figures.

However, there are other possible explanations which should also be considered. One is that the students used, relatively speaking, a large amount of vocabulary at this level because this is the level that best corresponds to their ability; another possible reason is due to differences in spoken and written language. Although a lecture can be considered planned speech (Luoma, 2004, p. 12), the lecturer is generally quite informal so this language might be less complex than an equivalent piece of academic writing. As a result, the written language of
the non-native speaker students in the test may, in some cases, be more formal than the spoken language of the native speaker lecturer.

Finally, the percentage of AWL types and tokens was analysed. These results reveal what at first glance appear to be different overall findings to those outlined above. While the mean values have always been higher for part one irrespective of input format according to the previous tables, these results show that, although the video first group performed better in terms of the percentage of AWL types and tokens produced on part one than part two (in accordance with previous findings), in this case the audio first group appeared to perform better on the second part in terms of the percentage of AWL types and tokens reproduced. The percentage values, being higher for the second part than the first, concur with the original input lecture, which had higher percentages in part two than in part one for both measures with 5.83 per cent (n=18) and 7.48 per cent (n=22) respectively for the AWL types and 4.98 per cent (n=59) and 8.2 per cent (n=79) respectively for the AWL tokens. However, when the average percentage of AWL types is divided by the mean word count in each case for the four groups, the findings are similar to other findings in as far as the average count of types and tokens is higher in part one for both the audio first (n=8.37 (types) and n=16.80 (tokens)) and the video first groups (n=8.52 (types) and n=15.27 (tokens)). Similarly, the audio first group reproduces more AWL types and tokens in part two (n=7.59 (types) and n=12.91 (tokens)) than the video first group (n=4.28 (types) and n=7.37 (tokens)). This demonstrates that, in line with the findings above, the audio first group maintains performance at almost the same rate between the first and second part but the video first group performs much better in part one than in part two.

The mean percentage values across the four groups range from 8.155 per cent (group one) to 11.916 per cent (group three) for the percentage of AWL types and 8.584 per cent (group one) to 13.027 per cent (group three). These mean figures are above the suggestion by Nation (2008, in O’Loughlin, 2013, p. 14) that ‘four percent or more of AWL tokens in responses to
a given item indicates that a test item elicits an adequate level of academic vocabulary for a test of academic proficiency. However, it should be borne in mind that these are mean values whereas the minimum values in all groups except group three is zero percent. The minimum for group three was 2.5 per cent for the AWL types and 3.49 per cent for AWL tokens. The maximum values, on the other hand, ranged from 20.83 per cent for AWL types (group four) to 40 per cent (group three) and 25 percent for AWL tokens (groups two and four) to 40 per cent (group four). The median values, which ranged from 7.27 per cent (group one) to 10.95 per cent (group two) for AWL types and 7.69 per cent (group one) to 11.36 per cent (group three) suggest that there are a lot of test takers below the mean. Due to the wide range across the test takers in the four groups, it is difficult to say how many test takers met the four per cent threshold. Nonetheless, these values provide some evidence of the test’s lexical validity as an EAP test.

6.3.2 RQ1: How does performance on an integrated listening-into-writing EAP task differ when test takers are presented with lecture input in the form of a podcast or a vodcast?

Returning to the RQ then, it can be concluded that test takers performed better when presented with lecture input in the form of a podcast when this is presented as the first of the two input formats. However, there is a statistically significant difference between performance in an EAP listening-into-writing task when the podcast is presented as the first input mode compared to the second. On the other hand, performance related to video input is very similar irrespective of the input order such that there was no statistically significant difference between the two. Both groups performed better in the first half than in the second half, irrespective of input format, which may be due to a possible fatigue effect in the later part of the task, which seems to be greater if the change in input goes from vodcast to podcast than when the first part is a podcast and the second is a vodcast. This could be because the audiovisual input is stimulating enough to maintain concentration whereas this is not the case for audio input alone. There is a small chance that the differences were due to the content of
the second half; however, since the content was analysed and found to be very similar across the two parts of the lecture, this seems less likely.

6.4 RQ2: To what extent are test takers influenced by the written word in an EAP lecture compared to the spoken word?

6.4.1 Qualitative analyses

6.4.1.1 Coder reliability

It is important to train coders and check the coding for consistency by asking another coder to code at least a sample of the data. Mackey and Gass (2005, p. 244) point out that there are ‘no clear guidelines ... as to what constitutes an acceptable level or interrater reliability’. However, they cite Portney and Watkins (1993) who define a percentage match rate of 75 per cent or over as ‘good’; however, ideally, 90 per cent or above should be aimed for. On this basis, it can be argued that the reliability, while not quite ‘good’, was acceptable. At 71.43 per cent, although the percentage did not quite reach the ‘good’ level on exact matches, it was comfortably within the ‘good’ range when the close matches (84.26 per cent) were also included.

The biggest issue with the coding scheme was the number of ‘similar matches’, which had quite a large impact on overall agreement. The vast majority of similar matches were situations where both coders had spotted the same content but interpreted the coding scheme slightly differently. The second coder often highlighted the word rather than the paraphrase while the first coder tended to highlight the paraphrase. For the consensus version, it was agreed that such a situation like this would be considered an exact word match.

Another aspect of this issue arose when there was overlap between what the speaker said and what was on the slide. For some of the key words, such as ‘Honeymoon Stage’, there was no alternative way of saying the word; consequently, the lecturer had to use that word despite the fact that it was on the screen. This led to an issue with the coding as it was both a word match from the spoken language and also from the written language. One coder tended to
mark such matches as matches from the slide if they fell within the video input while the other coder generally considered them exact word matches from the spoken input.

A further difficulty which arose was that one coder sometimes coded a word or a clause as a paraphrase from the slides while the other coder had considered this a paraphrase from the audio. For example: ‘how strong cultural differences are’ (4SA4) was coded as a paraphrased clause from the slide by one coder and a paraphrased clause from the audio by the other coder. The lecturer used the words ‘but if the two cultures are very different …’ while the words on the slide read ‘differences between the two cultures’. As such, in these cases, it was often very difficult to judge which of these was the true situation when the test takers had access to the video, so the consensus was that, unless it was a clear example of a paraphrase from the video, it would be counted as a paraphrase of the audio. An example of a clear paraphrase from the video was ‘mastering your language skills’ (KA2). This was coded as a paraphrase from the slide as the words on the slide were ‘linguistic mastery’ whereas the spoken words were ‘Then you need to learn the language’.

The remaining discrepancies arose when one coder identified a match that the other coder did not consider a match. This was possibly due to the fact that Coder A was more familiar with the content than Coder B while Coder B may have been interpreting the propositions more loosely and thus highlighting some points which Coder A felt did not represent a close enough match.

It was a shame that these issues did not come to light during the training phase. However, the fact that the coders could meet to agree on a consensus version for the problematic areas mitigated the effects of this to some extent, especially as the original coder went back through the papers to check for these coding discrepancies.

Although a less complex coding system which simply highlighted any match, irrespective of length (word, clause or sentence length), as a match, would have resulted in the higher
percentage match of 84.26 per cent, it was, nonetheless, still slightly disappointing, especially as coder training and sample coding had taken place with the two coders before the full set of data was coded. These findings do, however, highlight the need to find a coding scheme which is complex enough to provide the information required without being so complex that the coders cannot interpret it correctly (Stevenson, 2015, p. 329).

6.4.1.2 Qualitative Content Analysis: types of reproductions

6.4.1.2.1 Part one

As might have been expected from the results of the quantitative analyses performed for RQ1, the results for the number of reproductions across parts one and two highlighted the fact that the audio first group produced around twice as many matches as the video first group, irrespective of input format. These findings are unexpected as the video group also had access to the written cues on the PowerPoint slides in part one so it could reasonably be expected that they would have reproduced more matches for this part, yet despite having the additional support of the visual input, they still only reproduced approx. 53 per cent of the number of matches that the first group produced.

Again, reflecting the results from RQ1 above, the results revealed that both groups produced somewhat more matches in the first part than the second. One reason for this could be that there were more possible matches in part one than in part two. The other possibility as we have seen in the analysis for RQ1 above (see Section 6.3.1.2) is that the students’ attention may have dropped off in the second part as fatigue set in. Consequently, they may not have made many notes and thus not had as much to write about. While differences in text difficulty could also be a reason, the preparation in the design phase and the resulting textual analysis suggest that both parts are of a very similar difficulty level so this is unlikely to be the case.

These findings also appear to support other studies (Robin, 2007; Vandergrift, 2004) who found that lower level learners performed better with the help of the written texts but struggled when this support was removed as they did not have the ability to form the schema.
required for the audio-only input. Consequently, the lower level learners may have been helped by the video in the first part but may have struggled or been unable to cope in the audio part when they no longer had visual support.

Looking at the detailed analyses for part one, if the exact word matches from both the slides and the audio are added together, this results in a total number of word level matches of 327 matches for the video first group, which is still less than 2/3 of the word level exact matches reproduced by the audio group despite not having access to the video for support.

In part one, the video first group produced a similar number of SCLE matches as the audio first group produced ACLE matches. If we therefore add the ACLE matches produced by the video first group to the SCLE matches, the result is 93, which means that the video first group reproduced almost 50 per cent more clause level matches in total than the audio first group. This suggests that the video may, in fact, have provided some useful support at the clause level to the video first group.

6.4.1.2.2 Part two

For the second part, the number of audio matches is quite similar but, due to having access to the video input as well, the audio first group still produced somewhat more matches overall. However, the fact that only four students reproduced the one whole sentence could be a result of the sentence in part two being at the end whereas in part one, it was at the beginning. The students may have been engaged in writing content from the previous slide so, due to the speed of delivery, they may not have had enough time to write the sentence. This would be in line with findings from Ladas (1980, cited in Robinson et al., 1997, p. 260) who found that students could not take notes if lectures were delivered faster than 135 wpm. The effect of working memory may also have affected how much students could retain and write after the speaker had moved on. The so-called ‘phonological loop’, which is responsible for retaining and controlling the input of information (Baddeley and Hitch, 1974, cited in Brunfaut and
Révész, 2015, p. 146), and working memory, it has been suggested, may be vital for comprehension. This would explain why there were comparatively more word or clause length reproductions than sentence length ones across both input formats.

Another reason why fewer students wrote down the sentence could be a result of how important they perceived the sentence to be. If they did not consider it as important as some other aspects of the content, they may have made a conscious decision not to write it down. Finally, it could, of course also be evidence that even the audio first group was tiring at the end so the effort they had expended in listening to the lecture resulted in fatigue (McGarrigle et al., 2017, p. 95) leading to the completeness of the notes dropping (Locke, 1977, p. 93).

6.4.1.3 Qualitative Content Analysis: propositions

6.4.1.3.1 Part one

In part one, it is striking that the stages of culture shock were frequently reproduced by the audio first group, but this was much less the case for the video first group despite having access to a diagram showing this information. This may have been due to the explicit signposting in the spoken text which has been found to be useful for learners (Chaudron and Richards, 1986). On the other hand, after the most widely cited proposition, which both groups agreed on, the top three most widely reproduced propositions by the video first group: ‘lonely’, ‘feelings of anxiety and insecurity caused by new culture’ and ‘difficulty solving problems’ were not as popular with the audio first group; indeed, the first two of these were only produced by approximately half as many students in the audio first group. This is likely due to the fact that these were key points on the slide, too, albeit written differently. In fact, when the results of the reproductions from the slide are taken into account, it can be seen that the most widely reproduced proposition from the slide was ‘loneliness’ (n=9), (which is only three fewer than the total number of reproductions of this proposition by the video first group), followed by ‘[t]he anxiety that results from losing familiar signs and symbols of social intercourse’ (n=8), ‘depression’ (n=8), ‘insecurity’ (n=8) and ‘an inability to solve easy problems’ (n=8). These
findings suggest that the test takers were highly influenced by the slides at this point. This coincides with Locke’s (1977) findings that students omit far more of the information which is not written on the board compared to what is on the board.

With regard to the least frequent citations, the video first group had a diagram showing the 5 stages of culture shock depicted on the ‘U’ and ‘W’ curves, yet they did not reproduce propositions relating to the shape more frequently. This may mean that they are less influenced by a diagram so this would certainly be worth investigating in a future study.

The results for the individual reproductions seem to indicate, as one would hope, that the higher level students reproduce more propositions than the lower level students, with the higher level students reproducing a mixture of paraphrased and exact matches while the lower level students produced mainly exact matches but a much lower percentage thereof.

Moving on to the reproductions from the slides in part one, to which only the video first group had access, these findings demonstrate rather well the large occurrence of word level exact matches from the video input compared to the number of clause level matches. They also demonstrate the reproduction of the one full sentence, which was produced by 8 out of 15 students in the sample and 16 overall (as shown in Table 5.11 in Section 5.4.3.1 above). The three students (5SV20, SV15 and SV10) who reproduced 75 per cent, 50 per cent and 50 per cent respectively of the propositions from the slides, were clearly assisted by the presence of the written form as the first two of these students did not manage to reproduce any propositions in the second part when they no longer had access to the visual scaffolding.

While the third student did have some audio reproductions in both parts, this student produced more overall matches in part one than two thanks to the additional support of the slides. Nonetheless, this student still produced more audio-based matches (n=16) than video-based (n=10) matches in part one. They also produced 10 audio-based matches in the second part
so this student appears to be more independent than the other two students with regard to reliance on the written word.

Overall, the findings from the video input for part one suggest that the higher level learners rely much less on the written input while those in the middle tend to rely on it somewhat more. However, at the lower end of the scale, the students in this sample appeared to not use the slides either, although they did produce some exact and paraphrased matches (n=4, n=2 and n=1) respectively from the audio input.

6.4.1.3.2 Part two

For part two, it is clear that the audio first group produced far more propositions in total than the video first group. Here again, a large number of the reproductions were related to the content of the slides, for example, ‘be well prepared’, which was in the nominalised form on the slides: ‘preparation’ and, similarly, ‘be aware of culture shock’ which was ‘awareness’ on the slide. This suggests that the visual input may also have helped the audio group to some extent. Indeed, this would tie in with the results from RQ1, because, despite the possible onset of fatigue in the second half of the lecture when the audio first group saw the video input, this group nonetheless performed only slightly worse than the video first group did when they had access to the video input in the first half.

When we look at the reproductions from the video for the audio first group, we can see that, in contrast to the video first group, many of the reproductions come from the diagram, which in this case, is a graphic organiser. In contrast to the diagram, from which few students reproduced the content in part one, this may suggest that students pay more attention to graphic organisers. Alternatively, they may have just found this easier to read as the writing was bigger than on the diagram. Therefore, further research could investigate the impact of different types of visuals such as diagrams and graphic organisers.
With regard to the individual performances for the reproductions in part two, the test taker who was awarded the highest score in the audio first group only produced just over 35 per cent of the propositions while the third highest score was awarded to a test taker who reproduced nearly 60 per cent of the propositions. In the video first group, the top scorer produced over 55 per cent of the propositions while the next three had much lower percentages of the propositions. These results appear to demonstrate that, for the raters, there was more to getting a good score than merely regurgitating the content. Indeed, the top scorers demonstrated both ‘fluency’ and accuracy in their writing by adding in sentences and clauses which helped the text sound more natural and fluent, rather than being a stilted repetition of the key points from the lecture.

The video first group produced more video-based matches overall than the audio first group. This is striking as the video first group generally performed worse than the audio first group. However, this may have been a result of the input order such that the video first group took advantage of the visual input while the audio first group may have got used to taking notes from the audio input so may not have paid as much attention to the slides when they had access to them in the second part as the video first group did in the first part. Furthermore, if the video first group got used to relying on the slides and were then unable to process the aural input when that support was removed this might also explain why they performed worse, on the whole, in part two.

At the lower end of the scale, however, the test takers appear to merely combine words from the text into sentences which are very simplistic and, consequently, have a high proportion of matches from the text compared to the overall word count.
As can be seen from the graphs in Section 5.4.3.1.3, the results of this aspect of the analysis do not reveal any clear patterns. However, looking at the results in more detail, there are a number of points which can be made with respect to these results.

First of all, as we saw with the top test takers, there appears to be more to their essays than just mere reproduction of content from the lecture. Their content points are interspersed with linking phrases to make the writing more fluent but also the accuracy is generally very high. As a result, these test takers tend to have a high word count but relatively few propositions, thus resulting in a low ratio. However, this is not just the case for the top scorers; several other test takers had quite high word counts but also quite a high number of reproductions but this resulted in a ratio that was not terribly high due to the word count. For example, test taker 2KA9 had 157 reproductions in part one but as their word count for part one alone was 339, they had a ratio of 0.46. In part two, the same test taker had far fewer words (n=107) and also considerably fewer reproductions (n=62) thus resulting in a ratio of 0.58. The top scoring test taker in the video first group had a word count of 236 and 80 reproductions so, despite a reasonably high word count and a relatively large number of reproductions, their ratio was only 0.34 while test taker 5SV16 had only produced 2 propositions but in a 19 word text so this resulted in a ratio of 0.11. In contrast, test taker SV15 had a high ratio because they had reproduced a relatively high number of words (n=38) but they only had a low word count overall (n=46), resulting in a ratio of 0.83.

From this, we could conclude that this may not be a very worthwhile measure as the ratios do not give enough information about relative word counts and reproductions. Alternatively, we could conclude that it is necessary to do this analysis on the whole group to see if there are any more noticeable patterns across a bigger group of test takers.
6.4.2 RQ2: To what extent are test takers influenced by the written word in an EAP lecture compared to the spoken word?

