Application of the Construct of Coherence to Diagnostic Testing in English Medium Instruction in Higher Education

by Pasha Blanda

British Council's Master's Dissertation Awards 2023
Commendation

Student ID : 21048826

Institute of Education

UCL

Final assignment cover sheet 2021-22

Students: please complete the sections below and use this as the first page of your assignment. Do not put your name anywhere on this sheet or in your assignment.

Student number:	21048826
Programme name:	Applied Linguistics
Module name:	Dissertation Applied Linguistics and TESOL
Assignment title:	Application of the Construct of Coherence to Diagnostic Testing in English Medium Instruction in Higher Education
Word count:	15, 456
Date of submission:	01/09/2022

Declaration

By submitting your work, you are confirming that you have read and understood the guidelines on submissions, plagiarism, and late submission / word count penalties in the programme handbook and UCL Academic Handbook.

Student ID : 21048826

All summative assessments submitted will be considered as exemplar material for other students to show what type of assessment is expected. If you do not want your assessment to be included as exemplar material, you must inform the Module Administrator. To help us achieve this, please submit your electronic copy in Microsoft Word format. It will be anonymised before being shared.

Student ID : 21048826

Application of the Construct of Coherence to Diagnostic Testing in English Medium Instruction in Higher Education

MA Applied Linguistics

Student Number: 21048826

Word Count: 15, 456

Date of Submission: 01/09/2022

Student ID : 21048826

Declaration

By submitting my work, I am confirming that I have read and understood the guidelines on submissions, plagiarism, and late submission / word count penalties in the programme handbook and UCL Academic Handbook.

This report/dissertation may be made available to the general public for borrowing, photocopying or consultation without the prior consent of the author.

Acknowledgements

I would like to express my deep gratitude to my supervisor Dr. Talia Isaacs for her inexhaustible patience, enthusiastic encouragement and invaluable insight throughout the process of writing this dissertation. Without her instruction and conversation, this dissertation would not be completed in due course.

Thanks go to all the participants in this study for their time and effort. I would also like to thank all the staff with whom I have had the pleasure and privilege to learn from this year.

Finally, my heartfelt thanks go to my family, whose support in all respects has made this research possible.

Student ID : 21048826

Abstract

Instruction of content in higher education settings in L2 English is expanding rapidly globally (Macaro, Curle, Pun, An, & Dearden, 2018; Pecorari & Malmström, 2018).

This dissertation is an exploratory, proof of concept study aimed at considering how a test of coherence of student writing might be operationalised to satisfy needs of practicality and authenticity for English medium instruction. Three gaps in recent research on diagnostic testing are identified: communicative authenticity of tasks based on register, genre and discipline; the effect of source text cohesion in integrated task designs; effects of cultural and social differences between students.

To investigate effects of genre, register and discipline, a novel operationalisation of the construct of coherence for an integrated reading-writing task is used to analyse L2 writing (n = 396) in the in the British Academic Written English corpus (Alsop & Nesi, 2009). To investigate how text cohesion might be applied as an intervention for L2 students, a corpus of academic texts (n = 41) were summarised using an automated method based on 3 machine learning algorithms. Summaries were compared using cohesion based ease of reading measures based on previous research in cognitive psycholinguistics (Crossley, Greenfield, & McNamara, 2008). To investigate social and cultural differences between students, an integrated reading writing task was piloted with four participants, followed by a questionnaire relating to language use and proficiency.

Results suggest that the use of formulaic language may be affected by source use, suggesting further research into the construct of coherence may be warranted. Application of Latent Semantic Analysis to text summarisation showed statistically significant results in sentence level cohesion (p = <.001; d = 0.78). Piloting questionnaires suggest differences between participants in digital media consumption and informal conversation in L2.

Student ID : 21048826

Table of Contents

Student ID : 21048826

List of Tables and Figures

Figures

Tables

Student ID : 21048826

1 Introduction

Global trends in policy relating to the use of English in higher education settings have given rise to newly emergent needs for language instruction (Dafouz & Smit, 2021; Hyland & Jiang, 2018; Macaro et al., 2018). The position of applied linguistics research into English language teaching in higher education responding to these trends is increasingly one of collaboration with disciplines outside of its traditional confines, emphasizing the specificity of registers as they relate to discipline and genre (Airey, 2020; Doiz & Lasagabaster, 2021; Lin & Morrison, 2021). There is a growing body of literature that recognises the significance of these contextual factors for language testing research. Chapelle (2020b; 2015), for example, proposes that advances in technology and the context of globalisation are salient considerations for research in the field. Responding to a similar impetus, X. Xi (2017) proposes that one opportunity which technological advancement presents for language testers is the nature of the constructs that might be investigated. X. Xi (2017)writes “we need to shift some of our attention from discrete local phenomena (e.g., use of articles) to larger linguistic elements (e.g., discourse organization, overall coherence)” (p. 574).