The qualitative analysis showed that the majority of test takers, especially those around the A2 to B1 levels are influenced by the written word when it is available; however, the very low level learners (A0 and A1) do not seem to pay as much attention to the information on the PowerPoint slides as those at intermediate or above level, On the other hand, higher level learners seem to paraphrase more, so, while they reproduce the information which is on the slide, this is paraphrased so it is less clear whether it has come from the audio input or the slides. What appears to distinguish the higher-level learners, however, is how they combine the information from the lecture and other language such as linking phrases and introductions to make the text sound more natural and not just a list of reproduced facts. However, it is not entirely clear from this limited qualitative study what the patterns are in terms of the number of words reproduced as a ratio of the total number of words as the results of this part were very mixed, with high level learners producing both high and low ratios and, similarly, lower level learners producing both high and low ratios. A more in-depth look at this is required to establish if there are any general patterns across a larger number of test takers.

To summarise then, we can say that for lower level learners, it appears that the video input helps them if they can cope with the cognitive overload but this does not seem to be the case for the very low-level learners (A0 or A1). Nonetheless, the lower and intermediate level learners (A2 and B1 level) do seem to benefit a lot from the additional support provided by the video input as demonstrated by the high number of matches to the video in both groups. The higher-level learners do also benefit but these learners did not appear to pay as much attention to the video as the intermediate level learners. This was more evident in the audio first group so this could well have been a result of getting used to having only the audio and thus not ‘needing’ to look at the video.
6.5 RQ3: How do test takers’ perceptions of their performance on an integrated EAP listening-into-writing test presented in two different input formats compare to their observed scores as awarded by trained raters?

6.5.1 Quantitative analyses

The results for this question for the audio first group show that while nearly 50 per cent of the students perceived their best and worst performances correctly, a large proportion of students thought they performed best on part two whereas only seven actually did. This might suggest that students are not very adept at assessing their own performance.

However, since a large number of students actually performed equally well on both parts, these figures may be misleading. The issue here was that it had not been anticipated when preparing the questionnaire that students would score equally well on both parts, something which, in hindsight, was an oversight on my part. This is likely due to the fact that, in the exploratory study, test takers were asked which part they felt they performed best and worst on but they had four input formats to choose from. In addition, the EAP texts were only given one overall mark so it was not the scores that were used for the correlation between test takers’ perceptions and their performance but rather the word count so the issue of equal performance across different input formats did not arise. Consequently, when the questionnaire was amended from choosing the best and worst performances from four input media to only two for the main study, this was not noticed so there was no option for ‘I performed equally well on both parts of the writing task’ on the questionnaire. Therefore, further investigation was necessary to clarify this RQ.

I thus decided to add the word counts to the analysis as this might give the additional information necessary to answer the question and this was how the students’ perceptions were compared to observed performance in the exploratory study. This made quite a difference as all except two of the students in each group could then be allocated to an observed best and observed worst performance column.
For the audio first group, this increased the difference between perceived performance and observed performance because the perceived best for part one was 23 whereas the observed best in terms of word count was 48. Since 30 students considered part two to be their best performance, this was clearly an error of judgement since only 8 people performed best on part two, even taking account of the results of the word count analysis.

Nonetheless, it was still not clear from these results if the students who had said they performed best on part one or part two respectively were indeed the ones that did perform best on a given section. When the person-by-person matches were analysed, there were still over two-thirds of the group who were unable to self-assess their best performance compared to performance as measured by their scores and nearly half who were unable to self-assess their worst performance compared to performance as measured by the word count. While the figures were slightly better for the number who correctly perceived their worst performances compared to their performance as measured by the scores and word count, these were still only just over two-thirds and nearly half respectively.

For the video first group, the same issue occurred. There were 25 per cent more test takers who thought they performed better on part one compared to those that actually did while the number of test takers who perceived their performance to be best on part two was almost twice as high as the reality. Logically, the reverse was the case for the worst performances.

However, again, the large number of test takers who scored equally well on both parts may have skewed the results somewhat so the word count was also analysed and this resulted in a much closer link between students’ perceptions and their performance. The results show that there were 42 perceived best performances in part one compared to 43 best performances based on word count and 11 perceived best in part two compared to 13 observed best performances based on word count.
Nonetheless, to answer the RQ fully, it was important to do the person-by-person analysis, too. These findings revealed that two-fifths of the test takers correctly perceived their best performances compared to their scores while nearly two-thirds of test takers correctly perceived their best performance when compared to the word count. This meant, however, that over 50 per cent were unable to predict their performance as measured by the score while nearly a quarter were unable to predict best performance measured by the word count. The results for the worst performance were very similar except that two additional test takers did not respond to this question. It is not clear of course why a number of test takers did not respond. This could have been because they were not sure which parts they performed best or worst on or it could be that they felt they performed equally well on both parts but, as there was not an option for this, they did not select anything.

In summary, the results show that between half and two-thirds of the test takers were unable to judge their best performance compared to their score and between one-quarter and nearly half were unable to judge their worst performances. The lack of ability to perceive their performance is in line with findings from the exploratory study (see Section 3.4.4) and raises questions about self-assessment but this is beyond the scope of this study.

However, these findings do suggest that students seem to equate performance more with how many words they write rather than the content or language.

6.5.2 Qualitative analyses
The qualitative analyses show that there is no consensus regarding the preferred input format so, although students were not necessarily able to predict which part they performed better on, they had clear ideas about why they felt they had performed better or worse on a given section. 33 per cent of respondents said that the text / visual helped them to perform best while 14 per cent said that watching and listening is difficult and / or they were distracted by the video so they performed worse on the video section. On the other hand, 15 per cent said
listening only was easier with 21 per cent saying they performed worse because listening only is more difficult.

With regard to the other findings, the clear presentation structure was mentioned as a beneficial aspect of the presentation.

The comments about the speech rate were all from participants in the video first group who had audio only as the second part. Although there was a small difference in the speech rate between the two parts (whereby part one was 177.75 wpm and part two was 182.26 wpm), this could be because the audio only may have felt much faster once the visual scaffolding had been removed.

6.5.3 RQ3: How do test takers’ perceptions of their performance on an integrated EAP listening-into-writing test presented in two different input formats compare to their observed scores as awarded by trained raters?

To summarise, the results from this analysis were not very revealing when test takers’ perceptions were compared to their scores. This is possibly due to the high number of participants who scored equally well on both parts, so test takers’ perceptions of their performance were compared with the word count for each of the two-mini summaries and a person-by-person analysis was carried out to investigate whether the people who perceived their best performance as a given input mode did actually perform best on this part as measured by the score and / or the word count. This latter analysis revealed much more agreement between the test takers’ perceptions and the word count than between the test taker’s perceptions and the score in both groups.

In terms of the qualitative analysis, test takers had very clear reasons why they thought they had performed well or badly on a given part but there was no overall consensus between the participants as to whether the video helped or hindered performance as there was a large proportion of test takers who felt that they were helped by the video and, similarly, a large percentage who felt that the video was a distraction or was more difficult than listening only.
In short, on the whole, test takers’ perceptions of their performance are more often out of line with their performance as measured by their score than they are in line with observed performance. However, there is some evidence to suggest that test takers’ perceptions of their performance are more closely linked to the number of words they produce than the overall quality of the work.

In terms of the input format that they felt they performed better on, while there are clear individual preferences, there is no overall consensus as to whether it is better from the students’ perspective to have the input in audio or video format.

6.6 Post-task questionnaire results
The post-task questionnaire was designed to collect participants’ opinions on their experiences of taking the test in order to collect validation evidence to support the use of the test and to highlight areas for further improvement.

6.6.1 Quantitative results
While the results for Q1 suggest that 44.5 per cent of test takers felt they performed just as well as they would in any other listening test and 21.6 per cent felt they performed better or much better, 25.9 per cent felt they did worse or much worse. There could be a number of reasons for this. This could simply be due to their proficiency level, that is, it could be the weaker students who thought this, which would be reasonable, given that 28.5 per cent of the test takers were A2 level or below. Alternatively, the test takers may have struggled with the accent of the lecturer, which was northern English as compared to the Russian and Ukrainian non-native speaker accents that they were familiar with. Another possible reason is that the speech rate, which was slightly faster than the researcher might have wanted as the speed of delivery of the two parts (177.75 wpm for the first part and 182.26 wpm for the second part) was marginally above the students’ listening rates of around 150-175 and the second part was also slightly faster than average lecture speeds. Another option is that the test takers may have struggled with the note-taking task so this may have affected their
performance if they were not able to judge when they should take notes and when they should listen or if the combination of having to listen, process the input and reformulate the input to make the notes was too challenging (Lin, 2006, p. 2). However, this is just speculation and more research would be necessary to confirm these possible justifications.

The findings from Q2 suggest that the test was aimed at approximately the right level, that is, around the border between B1 and B2 since the majority of students found it quite difficult while the stronger students (B2+ and above) found it easy or very easy. Conversely, the lower level students (A1 and A2) found it very difficult. While it is recognised that no formal standard setting procedures have been done to link the test to the CEFR, these results do provide some level of concurrent validity (Weir, 2005) as they relate quite well to the (external, that is the university) language level groups that the participants are in: 31.3 per cent are B2 and C1 (so they would be the test takers who would be expected to find the test easy or very easy); 25 per cent are B2 and 13.8 per cent are A2 (so these would be expected to be the test takers who would find it quite difficult) and 14.7 per cent are A1 (so for these students it would be expected to be very difficult).

The results for Q3 demonstrate that very few people were very familiar with the topic, which had been the intention in choosing a general subject that would be interesting but that hopefully not too many people would know too much about. Those who did have background knowledge tended to bring that knowledge in to the summary and were awarded lower scores as the content they brought in was not what they were supposed to be summarising.

Q4 provides evidence that test takers felt that the topic had a beneficial impact on their performance.

Q5, Q6, Q7 and Q8 relate to RQ3 above so these will not be addressed again here.

Q9 investigates the factors that contributed positively to the test takers’ performances and reveals compelling evidence that the video did in fact have a positive impact on the vast...
majority of test takers (over 90 per cent) in some way. Furthermore, these results provide more evidence to support the fact that test takers were, in fact, positively influenced by the video, with nearly half of the respondents (44.8 per cent) finding the textual information helpful. The results of question 10 are rather disappointing as a special effort had been made to improve the quality of the recording. However, the researcher had no influence over the equipment used to play the recording and this could, unfortunately, have led to construct-irrelevant variance (Messick, 1989).

The amount of information the test takers had to write down was also considered to have had a negative impact on their performance. This is presumably because they had to write more than they wanted to, which could have led to the presumed fatigue effect that we saw in the results of RQ1. This would concur with Locke (1977) who found the completeness of the students’ notes decreasing during the course of the lecture. The test takers also said that the number of times they saw the video was negative but it is not clear from this question whether they would have liked to have watched it three times or just once. However, there was an open-ended question for students to give more details about the factors which they felt affected their performance and three students confirmed that they wanted to listen again. Four others said that there was too little information which we must take to mean that they had too little information to perform well in the writing test.

Q11 is a very similar question to Q2 but this is overall including the listening part rather than just the writing. As such, the figures are a little more towards the difficult end than for Q2, but this question again confirms that the test is at more or less the correct level for the target population. This provides concurrent validity evidence but again more work would need to be done if such a test was going to be operationalised as the cut scores would have to be set using standard-setting procedures.
6.6.1.1 Summary of quantitative results
The results of the questionnaire were mainly quite pleasing, especially the fact that many students felt they did as well or better than they usually would on the listening test. Although a large proportion of test takers found the test quite difficult, there was also a large proportion who found it quite easy, which also suggests that it seems to be pitched at approximately the correct level (although I recognise that there would have to be a formal standard setting and CEFR linking process to justifiably claim this). Furthermore, the fact that students felt that the topic was beneficial was also satisfying. However, these findings must be treated with caution since the findings for RQ3 suggest that students are not necessarily as good at judging their performance as they think they are. On the other hand, it was very disappointing that a large number of test takers complained about the quality of the video, especially as extra care had been taken to do a better version than was used in the exploratory study.

To summarise, the quantitative results suggest that the test appears to be targeting approximately the correct level. Furthermore, the performances seem to reflect the test takers’ level as stated in the feedback questionnaire. However, before such a test were operationalised, the recordings would need to be redone using proper equipment to ensure that the video and audio were of high quality to minimise construct-irrelevant variance affecting the test takers’ scores. Similarly, more work would need to be done to set cut scores before using such a test but this is beyond the scope of this project.

6.6.2 Qualitative results
The qualitative results show that over 12 per cent of the respondents thought the topic and the experience had a positive impact on their performance. While investigating the effect of topic knowledge on performance per se is beyond the remit of the study, it is encouraging to see that such a large percentage of test takers specifically considered that it had had a positive impact on their performance.
Unfortunately, nine students stated that they ran out of time to complete the task which will most probably have affected their scores on part two and therefore will have had some effect on the results overall. Although the length was reduced to accommodate the limited time available for the testing, this still seems to have been an issue for some. However, looking at the test takers who made these comments, some of them were the stronger students who wrote far more than the 350 words required while others were at the lower end of the ability range (A2). Nonetheless, this is an issue as this will likely have had an impact on their performances in part two, which could explain why people seemed to perform worse on this part. Having said that, although it is an issue which must be taken into account, there were only nine comments in total related to the timing so it must also be borne in mind that it was a relatively small percentage of test takers overall who felt this was an issue. On the other hand, it could be a useful lesson in time-management during tests for the test takers.

The issue of speededness in essays is a tricky one. Students need to be able to produce long stretches of writing within a reasonably short space of time, for example, during an exam, yet many, particularly post-graduates quite often have non-speeded writing tasks, such as take-home assignments, which may lead to better results once the speeded element is removed. Therefore, this needs to be considered and, for a future iteration of the test, the time could be extended to ensure that all test takers have a sufficient amount of time, irrespective of how much they write.

**6.6.2.1 Summary of qualitative results**

There were a large number of positive comments related to the topic and the usefulness of the visual support, the largest number of negative results related to the lack of time available while other negative aspects related to aspects such as speech rate, the test takers’ ability in English and the difficulty of the topic.

Nonetheless, half of the test takers felt that nothing else had had any effect on their performance.
6.6.3 Summary of feedback questionnaire results

To summarise, the feedback questionnaire was designed to provide validation evidence for the test and to highlight areas for improvement. Based on the feedback from test takers, the test appears to demonstrate concurrent validity and appears to be targeting the correct level. Test takers found the topic interesting and the visual support useful which lends support to the use of video as opposed to audio only for integrated tests of listening-into-writing.

However, the issues highlighted include the quality of the video and sound as well as the time allowed for the test. Speech rate was also highlighted by some test takers as being an issue so this could be slowed down a little in a future recording.

Once these issues are addressed, more work would need to be done to set cut scores but, on the whole, the feedback questionnaire suggests that a video-based listening-into-writing test is a promising task type for an integrated test of listening-into-writing.

6.7 Conclusion

In this section, I will draw together the findings from my study and those of other studies in the field and will consider how my study contributes to the field of research.

In terms of RQ1, this study corroborates the findings of several other studies. In line with Leveridge and Yang (2013, p. 202), these findings suggest lower-level learners will be helped by the text if the audio is too fast or too difficult. Similarly, the results of the main study corroborate those of Robin (2007) and Vandergrift (2004, p. 10) in that students in the video first group struggled with the listening comprehension when the visual support was removed after they had had access to it, resulting in a statistically significant difference in performance between the two groups when presented with input in the form of a podcast, yet this was not the case if the video input came after the audio input. Nonetheless, the performance of the audio first group also dropped off in the second part, thus suggesting that there may be a fatigue effect in line with Locke (1977) or an issue with the amount of time that participants had to complete the written task, which is consistent with Read’s (1990) findings.
The results of RQ1 add weight to other studies in the field by confirming the drop in performance when visual support is removed and the longer the lecture, the higher the chances of fatigue affecting performance. In an integrated test, this could manifest itself as fatigue from the input or from the output. If the former, this results in students not having enough to write in the output. If the latter, this means they may have understood the input but may have been too tired to write as much as they could. In both cases, fatigue can have a negative impact on their performance. Alternatively, they may run out of time, either because they were tired or, conversely, because they had too much to write in the time available, which will also have a negative impact on overall performance.

With regard to RQ2, it was found that the majority of the lower intermediate to intermediate level students, that is, those around A2-B1 level made use of the video. These findings are consistent with those of Cumming et al. (2005a). They also reproduced exact matches from the video along with paraphrased clause level matches from the audio. The very low level learners, on the other hand, did not appear to focus as much on the content of the video, perhaps due to the cognitive overload that the combination of reading, writing and listening results in. This is consistent with Piolat, Olive and Kellogg (2005). On the other hand, the results demonstrate that the higher-level learners rely much less on the video input and paraphrase much more. While Gruba (1994, p. 34) and Brett (1997, p. 48) both found that the students in their study did not watch some or all of the video, the current study did not include eye-tracking or videoing of the participants so this study cannot confirm whether the students did indeed pay attention to the video or not. In addition, the more proficient test takers add to the text by including sentences to connect the ideas and make the text flow more naturally, thus sometimes resulting in a better mark despite having fewer reproductions from the text. This could be attributed to Spivey and King’s (1989) constructivist perspective whereby writers employ the three discourse synthesis processes of organising, selecting and connecting (p. 7) and the higher-scoring test takers included ‘more local coherence’ by
establishing ‘clearer connections for the reader’ and ‘superior global coherence’ resulting in a ‘tighter text organisation’ (p. 21).