1.2 The Construct of Coherence in the Context of English Medium Instruction

In this essay, the term ‘English medium instruction’ will be used in a broad sense to refer to content teaching in higher education, where language instruction is not the primary focus. In the literature on English medium instruction in higher education, the relative importance of language development compared with content instruction has been subject to considerable discussion (Macaro et al., 2018; Pecorari & Malmström, 2018). Research highlights the seemingly conflicting expectations of language instruction held by policy makers, instructors and students (Airey, 2020; Rose, Curle, Aizawa, & Thompson, 2020).

The present study proposes that coherence is a construct at the heart of our understanding of the relationship of content and language and therefore warrants examination for this context. For the purpose of this investigation, the term ‘coherence’ will refer to a latent construct which can be inferred by test users by measuring test takers’ use of formulaic language, syntax complexity and

inter-sentence cohesion, when test takers are engaged in inferential thinking relating to domain specific knowledge. The contentions in this definition will be explored more fully in the literature review.

1.3 The Purpose Of This Dissertation

This study set out to investigate three research gaps suggested in recent research on post-entry diagnostic language testing in higher education. These recently emergent concerns may be summarised as:

- The authenticity of tasks based on register, genre and discipline (Duan & Shi, 2021; Wang & Xie, 2022);
- The effect of source text cohesion on test taker performance (Bilki & Plakans, 2022; Cai & Chen, 2022);
- The effects of cultural and social differences on the use of academic language, (Cai & Chen, 2022).

To this end, the present research proposes the usefulness of coherence as a construct of interest for diagnostic testing in the context of English medium instruction in internationalised higher education institutions. Theoretical considerations of coherence suggest it is a construct which might encompass the concerns which have emerged in recent, relevant research. The main goal of the current study is to determine the feasibility of a subject specific diagnostic language test and a method of text abbreviation which might be applied as a intervention on the basis of diagnosis.

1.4 The Structure of this Dissertation

The first section of this paper will examine existing research to establish the value of investigating diagnostic testing for this context and use. The review of literature will then consider how the construct of coherence has been operationalised in previous research, with the aim of demonstrating its value and relevance to the context. Having established an operational definition of the construct of coherence, the review will move from discussion around what is being tested, to

how to test for it. The review of literature will conclude with an explicit statement of the research questions.

The second section of the research paper will establish methods through which the research questions will be addressed. This section will include a brief outline of theoretical paradigms, followed by a specification of the methods used to answer the questions. There will also be a discussion detailing the data samples used for this research.

This paper will conclude with the reporting of data and a discussion of the implications of these results. This section will also include a discussion of the limitations of the study and suggest methods through which future research may address as yet unresolved, and newly emergent questions.

2 Literature Review

2.1 English Medium Instruction

The precise relationship between content teaching and language teaching in English medium instruction is too varied globally to prescribe a singular characterisation (Macaro et al., 2018; Pecorari & Malmström, 2018). Nevertheless, research investigating the agency of content instructors in these contexts has argued that instructors are positioned in sufficiently influential roles within institutions to implement effective language learning strategies to aid L2 English learners (Aizawa & Rose, 2019; Lanvers & Hultgren, 2018; Peng & Xie, 2021). However, one consistent theme across relevant research indicates that educators in these roles might limit their linguistic instruction to the introduction of specialised, technical vocabulary because they do not consider language instruction to be within the scope of their responsibility or expertise (Block & Moncada-Comas, 2022; McGrath, Negretti, & Nicholls, 2019). While instruction of technical vocabulary is certainly a fundamental aspect of content specific instruction in L2 English, research has demonstrated that English usage in academic registers to be both highly dependent on genre and discipline, as well as strong predictors of success for students using English as L2 (Casal, Shirai, & Lu, 2022; del Mar Sánchez-Pérez, 2021; Durrant, 2017). In order to evaluate how diagnostic testing might address the disparity between the

language teaching expectations of policy makers and content instructors, the subsequent section of this review will consider research into diagnostic testing in higher education contexts for students who use English as L2.

2.2 Defining Post Entry Diagnostic Language Testing

Two recent, widely cited reviews of significant scope for the context of English medium instruction in higher education only briefly mention the role which language testing plays in these contexts (Macaro et al., 2018; Pecorari & Malmström, 2018). Two relevant issues might be derived from these reviews which relate to diagnostic language testing: the limited success of pre-entry proficiency tests for predicting success in content learning, and the need for the development of appropriate interventions to accommodate the diagnosed needs of L2 English students.

The controversy in this field of research around the appropriate use of pre-entry language proficiency scores is not a recently emergent contention (Dooey and Oliver, 2002; Kokhan, 2012, 2013). This earlier research pointed to the effects of variables besides language proficiency as potential factors contributing to academic success, and highlighting that standardised, pre-entry language proficiency tests had not been validated for use as placement tests. The definition between standardised tests of language proficiency, and placement tests might be extended further to consider how research has defined a difference between placement tests and diagnostic tests.

Read (2008) reports on the development of the Diagnostic English Language Needs Assessment, in which he takes great care to parse the difference between the diagnostic test, standardised tests of language proficiency, and post-entry placement tests. Three salient differences derived from Read (2008) might be summarised as follows:

- A diagnostic test is not used for gatekeeping purposes
- It offers formative feedback to test takers
- It is tied to appropriate interventions to aid learners in improving their language use

Student ID : 21048826

Subsequent research into theoretical approaches to diagnostic language testing, based on the perspectives of various stakeholders has largely supported such a definition (Alderson, Brunfaut, & Harding, 2015; Fenton-Smith & Humphreys, 2015; Phakiti & Isaacs, 2021).

2.3 Recent Research Into Diagnostic Testing

Studies over the past two decades have provided important information on issues in post-enrolment assessment of language proficiency (Bo, Fu, & Lim, 2022; Murray, 2010). Several major issues remain underexplored.

Wang and Xie (2022) provide valuable insight into the discourse competence of undergraduates majoring in business studies, and propose future research should explore differences between disciplines in the extended writing of L2 students in higher education. Investigations into lexical development and the use of formulaic language also emphasize the specificity of discipline and genre, with recent research highlighting the need for future studies to investigate cultural and social differences in language use (Duan & Shi, 2021; Durrant, 2017; Xie & Lei, 2022).

Cai and Chen (2022) investigated the thinking skills of L2 students in higher education and recommend that future research considers variables which might affect the deployment of these skills. Related to these findings, research into test taker strategy responding to integrated task designs suggests sentence level cohesion of source texts may be one such variable. Bilki and Plakans (2022) suggest sentence level cohesion might affect the meaning making strategies of advanced learners, and propose future research apply a quasi-experimental design to investigate reading order effects and cohesion of source texts. Bilki and Plakans (2022) also recommend a statistical analysis of cohesion features in source texts to investigate meaning construction processes adopted by test takers.

Having summarised areas of interest for recent research in diagnostic assessment in the context of English medium instruction at higher education institutions, I will reiterate the purpose of the

present research which has been articulated in the introduction. Three areas remain relatively under-explored based on this recent research:

- The authenticity of tasks based on register, genre and discipline (Duan & Shi, 2021; Wang & Xie, 2022);
- The effect of source text cohesion on test taker performance (Bilki & Plakans, 2022; Duan & Shi, 2021);
- The effects of cultural and social differences on the use of academic language (Cai & Chen, 2022; Duan & Shi, 2021).

The next section of this review will consider how the construct of coherence has been operationalised in past research in order to argue for its relevance for the present study.

3 The Construct of Coherence

3.1 Defining the Construct of Coherence

Before arguing for the value of using the construct in a diagnostic test in English medium instruction, this section will consider how the construct has been operationalised in previous tests of English for academic purposes. Following a review of these operationalisations, I will reiterate a working definition for the purpose of this research.

In the CEFR, IELTS and TOEFL iBT, cohesion and coherence are measured together in the same scales, as a trait in the test takers' writing (Council of Europe, 2020; Educational Testing Service, 2021; The British Council, 2022). All three examples refer to explicit analeptic and proleptic reference as expected evidence of cohesion and coherence through the use of linking cohesive devices, and the degree to which semantic information is repeated either through the same word, or through synonymy. Nevertheless, a point of difference between the scales involves assessing the test takers' ability to infer an appropriate sequence and amount of information to introduce. For this construct, the CEFR and IELTS scales use the term "logic" whilst the TOEFL iBT scale uses "coherence." Whilst the TOEFL iBT limits the representation of this

construct to the portion of the test which measures the test takers' skills in integrating listening, reading and writing, IELTS only mentions the construct for assessing the test takers' writing proficiency. Conversely, the construct of "logic" is spread across various rating scales in the CEFR, but only ever used to aid the assessment of an adjacent construct. In summary, whilst such an operationalisation may be appropriate for standardised tests of language proficiency, recent research suggests a closer investigation into the communicative function of coherence may be more appropriate for diagnostic purposes.