However, there were no clear patterns regarding the Reproductions index because some of the high performing students had high percentages of reproductions compared to the word count as they had reproduced large numbers of propositions from the input text while others had low percentages due to the high amount of paraphrasing and additional linking structures. At the other end of the scale, there were high percentage matches due to a high number of reproductions compared to a low number of words and, similarly, low numbers of reproductions compared to low word counts. As such, given that the qualitative analysis was only done for a small sample of 30 of the test takers, there were no discernible patterns.

With regard to RQ2 then, this study provides more evidence of the different ways in which learners at different levels interact with the input media. The study provides more evidence of the cognitive overload experienced by lower level learners, while it demonstrates that higher level learners do not rely as heavily on visual support as they have the listening skills to cope without it. This study also demonstrates that the percentage of reproductions does not reveal any patterns linked to proficiency level as measured by the scores awarded by the raters due to varying numbers of reproductions and variation in the way the propositions are reproduced. However, in line with Spivey and King (1989), the raters’ scores suggest that producing a successful summary goes beyond merely reproducing the input to involving more organisational features in the writing.

For RQ3, the researcher wanted to investigate how the test takers’ perceptions of their performance on the integrated listening-into-writing test compared to their observed scores. For this, the test takers answered a follow-up questionnaire and found that, as with the exploratory study, the test takers did not perceive their best and worst performances in the same way as the raters did, yet there was more agreement when the test takers’ perceptions were compared to the word count for each of the mini-summaries. This suggests that the test
takers may consider the amount they could write as a determiner of success while raters take other aspects into account such as the language, cohesion and / or the number of propositions reproduced. Nonetheless, the actual level of agreement between the word counts and the test takers revealed that only between 20 per cent and 40 per cent of the students correctly perceived their best and worst performances.

The second part of the analysis for RQ3 looked at test takers’ comments regarding why they perceived a given part (one or two) to be their best or worst performance. Over 40 per cent of the respondents felt that they had performed better on the video part because they were helped by the visual input while 15 per cent of respondents thought that listening only was easier. 14 per cent felt that the video had hindered them or that this led to cognitive overload so they did not pay attention to the it. The questionnaire findings lend weight to the findings in RQ2 that many of the test takers paid attention to the video when it was available but some did not need it and others did not want to use it.

Previous research looking at test taker preferences regarding input format has produced conflicting results. The preference for video input is in line with other studies (Aldera, 2015; Brett, 1997; Pardo-Ballester, 2016; Suvorov, 2008; Wagner, 2010b) while the comments in favour of audio only input are in line with Coniam’s (2001) and Gruba’s (1994) findings. The comments in support of the use of textual input coincide with findings from Chang, Lei and Tseng (2011) and Montero Perez, Peter and Desmet (2014).

RQ3, along with the findings from previous studies, demonstrates that the question of which input format is best is a complex one. Students express different preferences for the input media, which may or may not be in line with their observed performances and their perceptions of how they perform. As such, there is still no definitive answer to this question.

The last set of results was the questionnaire. These results highlighted a number of issues. Participants in the video first group commented on the speech rate. Although it was only
slightly faster in the second half when they only had audio input, they may have perceived it to be much faster due to the lack of visual support. In line with Chang, Lei and Tseng (2011), the participants may not have built up the schema to process the audio input as they were relying on the scaffolding in the first part. Alternatively, this could have been because the speed was just slightly faster than the 100-180 wpm that most lectures are delivered at (Carver, 1982 in Robinson et al., 1997, p. 260. Another possibility is that this was due to the lack of familiarity with the speaker’s accent, which corresponds to Ockey, Papageorgiou and French’s (2016) findings. This is an important factor to consider going forward as the impact of not understanding the input and its effect on performance on the output is one of the concerns with integrated tests (Cumming et al., 2005b).

Another issue that was mentioned was the quality of the recording, which also needs to be addressed to avoid construct-irrelevant variance (Messick, 1989).

Finally, the time available was an issue for some test takers so thought needs to be given to either extending the time available or making the test non-speeded. and the latter may better reflect the test takers’ real abilities (Read, 1990, p. 118-119) but may not be practical for a university entrance exam.

On the other hand, the questionnaire results also highlighted the positive aspects of the test. In general, test takers were happy with the topic despite not being familiar with it, which had been the intention to avoid the advantage of considerable background knowledge in accordance with Clapham (1996). In fact, some of those who did possess background knowledge brought this into their summaries, which was not advantageous as the content they included had not been in the input text. These findings support Long’s (1990) study, which also found that, while background knowledge generally helped students to perform better in a post-test summary task, for a small group of participants, background knowledge led to a negative impact on their performance. On the other hand, it is important that test
takers do not feel that they are being disadvantaged by a topic (Jennings et al., 1999, p. 426) as this may result in construct-irrelevant variance (Messick, 1989).

Comments that the clear presentation structure was helpful concur with findings related to discourse markers in academic lectures (Chaudron and Richards, 1986; Rickards et al., 1997; Sanders and Noordman, 2000).

Finally, test takers were also satisfied with their performance on these tests compared to other tests, on the whole, although some found the essay quite or, indeed, very difficult. Given that some of the test takers were somewhat lower than the target audience, this is to be expected. Similarly, the test takers at the higher level found the test easy or very easy. Again, this is to be expected. The fact that the lower level test takers struggled while the higher level ones found it easy, along with the fact that the majority of test takers fell in the middle of these two extremes with some finding it quite easy and others finding it quite difficult, suggests that the test was targeting the correct level and that the test takers who took the test would reflect the target population in terms of level.

All of these points highlight areas to work on but also provide evidence to support the validity of the test. This is important because the testing instrument must be valid in order to be able to draw conclusions from the findings. If a test does not demonstrate the qualities of test usefulness (Bachman and Palmer, 1996) or does not have evidence of different types of validity (O’Sullivan and Weir, 2011; Weir, 2005), the conclusions we draw on the basis of the results may be incorrect. In this study, establishing the validity of the interpretation of the test scores is fundamental as it would not be possible to provide clear answers to the RQs if the test did not give valid results.
7. Limitations of the study and recommendations for future research

7.1 Introduction
This chapter will outline the limitations of the study and, on the basis of these limitations and the findings from the study, will suggest a number of areas for future research. Finally, it will discuss the implications of this study and the conclusions to be drawn from the project.

7.2 Limitations
One of the limitations was the sample used for the main study. The participants comprised an opportunity sample from each of the two universities. It would have been better to have had an entire population. Unfortunately, this was not possible as, at the time of data collection, some students were already doing their exams and participation in the study would have put additional pressure on them. Furthermore, there was a problem with the data collection at the Russian university and some students were allocated to the wrong group resulting in there being more video first group students there so, due to the timing of the data collection, this had to be made up with additional participants from the Ukrainian university. However, although this was not initially planned like this, it did, in fact, have the benefit of evening out the number of participants a bit more.

A second consideration is that, because there was only one input text, it was not really possible for students to demonstrate their ability to synthesise information from other sources, which is a feature of Overall Written Production at B2 level according to the Common European Framework of Reference for Languages (Council of Europe, 2011, p. 23). Having more texts would allow test takers to better demonstrate their abilities at the higher end of the CEFR scales as the combination of the task and the rating scales made it difficult for test takers to achieve the higher marks.

The next aspect to consider is the time allowed for the essay writing task. Participants were limited to 45 minutes for the task. However, as scores were lower in the second half of the
essays for both groups, it is not entirely clear whether this was due, partly or mainly, to a fatigue effect - during the listening or the writing – or whether the participants ran out of time, both of which could have skewed the results.

Another point to note is that, although the raters were comfortable with the CEFR and the rating criteria, as evidenced by the high rater reliability, the rating criteria were not designed specifically for the task as this would have been beyond the scope of the project. Instead, the CEFR Overall Written Production scales were used. However, this scale was quite limiting. While it might have been beneficial to have used the ‘Reports and Essays’ scale as well as the ‘Processing Text’ scales, this was unfortunately not practical due to the additional costs that would have been involved in rating the papers multiple times. It has been pointed out that the rating criteria must be clearly defined to take account of the constructs being tested (Chan, Inoue and Taylor, 2015; Yu, 2013b, p. 111) so it is a shame that the Companion Volume of the CEFR (Council of Europe, 2018) had not been published at the time the rating was carried out as the ‘Relaying specific information in writing’ and / or the ‘Processing text in writing’ from the new Mediation scales may have proven to be more appropriate.

Finally, with respect to RQ3, the initial findings were somewhat misleading as they suggested that many test takers were not good at assessing which part they performed best and worst on. This was an issue with the questionnaire in so far as there was not an option for ‘I think I did equally well in both parts’ so students had to make a choice while, in actual fact, a large proportion of both the audio first group and the video first group (n=25 and n=21 respectively) received the same Fair Average scores for both halves. It would have been better to include the option for students to say they felt they performed equally well on both parts. For this reason, it was decided to analyse the test takers’ perceptions compared to the word count and, although this was not the original intention, it did highlight the fact that test takers appeared to judge their performance on the basis of the amount they could write more than the quality of the writing.
7.3 Recommendations for future research

The section above highlighted the limitations of the study, some of which resulted in an unexpected positive contribution to the study while others raise issues which can feed into the recommendations for future research.

First, regarding the exploratory study, both noun phrase density and verb phrase density were slightly higher for the IELTS essay than the EAP essay. This could be due to the preparation that the students had had for IELTS essays and the fact that they had been taught to use nominalisations in such essays but simply forgot about the ‘rules’ of academic writing in the EAP task as the task requirements resulted in their focussing on the different types of processing required for the EAP essay. While this would potentially account for the lack of any statistically significant correlations, this is only speculation and merits further investigation so a further study could investigate the reasons for the differing performances between task characteristics and output on these two types of tasks.

In this study, the effect of the input medium and the learners’ source text use were analysed as two separate issues so as not to ‘cross-contaminate’ the data. However, having now analysed these aspects as two individual questions, a future study could look at how the type and frequency of source text use from different input media impacts on the scores. While this would have been a valuable aspect to consider in the main study, it was beyond the scope of the project.

Next, although the main study used a counterbalanced design, it would have been good to have two other groups: one that had both parts as audio only and one that had both parts as video only to investigate whether the audio second was still as difficult without any change of format or because of a fatigue effect. Similarly, it would enable investigation of the difficulty of the video formats when compared with each other.

Another idea for a future study would be to carry out the same study but with two lectures so that each group had audio first – video second and video-first – audio second. In this case, it
would be wise to not carry out both tests on the same day as there may be an even bigger fatigue effect than the one that was possibly a reason for the differences in scores in the main study presented here. This could be combined with the previous suggestion for future research by splitting the groups such that in addition to two groups who each get the two lectures in both formats, some test takers could watch both lectures in video format and others would listen to both in the audio format so that there would be no disjoint subsets and the whole analysis could be carried out using Multi-faceted Rasch Measurement. Furthermore, using more than one input text would enable the test taker to demonstrate their ability to ‘synthesize[e] and evaluat[e] information and arguments from a number of sources’ (Council of Europe, 2011, p. 23), as stated in the Overall Written Production descriptor at B2 level.

Another recommendation for future research is to carry out the research using test takers from different language and academic backgrounds. This was the original intention but, to eliminate language backgrounds as a factor which could have impacted on the study, the decision was made to keep to similar language and academic backgrounds for the main study.

The use of other types of visual input could also be investigated, as in the exploratory study. The main study focused on audio only versus audiovisual, which includes kinaesthetic, diagrammatic and textual support but a bigger study with a fully counterbalanced design to investigate the types of audiovisual assistance that are most beneficial would be of value. The findings from the qualitative investigation (RQ2) revealed that the video first group did not pay much attention to the content of the diagram in terms of propositions they reproduced whereas the audio first group reproduced far more propositions from their diagram, which was a visual organiser, so a similar study in which one group of test takers has textual input, another has diagrammatic input only and the third has graphic organisers would be useful to see how these input formats affect test takers differently.

On the other hand, a future project could look at using this study as a basis for developing and validating a task-specific rating scale for summarising information from a lecture.
Alternatively, this study could be replicated using the new mediation scales from the CEFR Companion Volume, which had not been published at the time the essays were being rated. It would be particularly worthwhile to see if the ‘Relaying specific information in writing’ and the ‘Processing text in writing’ scales are more appropriate for this task.

As there was a potential issue with the speededness of the writing task, a future study could consider either allowing longer to produce the essay or removing the speededness aspect completely so that participants are free to write as much as they wish in both halves, which may then produce different results. This would be particularly valuable for post-graduate students, who are generally not required to produce texts under time-constrained conditions. In line with Read (1990), having a speeded aspect to such a test may not be an authentic task, and could consequently result in scores which do not reflect the test takers’ level of ability, as they may be proficient but slow writers.

With regard to the Text Inspector findings, these analyses highlighted the differences between measures of lexical diversity, which are often considered a measure of how complex a text is. While type-token ratio can be calculated on texts of any length, it is particularly sensitive to text length (McCarthy and Jarvis, 2010, p. 381). On the other hand, when calculating group means using MTLD, care must be taken to highlight the fact that only texts over 100 words in length are used in the calculation. Bearing this in mind, however, MTLD has been found to be an effective tool for measuring lexical diversity (McCarthy and Jarvis, 2010, p. 387). Nonetheless, as suggested by McCarthy and Jarvis, 2010, p. 390) more research could be done to evaluate whether it is found to be consistently stable across different registers.

Another aspect of the Text Inspector analyses is that the descriptive statistics were only run on the analyses outlined in the Methodology chapter above. As a result, it was not clear whether the audio first group actually performed better on the CEFR measures than the video first group because only percentages were analysed and not the actual type or token counts.
in each case. Therefore, it would be good to run descriptive statistics on the type and token counts as well as the percentages.

7.4 Implications

This study has implications not only for testing but also for teaching. With regard to testing, perhaps the most important implication is the issue around the construct that is being tested in an integrated listening-into-writing task when video is (one of) the input format(s). This study has highlighted how the number and types of reproductions vary depending on whether the test taker has access to the video or not. If they do have access to textual information in a video, the focus tends to shift from the listening to the reading mode, which changes the construct being tested to reading / copying instead of listening.

On the other hand, regarding teaching, educators can consider the results of this study when deciding how to make content available for online study or in a flipped classroom or blended learning environment. They should consider that video may provide scaffolding to learners, especially at the lower levels, but at the same time, they need to be aware of the potential negative impact this may have on listening skills, so it may be wise to not have video exclusively, but to include audio formats from time to time, too. At the same time, they should be aware of the potential for a fatigue effect which, in this study, was found to be much higher when the audio was the second input format. This may have been exacerbated by the change of format but it may also have been a result of there being no visual scaffolding available to both maintain interest and support understanding. This is of particular relevance as schools, colleges and universities are forced to move to online teaching due to the spread of Covid-19, which may impact on the way teaching is delivered in the longer-term.

Another implication of the study is the use of type-token ratio as a measure of complexity. While this may be widely used to measure the complexity of a text along with the Flesch Reading Ease measure, I would advise that these be used with caution as the Text Inspector findings highlighted the differences between the Flesch Reading Ease, the type-token ratio
and MTLD. Bearing this in mind, MTLD might be useful for teachers and testers who want to assess the complexity of a text for their students or as a test task. Similarly, if measuring the complexity of student writing, care must be taken to ensure that the texts are all longer than 100 words.

7.5 Conclusions

Based on these findings, it would appear that video input could be the most beneficial form of input, especially since this provides visual input for those who want it but also means that those students who cannot cope with the cognitive overload that comes as a result of having to listen, watch, read and write can just listen without looking at the screen. Furthermore, the findings suggest that there is much less difference in performance across the two parts of the lecture when video input is provided irrespective of the input order. In contrast, there was a significant difference in performance across the two groups depending on whether the audio input came first or second. Furthermore, in terms of students’ preferences, it seems that providing audio input only is likely to disadvantage more students or test takers since fewer test takers appeared to prefer the audio only input. Conversely, those who expressed a preference for video input may struggle to concentrate or understand if the input does not provide visual scaffolding.

Taking into account the results of the questionnaire, which revealed that test takers were not all in agreement about which format is best, it may be wise to offer both options to students, and / or to vary input methods and to then investigate whether any pattern emerges regarding longer-term performance based on one or other of the input formats.

However, in a testing situation, it is important to consider the construct being tested. While the use of video may make for a more authentic test, this may affect the construct validity if visual input is also provided (Coniam, 2001) or, at the very least, one may need to reassess the definition of the construct (Ockey, 2007). If the aim is to test listening comprehension, then having textual input has an impact on the construct as shown by the large number of
words and phrases which were copied verbatim from the slides, changing the construct from testing listening skills to testing listening and (reading or) copying skills. If, on the other hand, the construct is to make an authentic assessment that reflects a TLU domain in which test takers are likely to have access to textual and diagrammatic PowerPoint input as well as nonverbal communication, such as body language or gestures, then the use of video may make the test more valid; however, further investigation is required to uncover the extent to which different types of visual input can support or potentially hinder (Leveridge and Yang, 2013) understanding of the input. To conclude, this study has attempted to address the question of the impact of input task characteristics, that is, input medium, on performance on an integrated EAP test. It has attempted to fill a gap in the research by bringing together the research on input formats in listening comprehension with the research that has been done regarding source text use in reading-into-writing research. Despite the study having a number of limitations, it has highlighted a number of areas for further research to improve our understanding of the listening-into-writing construct but, at the same time has found some answers to questions which had previously not been addressed to the best of the author’s knowledge.
Appendices
Appendix 1: Exploratory study - Information sheet and consent form for participants

Introduction to the research

This doctoral research will investigate the relationship between input task characteristics (textual, visual, kinaesthetic and auditory input) of listening comprehension texts, and performance on written output tasks.