Wang and Xie (2022) operationalise coherence as a subconstruct of discourse competence in academic writing for a diagnostic test. The researchers divide coherence into two aspects. "Global coherence," according to Wang and Xie (2022), refers to overall, conceptual structure, and rhetorical patterning which is appropriate to the communicative purpose of the text. "Local coherence," on the other hand, relates to the ratio of new information introduced in each sentence. Wang and Xie (2022) demonstrated that the sample of L2 learners in their study did indeed encounter problems with both of these types of coherence. As interventions, the researchers recommend a genre-based approach to teaching academic writing. They propose teaching coherence through comparative exercises to encourage students to notice differences. Such salient differences might be deduced from research into formulaic language (Durrant, 2017).

Building further on the communicative emphasis in Wang and Xie (2022), psycholinguistic investigations into the role of memory in text processing suggest a role for text comprehension in a test investigating coherence. Johns (1986) defined coherence as an element of pragmatics based upon its relationship to the prior knowledge of the reader and the use of sociolinguistic conventions. Kintsch's (1988) construction-integration model further formalised the role of prior knowledge in text comprehension strengthening the communicative aspect of the construct. However, the characterisation of the role of memory on text comprehension in the model proposed by Ericsson and Kintsch (1995) came under criticism in subsequent research. Gobet (2000) argued that Kintsch's

(1995) model only applies to conscious, explicit learning processes, as opposed to implicit, automatised learning.

Recent evidence into automatised processes which interweave productive and receptive processes suggest that L2 users occasionally process text through prediction-by-production (Ito, Corley, & Pickering, 2018). In prediction by production, a comprehender covertly imitates what they have already comprehended and infers underlying intentions. From this inference, along with background knowledge, the comprehender predicts upcoming utterances at the level of lexis form and semantics. This form of processing exists alongside prediction-by-association which derives comprehension according to the frequency and proximity of associated words. Recent research using eye-tracking methods is consistent with these findings (Kuperman et al., 2022; Nisbet, Bertram, Erlinghagen, Pieczykolan, & Kuperman, 2022). Consequently, an integrated reading-writing test might arguably be appropriate for a diagnostic test of coherence to take account of these processes.

Derived from these considerations, the term ‘coherence’ in the present study refers to a latent construct which can be inferred by test users by measuring test takers’ use of formulaic language, syntax complexity and inter-sentence cohesion, when test takers are engaged in inferential thinking relating to domain specific knowledge. The variables which will be used to measure these elements of coherence will be discussed in further detail in the methodology section of this paper.

3.2 The Position of Coherence for Diagnostic Testing in English Medium Instruction

This section of the review will argue that coherence is a construct at the heart of the stated purpose of this study for two reasons. First of all, previous empirical and theoretical investigations have suggested that the construct is helpful in theorising the relationship between both content knowledge and linguistic proficiency (Kintsch, 2018; Landauer & Dumais, 1997). On one hand, experimental evidence suggests that prior knowledge is essential in the processing of text and

forming a coherent mental representation of the information in that text (McNamara, 2001; McNamara & Kintsch, 1996). On the other hand, research suggests that the complexity of written expression is affected by topic familiarity (Tabari & Wang, 2022).

To date, there are few studies that have investigated the association between prior knowledge and the coherence of student writing. Langer (1983) investigated coherence as an element of a broader construct of writing quality, finding a positive correlation between the quality of 10th grade L1 English student writing and topic familiarity. Ahmed (2010) investigated coherence in the writing of L1 Egyptian students training to become teachers of English as a foreign language, reporting that prior knowledge strongly affected the students' written output. Due to the samples used in these studies, their findings might not be generalisable to the context of the present work (English medium instruction in higher education). Consequently, further investigation is warranted into the construct as an element of both reading comprehension and competence with written discourse.

The second reason why coherence is of exceptional interest to the present context relates to insights from research investigating how distributional patterns of lexical items in large corpora reflect the psycholinguistic processing of language users. Hoey (1991) proposes that coherence relates to the communicative dimensions of a text, and is dependent on the reader subjectively making inferences about the relationships of the illocutionary acts communicated in that text. In this sense coherence is closely related to cohesion in that the processing involved in inferring the illocutionary implications of a text is dependent on how units of information relate to one another across words, sentences and larger units of discourse. The repetition of lexical items is instrumental in establishing such relationships, and consequently, the frequency with which a reader anticipates encountering certain words is an essential element of coherence. In Hoey's (1991) model, coherence might be said to relate to the expectations language users hold regarding encountering certain words in close proximity.