The objectives are:

to investigate the relationship between performance on integrated EAP written tasks when varying types of input (textual, visual, kinaesthetic and auditory) are provided;

to investigate test-takers’ perceptions regarding the impact of the input media on their performance and the relation to test-takers’ sensory preference(s).

Methodology

The research questions are as follows:

1. Is there a relationship between listening ability and performance on an independent writing test and an integrated EAP listening-into-writing test?

2. To what extent does output performance on an independent EAP writing task predict output performance on an integrated EAP listening-into-writing test?

3. How does the input format of an academic lecture affect performance on an integrated EAP listening-into-writing task?

4. How do learners’ perceptions of their performance compare with observed performance on an integrated EAP listening-into-writing task?

The aim of the study is to find out about the relationships that exist between different input formats and written output.

All data will be anonymised and informed consent will be sought from all participants.
The study comprises four steps:

1. Independent IELTS academic listening test (25 mins)
2. Independent IELTS academic writing test (30 mins)
3. Integrated EAP listening-into-writing test (80 mins in total)
4. Post-task questionnaire (approx. 10 mins);

Test format

An independent (IELTS-type) listening comprehension test will assess general listening comprehension (25 mins);

A short independent (essay-)writing task will assess general writing competence (30 mins);

A specifically developed integrated EAP listening and writing task. Test-takers will watch/listen to a content-based video of a lecture (on a general academic topic) in different (kinaesthetic, visual, auditory and read/write) formats and produce a thematically-linked written output. The same auditory input will be provided in each case (80 mins).

There will be a break after you have done the first two tests (i.e. before you do the third test).
Information sheet for participants

First of all, thank you for showing an interest in participating in this study, which will be the basis of the researcher’s PhD thesis. Participation is completely voluntary and you may withdraw at any time.

Please see the information sheet provided for information about the specific tasks.

The data you provide (your test results, post-test questionnaires and, where applicable, the interview data) will be analysed to see if the type of input has an effect on your performance.

The participant with the highest score overall will win a £10 Amazon.co.uk token and there will be two runners-up tokens of £5 each. However, please note that the performance on these tasks does not affect your academic course in any way.

You are more than welcome to find out your test result. If you wish to do that, please send me an e-mail: Carolyn.westbrook@solent.ac.uk

It is intended that the finished research will be published in papers and journal articles as well as through the researcher’s PhD thesis. If you wish to have a copy of the results of the study, please send me an e-mail at the above address, too.

If you are willing to participate, please complete and sign the consent form overleaf and return it to the tutor/researcher.
Consent form for participants

First of all, thank you for agreeing to participate in this study, which will be the basis of the researcher’s PhD thesis.

Please read the consent information below and tick the boxes to confirm your agreement.

Then please sign your name below and write your name clearly.

I confirm that:

- I have been given clear and detailed information about the study I will be involved in.
- I understand that participation is completely voluntary and I may withdraw at any time.
- I agree to my data being used for the purposes of this study.
- I am aware that all data will be anonymised and any personal data will be treated as confidential.
- I am aware of how the results of this study will be disseminated (= passed on to other people) and I agree to this.
- I understand that I may request a copy of my results and of the results of the study by sending an e-mail to the researcher and that I have been given the researcher’s e-mail address.

Signed: ____________________________________________

Name: (please print) ________________________________

THANK YOU FOR YOUR PARTICIPATION AND YOUR HELP. THEY ARE GREATLY APPRECIATED!
Appendix 2: Exploratory study - IELTS Listening test with transcript and answers (UCLES, 2002).

STUDENT NAME _________________________________

STUDENT ID _________________________________

LISTENING TEST

Instructions

The time allowed for this test is 25 minutes. Dictionaries are not allowed.

Write in blue or black ink and write your answers in the spaces provided on the question paper. Do not write in red ink or pencil.
NAME: ____________________________

ID: ______________________________

WRITING

You should spend about 40 minutes on this task. Present a written argument or case to an educated reader with no specialist knowledge of the following topic:

Children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

To what extent do you agree or disagree with this opinion?

You should use your own ideas, knowledge and experience and support your arguments with examples and relevant evidence.

You should write approximately 250 words. You will lose points if you write much less than this and if you write much more than this.

Write your answer below.

___________________________________________________________________________

___________________________________________________________________________

___________________________________________________________________________

___________________________________________________________________________

___________________________________________________________________________

___________________________________________________________________________

___________________________________________________________________________
Appendix 4: Exploratory study - INTEGRATED EAP TEST

NAME __________________________________________

ID (e.g. B01) ______________________________________

You are going to do a test in which you will listen to a short lecture on the topic of “culture shock”. You should make notes on what you hear. After that, you will write a text on the topic of the lecture. The title of the essay is:

‘What is culture shock and how can we reduce the effects of culture shock?’

The lecture will be presented in four sections. You should take notes on each section in the spaces provided below.

You will have time at the end of each section of the lecture to check your answers.

You will hear each section twice.

You should write your text on the paper provided and you should include as much information from the lecture as possible. You will be given headings for your text to help you.

The lecture will take about 35 mins in total. You then have 45 minutes to write your essay.
LECTURE

Section 1

You are going to hear an explanation about what culture shock is.

Make notes below. Write down as much information as you can.

These notes will NOT be assessed as part of the listening test but you will need the information for the speaking test later.

You will hear this section TWICE. There will be a 30-second pause before you hear the section for the second time so you can check your work.

- What is culture shock?
- Definitions of culture shock
You now have 1 minute to check your answers.

When you have checked your answers, turn over and read the instructions for section 2.
Section 2

You are going to hear an explanation the symptoms of culture shock.

Make notes below. Write down as much information as you can.

These notes will NOT be assessed as part of the listening test but you will need the information for the speaking test later.

You will hear this section TWICE. There will be a 30-second pause before you hear the section for the second time so you can check your work.

Note: this section is audio only (you will not see anything in this section).

• Symptoms of culture shock

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
You now have 1 minute to check your answers.

When you have checked your answers, turn over and read the instructions for section 3.
Section 3

You are now going to hear an explanation of the five stages of culture shock.

Make notes on the five stages under the five headings below. Write down as much information as you can.

You will hear this section TWICE. There will be a 30-second pause before you hear the section for the second time so you can check your work.

• The Honeymoon Stage

• The Disintegration (or Distress) Stage
• The Reintegration Stage

• The Autonomy Stage

• The Independence Stage

You now have 1 minute to check your answers.

When you have checked your answers, turn over and read the instructions for section 4.
Section 4

You are going to hear an explanation about how long culture shock lasts and how you can reduce the effects of culture shock.

Make notes under the headings below. Write down as much information as you can.

You will hear this section TWICE. There will be a 30-second pause before you hear the section for the second time so you can check your work.

- How long does culture shock last?

- Reducing culture shock - Preparation
• Reducing culture shock - Awareness

• Reducing culture shock - Integration

You now have 1 minute to check your answers.

When you have checked your answers, turn over and complete the writing task.
WRITING TASK

Using the information you have written down from the lecture, you are now going to write a text for a non-specialist audience of about 500 words entitled:

What is culture shock and how can you reduce the effects of culture shock?

You should include the following areas from the lecture in your essay:

1. Introduction to culture shock.
   a. What is culture shock?
   b. Definitions of culture shock
2. Symptoms of culture shock
3. The five stages of culture shock
4. How long does culture shock last?
5. Reducing the effects of culture shock:

You will be marked on the language, content and structure of your essay.

You may make notes below but please cross out any notes that you do not want to be marked.

You have 45 minutes to write the essay. Write your essay below.

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
Appendix 5: Exploratory Study - Lecture Transcript: Culture Shock

NOTE: Highlighted sections are the key points.

PART ONE

Hello everybody and welcome to today’s lecture on the topic of culture shock. Well, today I’m going to be talking to you about what is culture shock; what do we mean by it? We’ll then look at the symptoms of culture shock, that is, how you may feel if you suffer from culture shock. We’ll then look at the 5 stages of culture shock and finally, I’ll tell you some things you can do to reduce the effects of culture shock so that maybe you don’t suffer as much as you might otherwise.

Ok, so let’s start with a definition. What do we mean by culture shock? Well, erm, when we go to a different country, we often find that things are done slightly differently in the new country to how things are done in our own country and I’m sure most of you here today will have experienced that to some extent, erm, and as a result, we may feel a little bit unsure of ourselves, maybe not too happy about how things work and this disorientation, this unhappiness that we experience when we don’t know how things work, this is what Dr Kalervo Oberg refers to as culture shock. And Dr Oberg, that’s O-B-E-R-G, decided upon this term, culture shock, some 50 or 60 years ago now so it’s not a new concept; it’s been around for a while. Erm, another famous researcher in terms of intercultural competence and intercultural communication is a gentleman called Dr Hofstede, that’s H-O-F-S-T-E-D-E, erm, and what Dr Hofstede said is that culture shock is like being a child again, so maybe a 1-year-old or a 2-year-old, erm, where everything we ever knew about the world no longer applies so if I take you from your country and your environment that you’re happy with, you understand, you know how things work, I take you out of there, transport you halfway across the world, which is what some of you have done and I put you into a new environment that you don’t know at all, it works in a completely different way, people talk to each other in a completely different
way, erm, suddenly you may feel very unsure of yourself, erm, and this disorientation, this insecurity, is what we mean by culture shock because you’re in a situation where everything that you once knew about the world has changed, so that’s what we mean by culture shock.

(406 words – 3:14 = 125.6 wpm)

PART TWO

Ok, so let’s move on to the second part of the lecture, so how might you feel if you’re suffering from culture shock? Well, there are quite a few symptoms that you might experience and these can involve, erm, depression, so you may feel depressed; you may feel lonely or sad. Erm, you may also feel that you don’t have any confidence - where once you used to be quite confident, now you’re not very confident any more. Erm, and you may also find that you can’t solve even the smallest of problems so something that is normally a very small problem becomes quite a big issue, quite a big problem, for you. You may also find that you suddenly feel insecure or unsure about things and you may even become very shy or withdrawn so suddenly you might not want to talk to people. Erm, another symptom is that you might feel very emotional, so more emotional than usual, and maybe cry, erm, at something that normally would never make you cry. Some people, when they experience culture shock, say that they feel ill so they constantly think they’ve got a headache or that they’re coming, they’re going to get a cold, or something or they feel very tired, or alternatively, they can’t sleep at all. Erm, you may find that you’re suffering from stress as well, because if you’re in a situation (like you are) where you have to speak a different language all the time, that can actually be quite a strain, it can be quite hard work and you’re suddenly surrounded by that language so those are some of the symptoms of culture shock, which sounds very negative really but it’s important to know that it’s not negative. In fact, it’s perfectly natural and it’s perfectly normal. And there’s nothing wrong with suffering from culture shock. In fact, and you may not believe this, it can actually be very positive in the long term and I say that because it can actually make you appreciate certain aspects, certain parts, of your own culture that you weren’t
perhaps aware of or weren’t conscious of before you came to the new culture. And it will be a positive learning experience; it will provide you with valuable skills for the future and you may also find that there are certain aspects of the new culture that you also value by the end of your time here. So those are some of the symptoms of culture shock and, as I say, it’s important to remember that it is perfectly normal. (430 words – 3:45: 114.7 wpm)

PART THREE

Ok, let’s have a look at the next part, which is the various stages of culture shock. If you look at this diagram, you can see that it looks a little bit like a “W”. That’s because the first “U”, on the left, represents the time in the new country and the second one, on the right, is when you return home to your own country.

So, let’s go through the five stages. The first one is the Honeymoon Stage - that’s up at the top left – and this is when you’ve only just arrived in the new country. You’re happy, you’re excited because everything that you’re seeing and learning is new, and also, because you’ve only just arrived, memories of friends and family back home are still close to your heart so you still feel quite happy.

However, this will soon come down to the Disintegration Stage, which is the stage at the bottom of the “U”, and at this point, this is where you discover that there are cultural differences and you may feel unhappy and lonely because you don’t know how to deal with those cultural differences so this stage is sometimes called the Distress Stage. You may also feel confused or uneasy because things are so different and, of course, you you you know, you probably don’t have friends or family close by that you can talk to, so that’s why you may feel lonely.

This then turns into the Reintegration Stage, which is about half way up the right hand side of the “U”, and, at this point, you start to dislike the new culture and you get together with people from your own country to complain about the new culture. Now that sounds pretty
negative, but, again, it is perfectly normal and it just means that you now value certain parts of your own culture.

After that, we go to the Autonomy Stage, which is a little bit further up on the right. And in the Autonomy Stage, you start to realise that, yes, things are done differently and you may not like how things are done but, in actual fact, you can live with it and it’s not a problem. You’ll find at this point that those little problems at the beginning that caused you such heartache and suddenly became big problems are now little problems again and you can easily deal with them.

And once you’ve passed that stage, you come into the Independence Stage, which is the part at the top of the “U” curve on the right. And in the Independence Stage, that’s where you’re perfectly happy with the new culture. You know how things work. You know how life works. You may not do everything in the same way as people in the new culture but you’re perfectly happy with the situation. You can live with it and you can deal with it. So that’s the Independence Stage.

Now that’s the end of the “U” curve but if we move on to the “W” part, that’s actually when you return home and you may not believe it, but you’ll go through exactly the same stages all over because you’ve a different person. You’ve adapted to a different way of life so you’re going back to a different way of life that you’re no longer used to when you return home. (562 words – 4:02=138.2wpm)

PART FOUR

Ok, so let’s come to the fourth and final part of today’s lecture, which is reducing the effects of culture shock. So you’re probably wondering how long culture shock will last. Well, it depends. It depends on many things. It depends on you as an individual. It depends on your attitude to the new culture (whether you like it or not). It depends on the situation that you find yourself in, so are you having a good time and making lots of friends or are you maybe
having problems, maybe with your classwork or maybe with your landlord. All of these things will impact on the length of time that it takes you to get over culture shock. It also depends on how different the new culture is from your own culture. Clearly, if you’ve travelled half way around the world to a completely different culture, it’s going to last longer, really, than someone who’s maybe just come from Europe, where things are relatively similar, and, of course, it depends on the steps that you take to minimise culture shock, so that’s what we’re going to look at.

There are three key stages and these are preparation, awareness and integration.

The first stage is preparation and that means prepare yourself well before you come to the new country. Now I recognise that, for you here now, it might be a bit late but you may also have actually prepared yourself before you came. You may have found out about the new country or about the new culture. Basically, anything that you can find out about the new country will help you to reduce the effects of culture shock. So what is the weather like? What time do the shops open and close? How do people talk to each other? Anything like that will help you with the effects of culture shock.

The next thing is to be aware. If you can, be aware of your own culture and the kinds of things in your own culture that might be different elsewhere but remember that culture shock exists. Just remembering that it exists and knowing about it will help you to reduce the effects. And also, be aware of the symptoms. If you know that you’re feeling depressed, then you can say, “Ok, is it because of culture shock?” and if it is, then that will help you to deal with it. If you’re feeling lonely, the same thing. Tell yourself you know it’s because of culture shock and it won’t last forever.

And the third thing is integration. Join a club or find a hobby. If you enjoy sports, maybe you play football or basketball, find a local football club or basketball club. Join in with the local community. Maybe you fancy learning something new – maybe you fancy learning to paint –
go to an evening class at a local college and mix with the native speakers. In other words, if you can integrate into the local community, that will help you and will help it will help you, let’s say, to reduce the effects of culture shock.

Now those are the three steps. The other thing to remember is that culture shock is perfectly normal and you’re not the only one to experience it.

So, that’s the end of today’s lecture. Thank you very much.

(555 words – 3:54: 142.3 wpm)
## Appendix 6: Exploratory study - IELTS cohesive devices

| IELTS cohesive device | B1 | B2 | B4 | B5 | B6 | B7 | B8 | B9 | B1 | B1 | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R1 | R1 | TOTALS LS |
|-----------------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|-----------|
| And                   | 3  | 10 | 5  | 3  | 2  | 2  | 2  | 2  | 5  | 8  | 11 | 7  | 12 |    |    |    |    | 3  | 5  | 2  | 3  | 10 | 4  | 6  | 103      |
| Or                    | 1  | 1  | 2  |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 3  | 1  | 2  | 4  | 5  |    | 32       |
| But                   |    |    |    | 2  |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 2  |    |    |    |    | 1  | 14       |
| So                    | 1  | 2  | 1  | 1  | 2  | 2  | 2  | 2  | 1  | 2  |    |    | 1  |    | 3  | 3  |    |    |    |    |    |    |    | 21       |
| Because               | 2  | 3  |    | 2  |    | 3  |    | 3  | 2  |    |    |    |    |    |    |    |    |    |    |    |    | 1  |    |    | 29       |
| As we know            |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 2        |
| In other words        |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 1        |
| There are             | 1  |    |    |    |    |    |    |    |    |    | 2  |    | 1  |    | 1  |    |    |    |    |    |    | 1  |    |    | 10       |
| In my opinion / To my mind / In my personal view / As far as I am concerned | 1  |    | 1  | 1  | 1  |    | 1  |    |    |    |    |    |    |    |    |    |    |    |    |    | 1  |    |    | 8        |
| I think               |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 4        |
| Furthermore           |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 4        |
| Also                  | 1  | 2  | 2  |    |    | 2  |    | 1  |    |    |    |    |    | 1  |    |    |    |    |    |    |    |    | 1  |    | 11       |
| Moreover              | 2  |    |    |    |    |    |    |    | 1  |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 1  |    | 5        |
| In addition           |    |    |    |    |    |    |    |    | 1  |    |    | 1  |    |    |    |    |    |    |    |    |    |    |    |    | 3        |
| Another               |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 1        |
| For one thing         |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 1        |
| For another (thing)   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 1        |
| For example / I'd like to give an example / For instance | 1  |    | 1  | 4  | 2  | 1  | 2  | 3  | 1  |    |    |    |    |    |    |    |    |    |    |    | 4  |    |    | 26       |
| Firstly               |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 1        |
| The first one ...     | 1  |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 1        |
| First of all          |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 1  |    |    | 3        |
| ... the second example / the second ... | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 1 |
| Thirdly |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Then |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 0 |
| On the one (hand) |   |   |   |   | 1 |   |   |   |   |   |   |   |   |   |   |   | 1 |
| However | 1 | 2 | 1 |   | 2 | 1 | 2 |   | 2 | 2 | 13 |
| Alternatively |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 1 |
| In contrast | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 4 |
| Conversely |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 1 |
| On the other (hand) | 1 | 1 | 1 | 1 |   | 1 | 1 | 1 |   |   | 7 |
| Although | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |   |   | 10 |
| In spite of |   |   |   |   |   |   |   |   |   |   | 1 |
| Therefore | 1 | 1 |   |   | 1 | 1 | 1 |   |   | 1 | 2 | 1 | 9 |
| As a result |   |   |   |   |   |   |   |   |   |   |   |   |   | 1 |
| To sum up / In summary |   |   |   |   | 1 | 1 | 1 |   |   |   |   | 4 |
| In conclusion |   | 1 | 1 | 1 | 1 |   |   |   | 1 | 1 | 1 | 1 | 9 |
| Normally |   |   | 1 |   |   |   |   |   |   |   |   |   |   | 2 |
| As a rule |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 1 |
| To support that opinion | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 1 |
| Due to (the fact that) ... | 1 | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 2 |
| **TOTAL No. OF COHESIVE DEVICES** | **11** | **24** | **18** | **12** | **14** | **19** | **11** | **17** | **33** | **33** | **18** | **22** | **1** | **11** | **18** | **10** | **1** | **5** | **35** | **18** | **15** | **355** |
## Appendix 7: Exploratory study - EAP cohesive devices

| EAP cohesive device | B1  | B2  | B4  | B5  | B6  | B7  | B8  | B9  | B10 | B1  | R1  | R2  | R3  | R4  | R5  | R6  | R7  | R8  | R12 | R13 | TOTALS |
|---------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------|
| And                 | 8   | 27  | 13  | 16  | 8   | 12  | 11  | 18  | 9   | 19  | 17  | 14  | 20  | 6   | 9   | 13  | 14  | 12  | 9    | 17   | 272   |
| Or                  | 5   | 1   | 1   | 2   | 2   | 6   | 10  | 12  | 1   | 2   | 1   | 4   | 1   | 1   | 6   | 6   | 1   | 62   |      |       |
| But                 | 1   | 1   | 2   | 3   | 2   | 1   | 1   | 10  | 1   | 2   | 2   | 2   | 2   | 2   | 2   | 2   | 2   | 20   |      |       |
| So                  | 1   | 3   | 2   | 1   | 2   | 4   | 3   | 1   | 1   | 1   | 3   | 3   | 3   | 3   | 3   | 3   | 3   | 38   |      |       |
| Because             | 2   | 5   | 5   | 1   | 7   | 3   | 2   | 5   | 3   | 3   | 2   | 2   | 2   | 2   | 2   | 2   | 7   | 1    | 3    | 59    |
| To my mind          |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     | 1    |       |       |
| There are (5 stages) ... / (Culture shock) consists of / shows / has (5 stages) | 1   | 3   | 4   | 1   |     |     | 1   | 4   | 1   | 1   | 1   | 3   | 3   | 3   | 3   | 3   | 3   | 3   | 38   |       |       |
| We experience 5 stages of CS |     |     |     |     |     |     |     |     | 1   |     |     |     |     |     |     |     |     |     |     | 1    |       |       |
| (The) first one / stage / step is | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 2   | 2   | 1   | 1   | 1   | 1   | 1   | 14   |       |       |
| The second (one is) | 1   | 1   |     | 1   |     | 2   | 2   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 11   |       |       |
| ... the second stage ... | 1   |     |     |     |     | 1   | 1   |     |     |     |     |     |     |     |     |     |     |     |     | 4    |       |       |
| The third / fourth / fifth one / stage is ... / Stage number three / four is ... | 2   | 2   |     | 1   |     | 3   |     |     |     |     |     |     |     |     |     |     |     |     |     | 14   |       |       |
| In this stage / In this / that one | 1   | 2   | 1   |     | 1   | 3   | 5   |     |     |     |     |     |     |     |     |     |     |     |     | 17   |       |       |
| Next                |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     | 1    |       |       |
| The next stage ...  | 1   | 1   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     | 4    |       |       |


<p>| After that |  |  |  |  |  |  | 1 | 2 | 1 | 1 | 10 |
| Then | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 |
| the last / final / stage / step is | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 15 |
| First(ly) / At first / First of all | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 11 |
| Secondly | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 6 |
| Thirdly | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| Fourthly | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 15 |
| Finally, | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 11 |
| Moreover | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 7 |
| Furthermore | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4 |
| Also | 1 | 1 | 1 | 3 | 3 | 1 | 4 | 1 | 15 |
| One (…) is | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 |
| (An)other (…) is | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 1 |
| Besides | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4 |
| Alternatively | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| In addition | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 |
| and so on | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 |
| However | 1 | 2 | 1 | 1 | 1 | 3 | 2 | 11 |
| On the other hand | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| While | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 5 |
| Although | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4 |
| Therefore/Thus | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4 |
| As a result | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4 |
| For example / For instance / such as | 4 | 3 | 2 | 1 | 1 | 1 | 2 | 3 | 5 | 2 | 3 | 1 | 1 | 5 | 2 | 35 |</p>
<table>
<thead>
<tr>
<th>Luckily</th>
<th>Normally</th>
<th>Moving (on) to</th>
<th>In short</th>
<th>According to</th>
<th>Due to (the fact)</th>
<th>In conclusion</th>
<th>TOTAL No. of COHESIVE DEVICES PER STUDENT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15 53 39 36 26 32 31 30 43 50 45 38 50 27 21 32 30 49 31 41 719</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 3 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
</tr>
</tbody>
</table>

TOTAL No. of COHESIVE DEVICES PER STUDENT: 719
## Appendix 8: Exploratory study - EAP lecture propositions

<table>
<thead>
<tr>
<th>EAP lecture propositions</th>
<th>B1</th>
<th>B2</th>
<th>B4</th>
<th>B5</th>
<th>B6</th>
<th>B7</th>
<th>B8</th>
<th>B9</th>
<th>B10</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
<th>R5</th>
<th>R6</th>
<th>R7</th>
<th>R8</th>
<th>R12</th>
<th>R13</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part one (seeing and hearing the speaker)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>things are done slightly different(ly)</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>unsure</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not happy about how things work</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Dr Kalervo Oberg's definition of culture shock</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>culture shock not (a) new concept</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>another intercultural competence and (intercultural) communication</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Dr Hofstede's definition of culture shock (like) being a child again</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>everything you knew about the world no longer applies</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>put you in(to) a new environment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>people talk in a different way</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>disorientation</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>insecurity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>
(everything) you know about the world has changed

Part two (hearing the speaker only)
how might you feel?
symptoms
depression / depressed
lonely
(feel) sad
(no) confidence
you can't solve even (the smallest of) problems
insecure
unsure
shy
withdrawn
may not want to talk to people
(more / very) emotional
cry (at something that would not normally make you cry)
(feel) ill
headache
(get) a cold
tired
can not sleep

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>(everything) you know about the world has changed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part two (hearing the speaker only)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>how might you feel?</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>symptoms</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>depression / depressed</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>lonely</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>(feel) sad</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>(no) confidence</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>you can't solve even (the smallest of) problems</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>insecure</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>unsure</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>shy</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>withdrawn</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>may not want to talk to people</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>(more / very) emotional</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>cry (at something that would not normally make you cry)</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>(feel) ill</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>headache</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>(get) a cold</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>tired</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>can not sleep</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>feel stress(ed) / stressful / suffer(ing) from stress</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>speak a different language every day (can be a strain / hard work)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>perfect(ly) normal (and (perfectly) natural)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>positive in the long term</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>make you appreciate aspects of your own culture / value aspects of your own culture</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>positive learning experience</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(gain) valuable skills for (the) future</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Part three (hearing the speaker and seeing the diagram)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 stages of culture shock</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>looks like a 'W' curve</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>'U' on the left is the new culture</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>'U' on the right: return to old culture</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Honeymoon Stage</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(you have) just arrived</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>feel happy</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>----------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>excited / exciting</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>everything (you're seeing and learning) is new</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>memories (of friends and family) are still close to your heart</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Disintegration Stage</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>discover that there are cultural differences</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(feel) unhappy</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(feel) lonely</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>you don't know how to deal with (them / the differences)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>also called Distress Stage</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(feel) confused</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>uneasy</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>probably don't have friends or family close by (that you can talk to)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Reintegration Stage</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>start to dislike (the new culture)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>you get together with people from your own country</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>complain about the new country</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>now value (parts of) your own culture</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>you can live with (it / them)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Autonomy Stage</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>realise things are done differently</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>may not like the way things are done</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>but you can live with it</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>little problems that were big problems are now little again</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>you can deal with (it / them)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>perfectly happy (with the new culture)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>you know how things work / how life works</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>may not do (every)thing in the same way as people (in) the new culture</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(go through the) same stages when you return home</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>you've become a different person</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>you've adapted to a different way of life</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>you're going back to a culture you're no longer used to</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Part four (hearing the speaker and seeing words on the slide)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reducing the effects of culture shock</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>how long culture shock will last</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>depends on the individual</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(depends) on the attitude (to the new culture)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>if you like it or not</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(depends on the) situation (you find yourself in)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>how different the new culture is (from your own culture)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>depends on the steps you take to minimise it</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>three key stages</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>preparation</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>prepare yourself before you come to the new country</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>anything you can find out will help (to reduce the effects of culture shock)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>weather</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shops</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>how people talk to each other</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>awareness</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>be aware (of your own culture)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>things (in your culture) that might be different elsewhere</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>remember that culture shock exists</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>knowing about it will help (to reduce the effects)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>be aware of the symptoms</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>if you are feeling lonely</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>tell yourself it is because of culture shock</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>it won't last forever</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>integration</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>join a club</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>find a hobby</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>enjoy sports</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>find a football club or a basketball club</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>join in with / integrate into the local community</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>(fancy) learn(ing) something new</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>
| go to an evening class  
(at a local college) | 0 |
| mix with native speakers | 1 1 3 |
| ... will help to reduce culture shock | 1 1 1 1 1 |
| remember that culture shock is perfectly normal and natural | 1 |
| you're not the only one to experience it | |
| TOTAL No. of EAP PROPOSITIONS | 24 36 65 66 31 58 48 43 55 52 49 49 53 24 60 58 32 63 63 61 990 |
**Appendix 9: Exploratory study - Percentage of propositions reproduced**

<table>
<thead>
<tr>
<th>Percentage of propositions</th>
<th>B1</th>
<th>B2</th>
<th>B4</th>
<th>B5</th>
<th>B6</th>
<th>B7</th>
<th>B8</th>
<th>B9</th>
<th>B10</th>
<th>B11</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
<th>R5</th>
<th>R6</th>
<th>R7</th>
<th>R8</th>
<th>R12</th>
<th>R13</th>
<th>MEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part one totals</td>
<td>21</td>
<td>29</td>
<td>71</td>
<td>43</td>
<td>21</td>
<td>36</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>43</td>
<td>29</td>
<td>21</td>
<td>57</td>
<td>43</td>
<td>36</td>
<td>36</td>
<td>43</td>
<td>36</td>
<td>36</td>
<td>21</td>
<td>34</td>
</tr>
<tr>
<td>Part two totals</td>
<td>23</td>
<td>31</td>
<td>58</td>
<td>81</td>
<td>23</td>
<td>62</td>
<td>54</td>
<td>35</td>
<td>62</td>
<td>73</td>
<td>54</td>
<td>38</td>
<td>58</td>
<td>46</td>
<td>65</td>
<td>77</td>
<td>31</td>
<td>77</td>
<td>65</td>
<td>69</td>
<td>54</td>
</tr>
<tr>
<td>Part three totals</td>
<td>36</td>
<td>44</td>
<td>72</td>
<td>64</td>
<td>36</td>
<td>46</td>
<td>51</td>
<td>54</td>
<td>49</td>
<td>46</td>
<td>49</td>
<td>41</td>
<td>51</td>
<td>15</td>
<td>49</td>
<td>62</td>
<td>46</td>
<td>56</td>
<td>54</td>
<td>51</td>
<td>49</td>
</tr>
<tr>
<td>Part four totals</td>
<td>3</td>
<td>19</td>
<td>33</td>
<td>39</td>
<td>22</td>
<td>53</td>
<td>31</td>
<td>28</td>
<td>47</td>
<td>25</td>
<td>33</td>
<td>56</td>
<td>28</td>
<td>0</td>
<td>53</td>
<td>25</td>
<td>0</td>
<td>44</td>
<td>56</td>
<td>56</td>
<td>33</td>
</tr>
<tr>
<td>TOTAL</td>
<td>21</td>
<td>31</td>
<td>57</td>
<td>57</td>
<td>27</td>
<td>50</td>
<td>42</td>
<td>37</td>
<td>48</td>
<td>45</td>
<td>43</td>
<td>43</td>
<td>46</td>
<td>21</td>
<td>52</td>
<td>50</td>
<td>28</td>
<td>55</td>
<td>55</td>
<td>53</td>
<td>43</td>
</tr>
</tbody>
</table>
Appendix 10: Exploratory study - Number of reproductions from the EAP slides

| EAP slides                        | B1 | B2 | B4 | B5 | B6 | B7 | B8 | B9 | B10 | B11 | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R12 | R13 | TOTALS |
|----------------------------------|----|----|----|----|----|----|----|----|----|-----|----|----|----|----|----|----|----|-----|-----|--------|
| Preparation                      | 1  |    | 1  | 1  | 1  | 1  | 1  | 1  | 1   | 1   | 1  | 1  |    |    |    |    |    |    |    |      |
| find out about the host culture  | 1  |    |    |    |    |    |    | 1  |    |    |    |    |    |    |    |    |    |    |    |      |
| find out about the host country  |    |    |    |    |    |    |    |    |    |    | 1  | 1  | 1  |    |    |    |    |    |    |      |
| Awareness                        | 1  | 1  | 1  | 1  |    | 1  | 1  | 1  | 1   |    |    |    |    |    |    |    |    |    |    |      |
| be aware of your own culture     | 1  | 1  |    |    |    | 1  |    | 1  | 1   |    |    |    |    |    |    |    |    |    |    |      |
| remember that culture shock exists| 1  | 1  | 1  | 1  | 1  | 1  |    | 1  |    |    |    |    |    |    |    |    |    |    |    |      |
| aware of the symptoms            | 1  | 1  | 1  | 1  | 1  | 1  | 1  |    |    | 1   | 1   | 1  | 1  |    |    |    |    |    |    |      |
| Integration                      | 1  | 1  | 1  | 1  | 1  | 1  | 1  |    |    |    |    |    |    |    |    |    |    |    |    |      |
| join a club                      | 1  |    | 1  | 1  | 1  | 1  | 1  | 1  |    |    |    |    |    |    |    |    |    |    |    |      |
| find a hobby                     | 1  | 1  |    | 1  | 1  | 1  | 1  | 1  |    | 1   |    | 1  | 1  |    |    |    |    |    |    |      |
| integrate into the local community| 1  | 1  | 1  | 1  | 1  |    | 1  | 1  | 1   | 1   |    |    |    |    |    |    |    |    |    |      |
| TOTAL                            | 0  | 0  | 3  | 6  | 3  | 8  | 4  | 1  | 11  | 0   | 5  | 8  | 5  | 0  | 9  | 5  | 0  | 6   | 8   | 11 | 93    |
Appendix 11: Exploratory study - Screenshot of input – part three

Appendix 12: Exploratory study - Feedback questionnaire

This short questionnaire is about your experience of taking the language test.

Your answers will help me to see how the kind of input (video, diagram, text, etc.) affected your test performance.

The questionnaire should not take more than 2-3 minutes to complete

If you do not understand any of the questions, please ask the tutor for help.

When you finish, please give your questionnaire to the tutor before you leave the room.

Please note: your answers to this questionnaire have NO impact on your test results so please feel free to give honest answers!!

Identification no. ______________________________________

Name (optional) ______________________________________

Nationality: ______________________________________

Course of study ______________________________________

Age (please tick): 18-21 □

22-25 □

26-30 □

31-40 □

>40 □
Gender (please tick):

Male ☐
Female ☐

Think about your performance on the first listening test. Do you think your results will demonstrate your listening ability accurately?

- I think I did much better than I normally would on a listening test ☐
- I think I did better than I normally would on a listening test ☐
- I think I did as well as I normally would do on a listening test ☐
- I think I did worse than I normally would do on a listening test ☐
- I think I did much worse than I normally would do on a listening test ☐

On the whole, how difficult did you find the first writing test?