In this way, the claims made in Hoey (1991) relate to how researchers might define differences between genres and disciplines in English for academic purposes. More recent investigations into such differences have largely corroborated these claims, demonstrating that collocations vary between academic genres and disciplines (Durrant, 2017). Furthermore, these insights might help generate interventions responding to the diagnosed needs of learners. Hoey (1991) for example, proposes that exploiting the repetition of lexical items across sentences may help compensate low lexical coverage in beginner learners. More sophisticated methods of summarisation applying collocational methods have also been tested with some success (Clarke & Lapata, 2010).

Nevertheless, to date, research on the use of collocational methods to support language processing are mixed. Words associated by collocation are not always associated by meaning, limiting the applicability of computational methods of abbreviation based on distributional methods of word-meaning representation (Budanitsky & Hirst, 2006; Carrell, 1982). Furthermore, research in this field has demonstrated that the benefits of processing speed are visible in L1 populations where words are related by meaning rather than collocation (Durrant & Doherty, 2010). Research into L2 populations using eye-tracking methods has demonstrated a theoretical need to separate processing fluency from processing accuracy, and emphasized the influence of cross-linguistic effects in processing text in L2 English (Kuperman et al., 2022; Nisbet et al., 2022). The present research might generate further informative data on this question.

Having defined the purpose of diagnostic testing, and the construct of coherence, the following sections of this review of literature will consider how a diagnostic test might be operationalised to integrate skills of reading and writing in order to develop a diagnostic test of coherence for L2 users of English for academic purposes.

Student ID : 21048826

4 How to Test for Coherence

4.1 Integrated Reading-Writing Tests for Diagnostic Purposes

This section of the review will consider the possibility of using integrated task designs for diagnostic testing. Applying integrated methods to diagnostic testing will be considered in relation to the 3 identified areas of interest for the present work. To reiterate, these three areas might be summarised as:

- The authenticity of tasks based on register, genre and discipline (Duan & Shi, 2021; Wang & Xie, 2022);
- The effect of source text cohesion on test taker performance (Bilki & Plakans, 2022; Cai & Chen, 2022);
- The effects of cultural and social differences on the use of academic language (Cai & Chen, 2022; Duan & Shi, 2021).

Controversy in research around the validity of integrated task designs is long standing. Early research focused on construct irrelevant variance arising from testing methods (Bachman & Palmer, 1982; Oller, 1979). Bachman and Palmer (1996) proposed that, in order to establish a valid target performance domain, test developers should consult with relevant stakeholders. Following this suggestion, Buck and Tatsuoka (1998) applied a procedure of first consulting with stakeholders to establish hypothesized cognitive processes qualitatively, and later applied statistical modelling to consider how much variance in test scores was explained by these respective methods. In their research they conclude that with such fine grained measures, a return to the integrated methods proposed by Oller (1979) may be desirable.

Cumming (2013) argued that the potential for integrated task designs for diagnostic purposes is considerable. Consistent with the argument of the present study, Cumming (2013) proposes that the rationale for such an application of this integrated task design is for students to demonstrate their ability in using academic registers which are “coherently relevant to their fields of

study" (p. 2). What this coherence entails is left underexplored, nevertheless, arising from the research presented in the Language Testing special issue on the testing of integrated skills, Cumming (2013) highlights newly emergent opportunities and threats to validity. These threats will now be considered in the context of subsequent complementary research.

Corresponding with the first of the areas of interest for the present work, Cumming (2013) claims that the appropriate use of integrated task designs may provide realistic literacy activities. In the context of diagnostic testing, this is consistent with one of the seven test qualities outlined in Phakiti and Isaacs (2021). Phakiti and Isaacs (2021) argue that authenticity is a crucial element of classroom based tests. The researchers propose that in order for a formative test to be authentic, it "should contain language samples that are natural, meaningful and relevant to real-world situations" (p. 13). How real world situations are defined in the context of English for academic purposes must therefore depend on the discursive conventions and content taught within academic disciplines. Cumming (2013) proposes that this places particular emphasis on the relationship of content knowledge and text comprehension. In turn, this relationship between writing skill and comprehension opens the possibility for confounding the distinct abilities.

Research responding to such confounding has applied the cognitive, construction-integration model as a basis for robust and clear construct definition (Sawaki, Quinlan, & Lee, 2013; Weigle, Yang, & Montee, 2013). This model proposes that the meaning-making process involved in reading can be derived from distributional patterns of lexical items (Kintsch, 1988, 2018). Using a combination of previously acquired knowledge, and the position of content-words in relation to others, the model proposes the different senses of a word are limited through a process of constraint satisfaction. The relationship of the operationalisation of previous knowledge and bottom-up text comprehension is particularly valuable for the present context in its ability to emulate the communicative function of language. Landauer and Dumais (1997) applied this model to accurately predict language acquisition in early L1 learners, and operationalise a construct of coherence. This study was limited, however, by the representation of the construct of coherence as

simply analeptic reference based on synonymy. Subsequent research has suggested that collocation may not adequately represent synonymy (Budanitsky & Hirst, 2006; Durrant & Doherty, 2010).

More recent research has built on what is known about test taker strategy and cognitive processing, highlighting the second area of interest for the present work (Bilki & Plakans, 2022; Cai & Chen, 2022). Nevertheless, research is generally consistent on the need for diagnostic tests needing to balance authenticity with practicality (Phakiti & Isaacs, 2021; Weigle et al., 2013). Cumming (2013) emphasizes that there is a risk in such cases for the inadvertent use of ill-defined constructs. Gebril and Plakans (2013) propose fluency, lexical sophistication, syntactic complexity, grammatical accuracy, verbatim source use, and direct and indirect source use as elements of the construct which integrated task design can test. Wolfersberger (2013) specifies further, emphasizing the role of task representation in classroom contexts. For the proposed context, expecting content instructors to have the time develop and validate tests which satisfy these demands seems unrealistic. Consequently, returning to the question of practicality, Sawaki et al. (2013) argue that the application of fine grained indices generated by automated essay scoring software may provide a valid means of striking this balance. These claims will be interrogated more closely in the next section of this review.

Finally, the effects of cultural and social differences on the use of academic language may also be confounding factors in formative assessment (Cai & Chen, 2022; Duan & Shi, 2021). Phakiti and Isaacs (2021) argue fairness is crucial in classroom based tests. The researchers go on to write that in order for a test to be fair, it may need to make “accommodations for students with special needs” (Phakiti and Isaacs, 2021: 13). The American Psychological Association Joint Committee on Testing Practices (2000) specify further that accommodations aiming to make a test fair should account for the needs of “those with diverse linguistic backgrounds” (p. 6). Cumming (2013) warns about the risks of integrated tasks requiring a threshold of competence for results to be comparable between test takers. Gebril and Plakans (2013) propose that cohesion and content play a more significant role in populations who score in the top ranges of the TOEFL iBT.

Consequently, these features are of particular interest for a post-entry diagnostic test (Bilki & Plakans, 2022).

There is some controversy in the literature regarding whether these considerations come at the expense of the validity of a test. Borsboom, Mellenbergh, and Van Heerden (2004) advocate for a strict definition of validity, proposing a separation between considerations of the consequences of a test and validity. However, to ask whether a formative test is valid would not make sense in Borsboom et al. (2004) conception of validity. The researchers argue that in models where the observed indicators are not caused by a latent variable, but rather cause that latent variable, the ontological claims which their concept of validity is intended to make become untenable. This would be the case in a test where the latent construct of coherence was inferred from a test taker's performance. While for summative assessment, a realist concept of validity might have value in its ability to differentiate between individuals, or between different stages of an individual's development, for the purposes described above its appropriateness is arguably limited.

4.2 Automated Essay Scoring Software for Diagnostic Language Testing

This section of the review will respond to the demands of authenticity and practicality which emerged as salient in the previous section. To reiterate, authenticity relates to the use of source texts which are "natural, meaningful and relevant to real-world situations" (Phakiti and Isaacs, 2021: p.13). Practicality refers to a diagnostic test's ability to be operationalised within the time and resource constraints endemic in educational settings. Sawaki et al. (2013) suggest automated measurement of fine grained indices by Natural language processing software may facilitate striking a balance between these two demands.

This is of particular interest for the present context because the expectations of language learning in English medium instruction courses varies between different stake holders. Research suggests instructors consider themselves content experts but do not feel language instruction falls within the remit of their expertise (Block & Moncada-Comas, 2022; McGrath et al., 2019).

Nevertheless, policy makers and students do consider English medium instruction to be valuable because of claims which are made about its benefits to language learning (Macaro et al., 2018; Pecorari & Malmström, 2018). Furthermore, research in cognitive psycholinguistics suggests a complex interconnected relationship between content knowledge and language use (Kintsch, 2018). Sawaki et al. (2013) suggest automating diagnostic language testing may be an appropriate response to these demands.