- Very difficult ☐
- Quite difficult but I could cope with it ☐
- Quite easy ☐
- Very easy ☐
How much did you know about the lecture topic before the test?

I knew a lot about the topic before  [ ]

I knew a little about the topic before  [ ]

I knew nothing about the topic before  [ ]

Do you think the topic of the test had a positive or negative impact on how well you performed?

I think the topic had a positive impact on my performance  [ ]

I don’t think the topic had any impact on my performance  [ ]

I think the topic had a negative impact on my performance  [ ]

I don’t know whether the topic had any impact on my performance  [ ]

Think of the different sections of the lecture. Which part do you think you performed best on?

Section 1 – hearing and seeing the speaker  [ ]

Section 2 – hearing the speaker only  [ ]

Section 3 – hearing the speaker and seeing the diagram  [ ]

Section 4 – hearing the speaker and seeing the words  [ ]
Why do you think you performed best on this section?

_________________________________________________________________________

Think of the different sections of the lecture. Which part do you think you performed worst on?

Section 1 – hearing and seeing the speaker

Section 2 – hearing the speaker only

Section 3 – hearing the speaker and seeing the diagram

Section 4 – hearing the speaker and seeing the words

Why do you think you performed worst on this section?

_________________________________________________________________________

Do you think any of the following had a negative effect on how well you understood the lecture? (please tick all that apply)

The sound quality

The number of times you saw the video

The amount of information you had to write down
Any other factors (please specify):
_____________________________________________________________________
_____________________________________________________________________

On the whole, how difficult did you find the essay based on the lecture topic?

Very difficult

Quite difficult

Quite easy

Very easy

Did anything affect (positively or negatively) how much you could write in the essay? Please give brief details:
_____________________________________________________________________
_____________________________________________________________________
_____________________________________________________________________

THANK YOU FOR TAKING THE TIME TO COMPLETE THIS QUESTIONNAIRE AND FOR BEING INVOLVED IN THE STUDY.
YOUR HELP AND FEEDBACK ARE VERY MUCH APPRECIATED. 😊
Appendix 13: Main study - Information sheet and consent form

Introduction to the research

This doctoral research will investigate the relationship between input task characteristics (video vs audio input) of listening comprehension texts, and performance on written output tasks.

The objectives are:

- to investigate how test takers perform when input is presented in audio format only compared to video format;
- to investigate the extent to which test takers are influenced by the written word in an EAP lecture compared to the spoken word.

Methodology

The research questions are as follows:

- Do test takers perform better when presented with video input for a lecture?
- To what extent do test takers rely on the written word compared to the spoken word?

The aim of the study is to find out how test takers’ performance on an integrated EAP test varies when the input (i.e. the lecture content) is presented in audio and video formats.

To assess the differences in performance, the data will be assessed qualitatively and quantitatively if numbers allow.

All data will be anonymised and informed consent will be sought from all participants.

The study comprises three steps:

- a quick lexico-grammatical placement test (approx. 30 mins)
- a language test comprising an academic lecture and a follow-up writing task (approx. 80 mins);
• a brief post-task questionnaire (approx. 5 mins);

Test format

A specifically developed integrated EAP listening and writing task. Test-takers will watch/listen to a short EAP lecture (on a general academic topic) presented in audio only and video formats and produce a thematically-linked written summary. The same auditory input will be provided in each case (90 mins).

The research will be conducted as follows: one x 30 min followed by one x 1.5 hour session.
Information sheet for participants – Main study

First of all, thank you for showing an interest in participating in this study, which will be the basis of the researcher’s PhD thesis. Participation is completely voluntary and you may withdraw at any time.

Please see the information sheet provided for information about the specific tasks.

The data you provide (your test results, post-test questionnaires and, where applicable, the interview data) will be analysed to see if the type of input has an effect on your performance.

The participant with the highest score overall will win a £10 Amazon.co.uk token and there will be two runners-up tokens of £5 each. However, please note that the performance on these tasks does not affect your academic course in any way.

You are more than welcome to find out your test result. If you wish to do that, please send me an e-mail: Carolyn.westbrook@solent.ac.uk

The initial findings from this research were presented at the 2011 Language Testing Forum and at EALTA 2013. It is intended that the finished research will be published in papers and journal articles as well as through the researcher’s PhD thesis. If you wish to have a copy of the results of the study, please send me an e-mail at the above address, too.

If you are willing to participate, please complete and sign the consent form overleaf and return it to the tutor/researcher.
Consent form for participants

First of all, thank you for agreeing to participate in this study, which will be the basis of the researcher’s PhD thesis.

Please read the consent information below and tick the boxes to confirm your agreement. Then please sign your name below and write your name clearly.

I confirm that:

- I have been given clear and detailed information about the study I will be involved in.
- I understand that participation is completely voluntary and I may withdraw at any time.
- I agree to my data being used for the purposes of this study.
- I am aware that all data will be anonymised and any personal data will be treated as confidential.
- I am aware of how the results of this study will be disseminated (= passed on to other people) and I agree to this.
- I understand that I may request a copy of my results and of the results of the study by sending an e-mail to the researcher and that I have been given the researcher’s e-mail address.

Signed: __________________________________________

Name: (please print) __________________________________________

THANK YOU FOR YOUR PARTICIPATION AND YOUR HELP. THEY ARE GREATLY APPRECIATED!
Appendix 14: Ethics form

UNIVERSITY OF BEDFORDSHIRE RS1 (Review)
(This form should be typewritten)

1. Name of Candidate: Carolyn Westbrook
   Student Ref No: 1030574

2. Faculty: Business
   Institute: BMRI / CRELLA

4. Title of report: Manipulating the relationship between input task characteristics and performance on written output tasks

5. Reviewer’s Comments and feedback on report: (see section 5 of the notes of guidance)

   - Does the report set out clearly the proposed plan of work?

   The proposed plan of work is clearly explained for the most part. In some areas, however, it was not entirely clear what the candidate is proposing, possibly owing to the word limit constraints, which unfortunately do not allow for a full presentation of the project.

   For example, the candidate could think about amending these areas:

   - the title is rather misleading – the candidate is surely aiming at investigating, and not manipulating, the relationship?
   - the literature review reads as a set of rather disconnected statements, but this is probably caused by the need for severe editing to fit the word limits.

   Another issue which will need consideration is the assumption in much of the literature, and arguably in this proposal, that all learners in fact have particular sensory preferences or sensory styles, whereas it is arguable possible that some people do not. I would recommend that the candidate start from a point of greater scepticism, or at least neutrality, and seek to investigate the area in a balanced way rather than appearing from the outset to act as advocate for the VARK cause. This would then mean modifying such statements as “I expect to find significant correlations…” A more neutral and academically unbiased approach in the eventual PhD thesis is likely to be more convincing.

   However, in terms of an initial submission, this proposal is interesting and innovative, and is of the required standard overall.

   - Is the scope of the project suitable? (i.e., not too broad or too specific; the aims and objectives are clearly defined and achievable etc.)

   The scope of the project is suitable and viable.

   - Are the methodology(ies) clearly explained and appropriate to the work done/ to be done?

   On the whole, the methodologies are explained clearly. The candidate could consider these areas, however:

   - it is not clear how many participants would be involved, but no doubt this will be made clearer in future drafts.
   - it is not clear whether all participants would complete the same integrated EAP listening and writing task in different sensory formats, or whether some would receive one version and others another version; in fact the mechanism of this aspect of the proposal is rather vague.

   Nonetheless, these issues can be remedied in the coming months.

   In summary, this proposal is of the required standard and the candidate should be allowed to proceed.

March 2011
### UNIVERSITY OF BEDFORDSHIRE RS1(Review)

(This form should be typewritten)

6. Recommendation of the reviewer: (see notes of guidance)

   i) the proposal is approved for onward transmission to URC

<table>
<thead>
<tr>
<th>Name of Reviewer</th>
<th>Stephen Bax</th>
<th>Signature</th>
<th>Date: 8/9/2011</th>
</tr>
</thead>
</table>

7. To the Reviewer: please forward to the Research Student Administrator on completion, retaining a copy for your own records.

8. Decision of Director of Institute - (Include instruction for candidate as appropriate)

   i) the proposal is approved for onward transmission to URC* (minor corrections may be required)

<table>
<thead>
<tr>
<th>Name of Director of Institute</th>
<th>Prof Jillian Farquhar</th>
<th>Signature</th>
<th>Date: 8/9/11</th>
</tr>
</thead>
</table>

9. To the Director of Institute: please return to the Research Student Administrator on completion, retaining a copy for your own records

10. Action taken by RGS:

    Copy of review forms to Candidate & Director of Studies (advising action): -

    i. RS1 and review forms forwarded to URC: or
    ii. Report to be re-submitted

<table>
<thead>
<tr>
<th>Date:</th>
<th>Where re-submission is required - complete next sections as appropriate.</th>
</tr>
</thead>
</table>

11. Resubmission - Reviewers comments

   Recommendation of the reviewer: *delete as applicable

   i) the proposal is approved for onward transmission to URC* (minor corrections may be required)

   ii) the proposal is not approved

<table>
<thead>
<tr>
<th>Name of Reviewer</th>
<th>Signature</th>
<th>Date:</th>
</tr>
</thead>
</table>

To the Reviewer: please forward to the Research Student Administrator on completion, retaining a copy for your own records.

13. Decision of Director of Institute:

   i) the proposal is approved for onward transmission to URC* (minor corrections may be required)

   ii) the proposal is not approved and the candidate should be counselled by the Director of Institute and the Head of the Research Graduate School.

March 2011
### UNIVERSITY OF BEDFORDSHIRE RS1(Review)

(This form should be typewritten)

<table>
<thead>
<tr>
<th>14. Name of Director of Institute:</th>
<th>Signature:</th>
<th>Date:</th>
</tr>
</thead>
</table>

To the Director of Institute: please forward to Research Student Administrator on completion, retaining a copy for your own records.

<table>
<thead>
<tr>
<th>15. Action taken by RGS:</th>
<th>Date:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copy of review forms to Candidate &amp; Director of Studies (advising action):</td>
<td></td>
</tr>
<tr>
<td>i. RS1 and review forms forwarded to URC:</td>
<td></td>
</tr>
<tr>
<td>ii. Co-ordinate with RGS Office with regard to counseling.</td>
<td></td>
</tr>
</tbody>
</table>
SECTION A    To be completed by the candidate

Registration No: 1030574

Candidate: CAROLYN WESTBROOK

Research Institute: BMRI

Research Topic: Manipulating the relationship between input task characteristics and performance on written output tasks

External Funding: NONE

The candidate is required to summarise in the box below the ethical issues involved in the research proposal and how they will be addressed. In any proposal involving human participants the following should be provided:

clear explanation of how informed consent will be obtained,

how will confidentiality and anonymity be observed,

how will the nature of the research, its purpose and the means of dissemination of the outcomes be communicated to participants,

how personal data will be stored and secured

if participants are being placed under any form of stress (physical or mental) identify what steps are being taken to minimise risk

If protocols are being used that have already received University Research Ethics Committee (UREC) ethical approval then please specify. Roles of any collaborating institutions should be
clearly identified. Reference should be made to the appropriate professional body code of practice.

Participants will be informed about the nature of the research, its purpose and the means of dissemination of the outcomes before they take part in any aspect of the study and they will be given detailed information regarding the tasks required of them. Participation will be voluntary and all participants who are willing to take part will be asked to sign an informed consent form and to give permission for their data to be used for the purposes of the study.

For the tasks involved in the study, namely completing a VARK questionnaire and a number of language tests, participants will be allocated an identification number and the researcher will use only this information when presenting and discussing the data. All personal data will be stored in digital form on the researcher’s computer, which requires a password to access data, and/or in hard copy in a locked filing cabinet in the researcher’s home and/or office.

Since participants will be involved in a language test, there will be some aspect of stress involved. However, it will be made clear that the results will not affect their academic course in any way and that the data is for research purposes only.

In terms of accessing participants, permission is being sought from the course leaders / Directors of Studies at the participants’ universities and ethical approval from those institutions will be completed as necessary. The participants with the best test performances in each institution will be entered win a book token.

Answer the following question by deleting as appropriate:

Does the study involve vulnerable participants or those unable to give informed consent (e.g. children, people with learning disabilities, your own students)?

Yes  No

Will the study require permission of a gatekeeper for access to participants (e.g. schools, self-help groups, residential homes)?

Yes  No

Will it be necessary for participants to be involved without consent (e.g. covert observation in non-public places)?

Yes  No
Will the study involve sensitive topics (e.g. sexual activity, substance abuse)? Yes
No

Will blood or tissue samples be taken from participants? Yes
No

Will the research involve intrusive interventions (e.g. drugs, hypnosis, physical exercise)? Yes  No

Will financial or other inducements be offered to participants (except reasonable expenses)? Yes  No

Will the research investigate any aspect of illegal activity? Yes  No

Will participants be stressed beyond what is normal for them? Yes  No

Will the study involve participants from the NHS (e.g. patients or staff)? Yes*  No

If you have answered yes to any of the above questions or if you consider that there are other significant ethical issues then details should be included in your summary above. If you have answered yes to Question 1 then a clear justification for the importance of the research must be provided.

*Please note if the answer to Question 10 is yes then the proposal should be submitted through NHS research ethics approval procedures to the appropriate COREC. The UREC should be informed of the outcome.

Checklist of documents which should be included:

<table>
<thead>
<tr>
<th>Project proposal (with details of methodology) and source of funding</th>
<th></th>
</tr>
</thead>
</table>
Documentation seeking informed consent (if appropriate)

Information sheet for participants (if appropriate)

Questionnaire (if appropriate)

(Tick as appropriate)

Signature of Applicant: Date: 24.05.2011

C. WESTBROOK

Signature of Director of Studies: Date:

This form together with a copy of the research proposal should be submitted to the Research Institute Director for consideration by the Research Institute Ethics Committee/Panel

Note you cannot commence collection of research data until this form has been approved

SECTION B  To be completed by the Research Institute Ethics Committee:

Comments:

Approved
Signature Chair of Research Institute Ethics Committee:

Date:

This form should then be filed with the RS1 form

If in the judgement of the committee there are significant ethical issues for which there is not agreed practice then further ethical consideration is required before approval can be given and the proposal with the committees comments should be forwarded to the secretary of the UREC for consideration.

There are significant ethical issues which require further guidance

Signature Chair of Research Institute Ethics Committee:

Date:

This form together with the recommendation and a copy of the research proposal should then be submitted to the University Research Ethics Committee
Appendix 15: Main study - Quick Placement Test (OUP, 2001)
Appendix 16: Main study - INTEGRATED EAP TEST

NAME

____________________________________

ID (e.g. 98764321)  __________________________________

You are going to do a test in which you will listen to a short lecture on the topic of “culture shock”. You should make notes on what you hear. After that, you will write a summary of the lecture. The title of the summary is:

‘What is culture shock and how can we reduce the effects of culture shock?’

The lecture will be presented in two sections, starting with an audio only section first then a video in the second part. You should take notes on each section in the spaces provided below.

You will hear the lecture twice. You will have 30 seconds to check your answers before the second playback.

You should write your text on the paper provided and you should include all the information from the lecture that you think is relevant.

The lecture will take about 35 mins in total. You then have 45 minutes to write your summary.
LECTURE – Please take notes here.

Section 1 – Audio playback

Make notes below. Write down as much information as you can.

These notes will NOT be assessed as part of the listening test but you will need the information for the speaking test later.

Note: this section is audio only - you will not see anything in this section.
Section 2 – Video playback

Make notes below. Write down as much information as you can.

These notes will NOT be assessed as part of the listening test but you will need the information for the speaking test later.
WRITING TASK

Using the information you have written down from the lecture, you are now going to write a summary for a non-specialist audience of at least 350 words entitled:

What is culture shock and how can you reduce the effects of culture shock?

You will be marked on the content, language and structure of your summary.

You may make notes below but please cross out any notes that you do not want to be marked.

You have 45 minutes to write the summary. Write your summary below.

_________________________________________________________________________

_________________________________________________________________________

_________________________________________________________________________

_________________________________________________________________________

_________________________________________________________________________

_________________________________________________________________________

_________________________________________________________________________

_________________________________________________________________________

_________________________________________________________________________

_________________________________________________________________________

_________________________________________________________________________
Appendix 17: Main study - PowerPoint Lecture Slides

Culture shock

- Definition:
The anxiety that results from losing familiar signs and symbols of social intercourse.
  (Oberg, 1954)

Symptoms of culture shock

- If you suffer from culture shock, you may experience:
  - depression
  - loneliness
  - insecurity
  - a lack of confidence
  - an inability to solve easy problems

Stages of culture shock

How long will culture shock last?

- It depends on ...
  - you
  - your attitude
  - your situation
  - differences between the two cultures
  - steps you take to minimise culture shock
Dealing with culture shock

Understanding of 'culture'

Reduction of the effects of culture shock

Integration

Preparation

Awareness

Linguistic mastery

REMEMBER:
Culture shock is a natural phenomenon and you are not the only one to experience it!
Appendix 18: Main study - Lecture Transcript: Culture Shock

Hello, everybody and welcome to today’s lecture on culture shock. Well, first of all, I’m going to be defining what we mean by culture shock; we’ll then look at the symptoms of culture shock, that is, what we are going to experience if we are going through culture shock; we’ll then look at the five stages of culture shock and after that, I’ll talk to you about how long culture shock might last; and then, finally, we’ll look at some ways that we can minimise culture shock.