To date, several studies have investigated the application of automated written assessment to diagnostic purposes. Chapelle et al. (2015) investigated the validity claims of two examples of automated writing evaluation software. Of particular interest to the present study is the investigation of the Intelligent Academic Discourse Evaluator because it was designed to offer feedback regarding discipline specific academic writing. The researchers applied a mixed method design to determine if the feedback provided to students helped them focus on how meaning is constructed in research articles. The research indicated that the color-coded modified input method used by the software helped students reflect on their meaning construction, though they also found that the software might encourage over-dependence on certain lexical formulations. The versatility of automatically modified input for the students to learn from which was generated by the software as a timely and personalised intervention is especially relevant for the present work. Nevertheless, investigating the cohesive features of the input generated by the software was beyond the scope of Chapelle et al. (2015).

M. Chen and Cui (2022) investigated the application of automated writing evaluation and peer feedback on the cohesion and coherence of student writing. The study focused on students responding to a continuation task. The study found that peer feedback was more effective in eliciting improvements in the coherence of student writing of a re-draft of their work compared to the feedback offered by the software. The researchers recommend that future researchers explore how chains of semantic reference interact in texts. The method of textual abbreviation proposed by Hoey (1991) is one method of doing so, where lexical items are conceptualised as edges in a network

which link sentences represented as nodes. Hoey (1991) proposes that by focusing on repetitions of words, learners can compensate for low lexical coverage through a process of syllogistic inference. Kim, Nam, and Crossley (2022) recently investigated a similar process in L2 listening comprehension. They found that inferencing abilities for listening comprehension are transferable to L2. They also found that working memory capacity had a significant influence on their results. Finally they found that linguistic knowledge had a significant effect on both shorter and longer passages. The present work will apply the method described in Hoey (1991), and may generate insight to see if the role of inferencing in listening comprehension is comparable to comprehension of written academic discourse.

5 Summary

English medium instruction at internationalised higher education institutions presents a challenging context for the design of diagnostic language testing software. Among these challenges is the relationship of language instruction and content instruction. Research suggest conflicting views held by stakeholders, with content instructors not feeling sufficiently informed to focus on teaching English for academic purposes (Block & Moncada-Comas, 2022; McGrath et al., 2019). Fenton-Smith and Humphreys (2015) report that language experts consider diagnostic testing to be an effective strategy to aid L2 learners in the use of English for academic purposes. Nevertheless, Phakiti and Isaacs (2021) argue that in order for classroom based testing to be effective it must be valid, reliable, authentic, practical, fair, and ethical.

Recent evidence suggests that three areas remain under-explored in research investigating post entry diagnostic testing:

- The authenticity of tasks based on register, genre and discipline (Duan & Shi, 2021; Wang & Xie, 2022);
- The effect of source text cohesion on test taker performance (Bilki & Plakans, 2022; Cai & Chen, 2022);

- The effects of cultural and social differences on the use of academic language, (Cai & Chen, 2022; Duan & Shi, 2021).

Based on these three areas, I have formulated the following 3 research questions.

6 Research Questions

- RQ1: Is there a difference in the coherence of L2 English students' writing in genres which make reference to external sources compared to genres which do not, as represented in the British Academic Written English corpus?
- RQ2: Is there a difference in cohesion-based reading ease measures between texts which have been abbreviated using Latent Semantic Analysis, Word2Vec, or Latent Dirichlet Allocation?
- RQ3: Is there a difference in the coherence of L2 Student's writing in a pilot trial of a classroom-based test of coherence when responding to a text abbreviated using Latent Semantic Analysis?

Student ID : 21048826

7 Methodology

This chapter will provide a framework and justification for the chosen research design. In this chapter I will address the methodological approach, the methods of data collection, the procedure of the research and the data analysis.

7.1 Paradigms

It seems trivial to acknowledge that epistemological and ontological positions which are commonly held in specific fields of investigation within applied linguistics are likely to have emerged from the questions which that community of researchers are concerned with answering. In language testing, questions of reliability and validity have been central pillars of investigation around which attitudes which evaluate constructivist and positivist answers to practical problems have formed (Cronbach, 1971; Cureton, 1951; Kane, 2010; Messick, 1989). There is a long discussion and continuing debate regarding how to test fairly, and what must be considered, emphasized or sacrificed in order to develop a test which is practical and useful (Chapelle, 2020a). For example, the validity of psychometric approaches to educational measurement, which might be more at home in positivist frameworks, continues to be interrogated by researchers (Chalhoub-Deville, 2016; Phakiti & Isaacs, 2021). Equally, despite literature in the field increasingly emphasizing social and ethical contexts, investigating the reliability of test scores remains fundamental (Borsboom et al., 2004).