(0:00-0:33 Introduction) (93 words)

So, what do we mean by culture shock? Well, sometimes, when we go to a new country, we find that things are done differently in the new country to how things are done in our own country home and that can leave us feeling a little bit confused or a little bit insecure because we don’t know how things work in the new culture. Erm, and this anxiety that we feel when we don’t know how society works is what Dr Kalervo Oberg, that’s O-B-E-R-G, defined as culture shock some 50 or 60 years ago. Another researcher in intercultural competence and intercultural communication is a gentleman called Dr Hofstede, that’s H-O-F-S-T-E-D-E, erm what he said was that culture shock is like being an infant again, so like being a child about one or two years old, where everything that we once knew about the world and how it worked, has changed and again, this can leave us feeling quite insecure and unsure of ourselves and this is what erm, Dr Hofstede defines as culture shock.

So, when we’re going through the culture shock, what are the symptoms? What are we going to feel? Well, we ... we’re probably going to feel depressed, unfortunately, erm, simply because we’re in this situation where we don’t know how things work. We could be in a new country on our own so maybe we don’t have our friends or family nearby so we may start to feel a bit depressed, we may be a bit lonely or we may feel sad. Erm, if we’re feeling lonely and sad and depressed and especially if we don’t know how things work in the new culture, then that
can leave us feeling rather unsure or insecure as I mentioned earlier. And that may ... may, or probably will, mean that we are lacking in confidence, that we suddenly aren't very confident any more, even if we are very confident in our normal, home culture, we may become quite shy or withdrawn in the new culture. Erm, and we may also find that we suddenly become unable to solve problems, so what would normally, in normal circumstances be just a small problem, suddenly becomes quite a major – quite an important - problem and we can't deal with it.

So those are some of the symptoms of culture shock. Now that sounds ... probably sounds quite negative but it’s ... it’s not, it’s actually important to remember that it is perfectly normal and it’s perfectly natural and everybody goes through those stages.

So what stages are we talking about? Well, there are five stages of culture shock and when we talk about culture shock, we usually refer to culture shock as ‘W’ curve or as a ‘U’ curve. The ‘U’ curve, so, in other words, the first half of the ‘W’ refers to the time spent in the new country, and if we see it as a ‘W’, then the first half is the new country and the second half of the ‘W’ is when we return home. And in each case, we have five stages.

Now the first stage, erm, in the ‘U’ curve is the Honeymoon Stage and that’s up at the top left of the first ‘U’, if you like, in a ‘W’. And, in the Honeymoon Stage, that’s when we’ve just arrived and we’re very excited to be there and very happy to be there and everything is new and exciting and wonderful. And also, the memories of our friends and family are still quite close to our heart, so, at this point, we feel quite happy.

However, this will soon turn into the Disintegration Stage, or the Distress Stage, and this stage is down at the bottom of the ‘U’ curve, erm ... and at this point, we feel rather unhappy, erm, because we’ve started to discover that there are cultural differences, that things a ... are done differently and we don’t necessarily like it, we don’t understand it, maybe we’re not sure of
why things are done differently. So, in the Disintegration Stage, or in the Distress Stage, as it’s called, we might ... well, we will feel quite unhappy.

But this soon changes into the Reintegration Stage – that’s about half way up on the righthand side of the ‘U’, and in the Reintegration Stage, we get you ... we get to the point where we want to be together with people from our own country and our own culture, and we complain about the new country and the new culture and the people and so on. Erm, now that really does sound very negative but, again, it’s perfectly normal and it’s perfectly natural, and it’s actually positive in that it ... it means you’ve realised what you value about your own culture. So, you’ve noticed these differences and you’ve suddenly decided this is much nicer in my culture; I like it in my culture so you’ve discovered your own positive aspects.

After that, this becomes the ... the fourth stage then is the Autonomy Stage. And in the Autonomy Stage, you realise that things are done differently. You may not necessarily do things in the same way and you may not like the way things are done but you’re beginning to get to grips with the culture.

And, once you’ve done that, once you’ve got to grips with the culture, you move in to the final stage, which is the Independence Stage. And at this point, you’re perfectly happy in the new culture, you can live with it, you can deal with it, you may not do things in the same way as things are done in the new culture but it’s not a problem for you.

So that’s the ‘U’ curve, and if we move on to the “W” curve then, as I said before, this is when we go back home and we go through exactly all the same stages all over again, simply because, while you’ve been away, you’ve changed. You’ve changed as a person: you’ve adapted to a new culture and a new way of life and you have to go through all of those again when you go back home to readapt to your own culture. So those are the five stages of culture shock.
And you’re probably wondering at this time how long this is going to last so let’s look at that now.

Length 6:14 (1108 words: 177.75 wpm)

Well, in short, it depends on a number of factors – there is no quick and simple answer. First of all, it depends on you as an individual: how are you ... how do you as an individual react to the new culture? It depends on your attitude to the new culture, so, do you want to integrate? Do you want to learn about the new culture? Or do you prefer to just stay back and be, ern, ... if you like, remain in your own culture? If that’s the case, then it’s going to take longer than if you want to integrate into the new culture and if you do everything you can to integrate into the new culture. It also depends on your situation, so on the situation that you find yourself in. So, let’s imagine that you’ve gone to the new country ... country, ern, you’ve made lots of new friends, you’ve got, ern, great flatmates, a nice place to live, everything is going well with your studies or with your work or whatever it is you’re doing there, then obviously you’re going to have a good time and culture shock will not last as long as if you go and you’re having a bad time, so, if you don’t have any friends or you’ve problems with your landlord or maybe you’re struggling with your studies or something, then, obviously, it’s going to put you in an awkward position and culture shock will last a little bit longer, so it’s important to be aware of that. But it can also depend on, ern, the differences between the two cultures so if you go to a culture that is very similar to your own culture then you’ll get through it quite quickly whereas if the new culture is quite different to your own culture, ern, then it will take longer to get past culture shock. Erm, finally, it also depends on the steps you take to minimise culture shock so there are various things we can do to ... to help us get over culture shock and that’s what we’re going to look at now.
So there are really 5 things we can do to deal with culture shock or to reduce the effects of culture shock. The first thing to do though is to get an understanding of culture: what do we mean by culture? What is culture? I don’t mean culture as in the Arts, music, theatre and so on. What I mean is the culture which is the ways that we behave as individuals and we learn those ways ... we learn our culture from the people around us, from the society that we’re brought up in so it’s not something we’re born with; it’s not that you’re born with the culture of your country and I’m born with the culture of my country but it’s something that we grow up with and we develop and learn. But once we’ve developed that culture, once we’ve learnt that culture, it does become quite embedded inside us, erm, and then it can affect us quite se ... severely or quite seriously when suddenly our new culture is different; hence why we might ... might ... may feel depressed. Erm, but once we’ve got a general understanding of culture, we can start thinking about where it is we’re going and the first thing to do is to prepare yourself well before you go, so find out as much as you can about the new country, about the new culture, about the people, about how life works there: simple things like when do the shops open and close? How do people address each other? All these kinds of things that you can find out before you go will help you to feel better and to get past culture shock more quickly.

Then you need to learn the language: if you speak the language not necessarily absolutely fluently but if you can at least communicate in the language, then that will reduce the misunderstandings and the frustration caused by not understanding the language so some e ... some level of understanding and ability in the language is very useful.

The fourth step then is to be aware of culture shock: be aware that it exists and be aware of the symptoms, be aware of what you’re going to go through but also, try to be aware of your own culture so what ... what is your own culture like? If you’re aware of your culture, then it will help you to contrast w ... with the differences in the new culture.
And finally, erm, integration. Erm, you need to integrate into the new culture. The more you can integrate, the better you will develop an understanding of the culture and the sooner you’ll get over culture shock but also, if you integrate, it helps you to improve your language as well and it gets you new friends, so if you like a sport, or you have a hobby, find a club where you can practise your sport or your hobby and if you don’t have one, get one! You know, maybe take up something new, something you’ve always wanted to do. Find a club where you can practise a new hobby and get to know new people, get to practise speaking the language and also learn more about the culture and the way of life in that country.

So those are the five stages or the five ways, if you like, that we can deal with culture shock. So that’s really everything that I want to say to you today. It just remains for me to say, or to remind you, that culture shock is perfectly normal. It is perfectly natural. It is a positive learning experience for the long-term and, if you’re sitting in a room full of foreign students, I can almost guarantee that you won’t be the only person going through culture shock because everybody will be going through culture shock, maybe at different stages but everyone will go through culture shock when they go to a new country. Thanks.

Length: 5:46 (1048 words: 182.26 wpm)

Ok, so that’s everything that I’d like to say to you today. Thank you very much for your attention. Erm, thank you for listening and if you have any questions, I’d be happy to answer them.

(Length: 0:09: Conclusion) (39 words)
## Appendix 19: Main study - Text Inspector analysis of lecture language and slides

<table>
<thead>
<tr>
<th>Operation</th>
<th>Data</th>
<th>Lecture Part 1</th>
<th>Lecture Part 2</th>
<th>Lecture Slides Part 1</th>
<th>Lecture Slides Part 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics</td>
<td>Sentence count</td>
<td>34</td>
<td>31</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Statistics</td>
<td>Token count</td>
<td>1108</td>
<td>1048</td>
<td>54</td>
<td>61</td>
</tr>
<tr>
<td>Statistics</td>
<td>Type count</td>
<td>290</td>
<td>277</td>
<td>42</td>
<td>44</td>
</tr>
<tr>
<td>Statistics</td>
<td>Type/token ratio</td>
<td>0.26</td>
<td>0.26</td>
<td>0.78</td>
<td>0.72</td>
</tr>
<tr>
<td>Statistics</td>
<td>Avge syllables per sentence</td>
<td>43.97</td>
<td>45.19</td>
<td>54.5</td>
<td>36.67</td>
</tr>
<tr>
<td>Statistics</td>
<td>Avge syllables per word</td>
<td>1.35</td>
<td>1.34</td>
<td>2.02</td>
<td>1.8</td>
</tr>
<tr>
<td>Statistics</td>
<td>Flesch Reading Ease</td>
<td>59.61</td>
<td>59.43</td>
<td>8.66</td>
<td>33.64</td>
</tr>
<tr>
<td>Statistics</td>
<td>Flesch-Kincaid Grade</td>
<td>13.04</td>
<td>13.37</td>
<td>18.76</td>
<td>13.62</td>
</tr>
<tr>
<td>Statistics</td>
<td>Average Sentence Length</td>
<td>32.59</td>
<td>33.81</td>
<td>27</td>
<td>20.33</td>
</tr>
<tr>
<td>Lexical Diversity</td>
<td>Lexical diversity (VOCD)</td>
<td>71.86</td>
<td>62.8</td>
<td>71.16</td>
<td>54.36</td>
</tr>
<tr>
<td>Lexical Diversity</td>
<td>Lexical diversity (MTLD)</td>
<td>39.76</td>
<td>40.25</td>
<td>44.27</td>
<td>59.14</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>A1 type %</td>
<td>45.97</td>
<td>52.45</td>
<td>28.57</td>
<td>51.16</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>A2 type %</td>
<td>20.47</td>
<td>16.43</td>
<td>9.52</td>
<td>2.33</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>B1 type %</td>
<td>10.07</td>
<td>13.99</td>
<td>19.05</td>
<td>25.58</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>B2 type %</td>
<td>8.72</td>
<td>8.74</td>
<td>16.67</td>
<td>0</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>C1 type %</td>
<td>2.01</td>
<td>1.4</td>
<td>4.76</td>
<td>11.63</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>C2 type %</td>
<td>0.67</td>
<td>0</td>
<td>2.38</td>
<td>0</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>Known Words type %</td>
<td>0.67</td>
<td>0.7</td>
<td>4.76</td>
<td>4.65</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>Unlisted type %</td>
<td>11.41</td>
<td>6.29</td>
<td>14.29</td>
<td>4.65</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>A1 token %</td>
<td>64.09</td>
<td>67.71</td>
<td>28.3</td>
<td>48.28</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>A2 token %</td>
<td>13.76</td>
<td>11.54</td>
<td>15.09</td>
<td>1.72</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>B1 token %</td>
<td>4.71</td>
<td>5.06</td>
<td>15.09</td>
<td>18.97</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>B2 token %</td>
<td>3.68</td>
<td>3.44</td>
<td>13.21</td>
<td>0</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>C1 token %</td>
<td>0.85</td>
<td>0.91</td>
<td>3.77</td>
<td>8.62</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>C2 token %</td>
<td>0.19</td>
<td>0</td>
<td>1.89</td>
<td>0</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>Known Words token %</td>
<td>3.3</td>
<td>6.07</td>
<td>11.32</td>
<td>18.97</td>
</tr>
<tr>
<td>Lexis: EVP</td>
<td>Unlisted token %</td>
<td>9.43</td>
<td>5.26</td>
<td>11.32</td>
<td>3.45</td>
</tr>
<tr>
<td>Lexis: AWL</td>
<td>AWL All Types %</td>
<td>5.83</td>
<td>7.48</td>
<td>11.9</td>
<td>15.56</td>
</tr>
<tr>
<td>Lexis: AWL</td>
<td>AWL All Tokens %</td>
<td>4.98</td>
<td>8.2</td>
<td>12.96</td>
<td>18.03</td>
</tr>
</tbody>
</table>
Appendix 20: Main study - Text Inspector CEFR analysis

Part one

This tool analyses your text according to the English Vocabulary Profile from Cambridge University Press. Click here for an explanation. If you want to correct any analysis, go to insert below, change the tag, then click Update.

Summary

<table>
<thead>
<tr>
<th>Word List</th>
<th>Types</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>137 (45.97%)</td>
<td>606 (44.09%)</td>
</tr>
<tr>
<td>A2</td>
<td>41 (20.47%)</td>
<td>146 (12.76%)</td>
</tr>
<tr>
<td>B1</td>
<td>30 (10.07%)</td>
<td>50 (4.71%)</td>
</tr>
<tr>
<td>B2</td>
<td>26 (8.72%)</td>
<td>59 (5.68%)</td>
</tr>
<tr>
<td>C1</td>
<td>4 (2.01%)</td>
<td>9 (0.85%)</td>
</tr>
<tr>
<td>C2</td>
<td>2 (0.67%)</td>
<td>2 (0.19%)</td>
</tr>
<tr>
<td>Known Words</td>
<td>2 (0.67%)</td>
<td>55 (3.50%)</td>
</tr>
<tr>
<td>Unlisted</td>
<td>34 (11.41%)</td>
<td>100 (9.43%)</td>
</tr>
</tbody>
</table>

Statistics and details

| A1 | 137 types / 606 tokens | 45.97% / 64.09% (Show words) |
| A2 | 41 types / 146 tokens  | 20.47% / 13.76% (Show words) |
| B1 | 30 types / 50 tokens   | 10.07% / 4.71% (Show words)  |
| B2 | 26 types / 59 tokens   | 8.72% / 5.68% (Show words)   |
| C1 | 4 types / 9 tokens     | 2.01% / 0.83% (Show words)   |
| C2 | 2 types / 2 tokens     | 0.67% / 0.23% (Show words)   |
| Known Words | 2 types / 55 tokens | 0.67% / 2.00% (Show words) |
| Unlisted  | 34 types / 100 tokens | 11.41% / 9.42% (Show words) |
so A A what A do A we A mean A by A culture shock well A sometimes A when A we A go A to A
A country A we A find A that A things A are A done A differently A in A the A new A country A our A own A home A and A that A
A can A leave A us A feeling A a A little A bit A confused A or A a A little A bit A insecure A because A we A don't A know A how A things A work A in A the A new A cultural A
A A A A this A anxiety A that A when A we A don't A know A how A things A work A in A the A new A culture A A A A how A A A A
60 years A ago A another A researcher B called A Hofstede that A's A ho A f s A t e d e r m A
like A being A a A child A about A one A or A two A years A old A where A A everything A
that A we A once A knew A about A the A A world A and A how A it A worked A has A
changed A and A again A this A can A leave A us A feeling A quite A insecure A and A unsure A of A ourselves A and A this A is A what A erm A dr A Hofstede defines A as A culture shock A A A A
when A we're A going A to A feel A well A A we A we're A probably A going A to A feel A depressed A unfortunately A
erm simply A because A we're A in A this A situation B where A we A don't A know A how A
A things A work A in A the A new A country A on A our A own A A A
A may A or A not A have A A our A old A A family A nearby A so A we A may A or A not A start A
A to A feel A A bit A depressed A we A may A or A not A have A A small A or A large A or A
A we A may A or A not A feel A A little A insecure A especially A if A we A don’t A know A how A things A work A in A the A new A culture A then A that A
A can A leave A us A feeling A rather A unsure A or A insecure A as A I A mentioned A earlier A
and A that A may A or A not A will A or A probably A will A make A A mean A that A we A are A or A
A confidence A that A we A are A suddenly A aren't A very A confident B any A more A even A much A if A we A are A
A very A or A not A withdrawn C in A the A new A culture A and A we A may A or A not A find A that A
A we A suddenly A become A usable A to A solve A problems A to A what A would A normally B
in A the A normal A circumstances B be A just A a A small A problem A suddenly B becomes A quite A
A a A major B quite A an A important A problem B and A we A can't A deal B with A it A A to A
A those A are A some A of A the A A symptoms B of A culture shock now A that A we A probably A
A sounds A quite A negative A but A it A's A it A's A not A it A's A actually A important B to A
A remember A at A that A it A is A perfectly A normal A and A it A's A perfectly B normal B and A
A everybody A gets A through B those A A stages A so A what A A stages A are A we A talking A about B
A well A there A are A five A stages A of A culture shock and A when A we A talk A about A
A culture shock we A usually A refer A to A culture shock as A well A curve B or A as A at A B curve C
A the A u C curve D so A in A the A other A words A the A first A half A of A the A A A
A and A then A we A see A it A A as A A will A refers A to A
A the A first A half A of A the A new A country A and A that A if A we A see A it A as A a A A
A and A then A the A first A half A of A the A new A country A and A we A second A half A of A
A the A second A half A of A the A new A country A and A we A return A home A and A in A each A case A we A have A five A
A stages A so A in A the A first A stage A culture shock in A the A u C curve D is A in A the A A
A and A we're A very A excited A to A be A there A and A very A happy A to A be A A
A and A everything A is A new A and A exciting A and A wonderful A and A also A the A A
Part two
Input
Appendix 21: Main study - Text Inspector AWL analyses

Part one

### Summary

<table>
<thead>
<tr>
<th>Word List</th>
<th>Types</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWL Total</td>
<td>18 (5.83%)</td>
<td>51 (4.99%)</td>
</tr>
<tr>
<td>AWL 1</td>
<td>4 (1.29%)</td>
<td>4 (0.39%)</td>
</tr>
<tr>
<td>AWL 2</td>
<td>2 (0.65%)</td>
<td>3 (0.29%)</td>
</tr>
<tr>
<td>AWL 3</td>
<td>1 (0.32%)</td>
<td>1 (0.10%)</td>
</tr>
<tr>
<td>AWL 4</td>
<td>1 (0.32%)</td>
<td>1 (0.10%)</td>
</tr>
<tr>
<td>AWL 5</td>
<td>1 (0.32%)</td>
<td>1 (0.10%)</td>
</tr>
<tr>
<td>AWL 6</td>
<td>1 (0.32%)</td>
<td>1 (0.10%)</td>
</tr>
<tr>
<td>AWL 7</td>
<td>1 (0.32%)</td>
<td>1 (0.10%)</td>
</tr>
<tr>
<td>AWL 8</td>
<td>1 (0.32%)</td>
<td>1 (0.10%)</td>
</tr>
<tr>
<td>AWL 9</td>
<td>1 (0.32%)</td>
<td>1 (0.10%)</td>
</tr>
<tr>
<td>K1</td>
<td>6 (1.94%)</td>
<td>12 (1.17%)</td>
</tr>
<tr>
<td>K2</td>
<td>4 (1.29%)</td>
<td>5 (0.49%)</td>
</tr>
<tr>
<td>K3</td>
<td>6 (1.94%)</td>
<td>7 (0.68%)</td>
</tr>
<tr>
<td>K4</td>
<td>2 (0.62%)</td>
<td>8 (0.70%)</td>
</tr>
<tr>
<td>Unlisted</td>
<td>270 (87.38%)</td>
<td>941 (91.89%)</td>
</tr>
</tbody>
</table>

### Statistics and details

<table>
<thead>
<tr>
<th>Word List</th>
<th>Types</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWL 1</td>
<td>4 types / 4 tokens</td>
<td>1.29% / 0.29% (Show words)</td>
</tr>
<tr>
<td>AWL 2</td>
<td>8 types / 18 tokens</td>
<td>5.69% / 3.71% (Show words)</td>
</tr>
<tr>
<td>AWL 3</td>
<td>2 types / 2 tokens</td>
<td>0.64% / 0.29% (Show words)</td>
</tr>
<tr>
<td>AWL 4</td>
<td>1 types / 1 token</td>
<td>0.22% / 0.10% (Show words)</td>
</tr>
<tr>
<td>AWL 5</td>
<td>1 types / 1 token</td>
<td>0.31% / 0.10% (Show words)</td>
</tr>
<tr>
<td>AWL 6</td>
<td>1 types / 1 token</td>
<td>0.32% / 0.10% (Show words)</td>
</tr>
<tr>
<td>AWL 7</td>
<td>1 types / 1 token</td>
<td>0.32% / 0.10% (Show words)</td>
</tr>
<tr>
<td>AWL 8</td>
<td>1 types / 2 tokens</td>
<td>0.32% / 0.23% (Show words)</td>
</tr>
<tr>
<td>AWL 9</td>
<td>1 types / 2 tokens</td>
<td>0.32% / 0.23% (Show words)</td>
</tr>
<tr>
<td>K1</td>
<td>6 types / 12 tokens</td>
<td>1.94% / 1.17% (Show words)</td>
</tr>
<tr>
<td>K2</td>
<td>4 types / 5 tokens</td>
<td>1.29% / 0.49% (Show words)</td>
</tr>
<tr>
<td>K3</td>
<td>6 types / 7 tokens</td>
<td>1.94% / 0.62% (Show words)</td>
</tr>
<tr>
<td>K4</td>
<td>5 types / 8 tokens</td>
<td>1.63% / 0.78% (Show words)</td>
</tr>
<tr>
<td>Unlisted</td>
<td>270 types / 941 tokens</td>
<td>87.38% / 91.89% (Show words)</td>
</tr>
</tbody>
</table>
Part two
Appendix 22: Main study - Feedback questionnaire

This short questionnaire is about your experience of taking the language test.

Your answers will help me to see how the kind of input (video vs audio) affected your test performance.

The questionnaire should not take more than 2-3 minutes to complete. If you do not understand any of the questions, please ask the tutor for help. When you finish, please give your questionnaire to the tutor before you leave the room.

Please note: your answers to this questionnaire have NO impact on your test results so please feel free to give honest answers!!

<table>
<thead>
<tr>
<th>Identification no.</th>
<th>___________________________</th>
<th>Name (optional) ______________________________</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nationality:</td>
<td>___________________________</td>
<td>Course of study ______________________________</td>
</tr>
<tr>
<td>Age (please tick):</td>
<td>18-21 ☐</td>
<td></td>
</tr>
<tr>
<td></td>
<td>22-25 ☐</td>
<td></td>
</tr>
<tr>
<td></td>
<td>26-30 ☐</td>
<td></td>
</tr>
<tr>
<td></td>
<td>31-40 ☐</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;40 ☐</td>
<td></td>
</tr>
<tr>
<td>Gender (please tick):</td>
<td>Male ☐ Female ☐</td>
<td></td>
</tr>
<tr>
<td>Language level:</td>
<td>A1 ☐</td>
<td>B2 ☐</td>
</tr>
<tr>
<td></td>
<td>A2 ☐</td>
<td>C1 ☐</td>
</tr>
<tr>
<td></td>
<td>B1 ☐</td>
<td>C2 ☐</td>
</tr>
</tbody>
</table>
Think about your performance while listening to the lecture. Do you think your results will demonstrate your listening ability accurately?

- I understood much more than I normally would on a listening test
- I understood more than I normally would on a listening test
- I did as well as I normally would do on a listening test
- I did worse than I normally would do on a listening test
- I did much worse than I normally would do on a listening test

On the whole, how difficult did you find the writing test?

- Very difficult
- Quite difficult but I could cope with it
- Quite easy
- Very easy

How much did you know about the lecture topic before the test?

- I knew a lot about the topic before
- I knew a little about the topic before
- I knew nothing about the topic before
Do you think the topic of the test had a positive or negative impact on how well you performed?

- I think the topic had a positive impact on my performance
- I don’t think the topic had any impact on my performance
- I think the topic had a negative impact on my performance
- I don’t know whether the topic had any impact on my performance

Think of the two sections of the lecture. Which part do you think you performed best on?

Section 1 – audio only  □  Section 2 – video  □

Why do you think you performed best on this section?
___________________________________________________________________________
___________________________________________________________________________

Think of the different sections of the lecture. Which part do you think you performed worst on?

Section 1 – audio only  □  Section 2 – video  □

Why do you think you performed worst on this section?
___________________________________________________________________________
___________________________________________________________________________
Think of the video you saw. Do you feel any of the following aspects had a positive effect on your understanding of the lecture? (Tick all that apply)

- Seeing the speaker
- Seeing some text related to the content
- Seeing a diagram related to the content

Do you think any of the following had a negative effect on how well you understood the lecture? (please tick all that apply)

- The sound quality
- The video quality
- The number of times you heard the lecture
- The amount of information you had to write down

Any other factors or any comments (please write your comments below):

_____________________________________________________________________
_____________________________________________________________________

On the whole, how difficult did you find the summary based on the lecture topic?

- Very difficult
- Quite difficult
- Quite easy
- Very easy
Did anything affect (positively or negatively) how much you could write in the essay? Please give brief details:

_____________________________________________________________________

_____________________________________________________________________

_____________________________________________________________________

THANK YOU FOR TAKING THE TIME TO COMPLETE THIS QUESTIONNAIRE AND FOR BEING INVOLVED IN THE STUDY.
YOUR HELP AND FEEDBACK ARE VERY MUCH APPRECIATED. ☺
Appendix 23: Main study - Example of a coded paper

What we mean by culture shock? Sometimes, for example when we travel to another country, we can find differences between how things are done here and in our home. We can feel a little bit confused or insecure because we don’t know how things work in a new culture. Dr. Oberd says that this anxiety that we feel when we don’t know how society works is a cultural shock. Dr. Holsted claims that cultural shock is like being a child of one or two years old when everything we know about the world has changed. This can make us feel insecure about ourselves.

There are many symptoms of culture shock. For instance, you feel depressed, lonely, sad, unsure, insecure, unconfident, shy, unable to solve problems. But there is nothing special or wrong in this situation.

There are five stages of culture shock. The first one is “honeymoon”. You feel excited, happy, everything seems exciting and wonderful. The second one is “disintegration” (distress). You feel unhappy, you probably started to discover that many things differ. The third stage is “reintegration”. You complain about everything: people, new country; you also realize things that you found about your own culture. The fourth stage is “autonomy”. You realize that you don’t have to do things in the same way. Finally, the fifth stage is “independence”. Now you can deal with the new culture, it’s not a problem for you anymore.

There are five things that we can do to reduce culture shock:

- try to understand of “culture”, of ways that we behave
- prepare yourself
- learn the language
- to be aware of culture shock
- integrate

Culture shock is a positive learning experience, a normal part of human life and everyone once will go through it.
Appendix 24: Main study - Culture shock lecture: expert panel propositions

5 sub-headings:

- Definition
- Symptoms
- 5 stages of culture shock
- Duration of culture shock
- Minimising culture shock

Definition: feelings of anxiety and insecurity caused by new culture Oberg; (defined 50/60 years ago); Hofstede; like being an infant again; leading to being insecure; unsure; (everything you once knew has changed);

Symptoms – feeling depressed; don’t know how things in the new culture work; lonely and sad (no family and friends); insecurity; lacking and confidence; shy/withdrawn in new culture; difficulty solving problems;

Normal and natural; applies to everyone;

Five stages:

W or U curve; U: time in new country; W: adds return home;

1. Honeymoon stage: top left of U curve; just arrived; excited; happy; everything is new; close memories of loved ones back home;
2. Disintegration (or Distress) stage; bottom of the U curve; (you realise) things done differently in new culture / (notice cultural differences); don’t like that things are different; unhappy;
3. Re-integration stage: half way up right hand side; want to be with / get together with people from old culture; criticise new culture; positive as you realise what you value from your culture;
4. Autonomy stage: (position on U curve not mentioned); realise things are done differently / realise differences; get to grips with new culture / coming to terms with culture; might still not like it;
5. Independence stage: happy with / accept new culture; no problem with it / can live with it;

W curve: return to old country; go through everything / stages of culture shock again when you re-adapt to your old culture;
Duration:

- depends on individual; engage with new culture or try to hold on to old one;
- attitude: takes longer if you don't try to integrate;
- situation: if all is positive, then culture shock doesn't last as long / going well? = shorter; lasts longer if things are not good / bad? = longer;
- degree of difference between two cultures: if similar cultures, you get through it quickly; culture shock lasts longer if cultures are very different;
- steps you take to minimise it

Minimising / reducing culture shock:

- 5 steps / things to reduce effects:
  - Understand what culture is; how we behave as individuals; depends on society around us; becomes / it is embedded; affects us;
  - Be well prepared; (find out about the new culture / find out how life works before going);
  - Learn the language; (to reduce misunderstandings and frustrations);
  - Be aware of culture shock; and your own culture;
  - Integration: develop (better) understanding of the new culture; helps with language; helps with friends / make new friends; find a club or hobby;

Culture shock: normal; natural; positive (learning / long-term) experience;
### Appendix 25: Main study - Propositions reproduced from part one: both groups and totals

<table>
<thead>
<tr>
<th>Propositions - part one</th>
<th>Definition:</th>
<th>3</th>
<th>3</th>
<th>6</th>
<th>feelings of anxiety and insecurity caused by new culture</th>
<th>5</th>
<th>9</th>
<th>14</th>
<th>Oberg;</th>
<th>2</th>
<th>2</th>
<th>4</th>
<th>Hofstede;</th>
<th>7</th>
<th>3</th>
<th>10</th>
<th>like being an infant again;</th>
<th>6</th>
<th>4</th>
<th>10</th>
<th>leading to being insecure;</th>
<th>3</th>
<th>3</th>
<th>6</th>
<th>unsure;</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symptoms</td>
<td>defination</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>feeling depressed;</td>
<td>11</td>
<td>14</td>
<td>25</td>
<td>don’t know how things in the new culture work;</td>
<td>6</td>
<td>3</td>
<td>9</td>
<td>lonely</td>
<td>6</td>
<td>12</td>
<td>18</td>
<td>sad</td>
<td>6</td>
<td>1</td>
<td>7</td>
<td>(no family and friends);</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>insecurity;</td>
<td>4</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>lacking and confidence;</td>
<td></td>
<td></td>
<td></td>
<td>shy/withdrawn in new culture;</td>
<td>6</td>
<td>1</td>
<td>7</td>
<td>difficulty solving problems;</td>
<td>7</td>
<td>9</td>
<td>16</td>
<td>Normal</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>natural;</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>applies to everyone;</td>
<td></td>
<td></td>
<td></td>
<td>Five stages:</td>
<td>9</td>
<td>7</td>
<td>16</td>
<td>W curve</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>or U curve;</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1. Honeymoon stage:</td>
<td>11</td>
<td>7</td>
<td>18</td>
<td>U: time in new country;</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>W: adds return home;</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>top left of U curve;</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>just arrived;</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>happy;</td>
<td>8</td>
<td>5</td>
<td>13</td>
<td>everything is new;</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>close memories of loved ones back home;</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>excited;</td>
<td>9</td>
<td>2</td>
<td>11</td>
<td>2. Disintegration (or Distress) stage;</td>
<td>9</td>
<td>7</td>
<td>16</td>
<td>bottom of the U curve;</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(you realise) things done differently in new culture / (notice cultural differences);</td>
<td>8</td>
<td>7</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>unhappy;</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>don’t like that things are different;</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>unhappy;</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>3. Re-integration stage:</td>
<td>8</td>
<td>7</td>
<td>15</td>
<td>half way up right hand side;</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>want to be with / get together with people from old culture;</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>criticise new culture;</td>
<td>5</td>
<td>4</td>
<td>9</td>
<td>positive as you realise what you value from your culture;</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4. Autonomy stage:</td>
<td>9</td>
<td>5</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>realise things are done differently / realise differences;</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>get to grips with new culture / coming to terms with culture;</td>
<td>8</td>
<td>5</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>might still not like it;</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Independence stage:</td>
<td>10</td>
<td>6</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>happy with / accept new culture;</td>
<td>7</td>
<td>4</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no problem with it / can live with it;</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W curve: return to old country;</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>go through everything / stages of culture shock again when</td>
<td>6</td>
<td>1</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you re-adapt to your old culture;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTALS PART ONE</td>
<td>238</td>
<td>169</td>
<td>407</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix 26: Main study - Propositions reproduced from part two: both groups and totals

<table>
<thead>
<tr>
<th>Culture shock lecture - Expert Panel propositions</th>
<th>TOTAL</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Propositions - part two</strong></td>
<td>AUD FIRST</td>
<td>VID FIRST</td>
</tr>
<tr>
<td>Duration:</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>· depends on individual;</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>engage with new culture or try to hold on to old one;</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>· attitude:</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>takes longer if you don’t try to integrate;</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>· situation:</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>if all is positive, then culture shock doesn’t last as long / going well? = shorter;</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>lasts longer if things are not good / bad? = longer;</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>· degree of difference between two cultures:</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>if similar cultures, you get through it quickly;</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>culture shock lasts longer if cultures are very different;</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>· steps you take to minimise it</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Minimising / reducing culture shock:</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>· 5 steps / things to reduce effects:</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>· Understand what culture is;</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>how we behave as individuals;</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>depends on society around us;</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>becomes / it is embedded; affects us;</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>· Be well prepared;</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>(find out about the new culture / find out how life works before going);</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>· Learn the language;</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>(to reduce misunderstandings and frustrations);</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>· Be aware of culture shock;</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>and your own culture;</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>· Integration:</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>develop (better) understanding of the new culture;</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>helps with language;</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>helps with friends / make new friends;</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>find a club</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>find a hobby;</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>normal;</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>natural;</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>positive (learning / long-term) experience;</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>TOTALS PART TWO</strong></td>
<td>145</td>
<td>86</td>
</tr>
</tbody>
</table>
Reference list


Lin, M.C. (2006). The Effects of Note-taking, Memory and Rate of Presentation on EFL Learners’ Listening Comprehension. EdD Dissertation. La Sierra University. Available at: https://search.proquest.com/openview/89e0f37cc0c41e9cd8f82cd251cf85fa/1?pq-origsite=gscholarandcbl=18750anddiss=y (Accessed: 28 October 2019).