Ultimately, the degree to which hypotheses can be empirically tested depends on the ability to make assumptions (Orman Quine, 1976). Therefore, research in its early stages, such as the present work, must remain tentative in articulating findings since it is ultimately attempting to balance holistic pragmatism with generalisability, validity and reliability (Orman Quine, 1976). Consequently, interest in the field arguably demonstrates that the need for careful consideration of the what, why and how of language testing is becoming increasingly complex and nuanced (Chapelle, 2020b; Shohamy, 1990). Positions of language testers responding to these questions is as varied as it is long standing (Bachman & Palmer, 1996; Oller, 1979). Consequently, the present work

takes coherence as the construct of investigation, as opposed to a general language ability, but proposes that integrated testing may be the most appropriate method of its measurement.

8 RQ1: Is There a Difference in the Coherence of L2 English Students' Writing in Genres Which Make Reference to External Sources Compared to Genres Which Do Not, as Represented in the British Academic Written English Corpus?

8.1 Research Design

The first research question relates to the first outlined research gap which emerged from the literature review. The most recent research into diagnostic testing of English for academic purposes suggests that The authenticity of tasks based on register, genre and discipline is of particular interest to the field (Duan & Shi, 2021; Wang & Xie, 2022). In response to this, the present research aims to operationalise the construct of coherence to include the use of formulaic language.

Previous research has demonstrated the specificity of multi-word sequences as they relate to genre and register (Y. Chen & Baker, 2016; Durrant, 2017; Simpson-Vlach & Ellis, 2010). Consequently, the present research intends to investigate if there is a difference In the mean strength of association in bigrams in the British Academic Written English Corpus, by comparing genres of L2 writing which make use of sources to genres which do not make use of sources.

The present study proposes the construct of coherence is situated as an especially important area of investigation for the proposed context and use. Wang and Xie (2022) divide the construct into two elements: local coherence and global coherence. Local coherence relates to the ratio of new information in each progressing sentence. Global coherence relates to the overall conceptual structure of the text. This is a view of coherence frequently held in applied linguistics (Halliday & Hasan, 2014; Landauer & Dumais, 1997).

Such approaches, however, may have under-represented some pragmatic communicative elements of the construct. Ito et al. (2018) propose that the cognitive process involved in communicative language use are deeply interwoven, with the inferencing of a general conceptual

mental model occurring concurrently with decision making in productive processes. Research has investigated the role of inferencing on the construction of mental representations of information through text, revealing a close relationship between prior knowledge of content, language proficiency and comprehension (Kintsch, 2018; McNamara, 2001). Consequently, the first research question in this paper aims to investigate if there is a difference in the means of 3 variables which form a latent construct of coherence.

Clarke and Lapata (2010) tested coherence by measuring how much information was preserved from a unabridged text compared to an abridged version through the use of a quiz coded by human raters. While such an approach is laborious and not appropriate for the present study, the concept of "centering" which appears in the study is what they argue maintains the coherence of the abridged text. Centering may be summarised as the degree to which a unit of information is repeated in different contexts in the text. Hoey (1991) uses the same metaphor and explains that sentences which are more central are more connected to other sentences in that text through lexical patterning. Conversely, sentences which are marginal bear fewer connections.

While a language user is engaged in the automatised inferencing and predicting involved in text comprehension, effects are likely to be demonstrable in other elements of the construct of coherence. Based on previous research, one hypothesis of the present study is that inter-sentence cohesion, the use of formulaic language, and grammatical complexity may be affected by the use of source texts.

This part of the study was exploratory, quantitative, and corpus based. A sample of L2 English student work representing two genres was selected from the British Academic Written English corpus (Alsop & Nesi, 2009). The sample was split according to discipline with life sciences and physical sciences representing "hard" academic discursive practices and "soft" academic discourse being represented by the social sciences, arts, and humanities, according to a taxonomy proposed by Durrant (2017). Means were compared for 3 variables intended to measure inter-

sentence cohesion, the use of formulaic language, and grammatical complexity in the students' writing.

8.2 Corpus

The British Academic Written English corpus was developed at the Universities of Warwick, Reading, and Oxford Brookes as part of the project An investigation of genres of assessed writing in British Higher Education (Alsop & Nesi, 2009). The corpus is a collection of 2,897 texts written by students and curated by language experts as examples of successful academic writing. This selection of student writing is appropriate for the present study because previous research indicates that language proficiency at the level of discourse is a more important differentiating factor for proficient users (Gebril & Plakans, 2013).

Furthermore, while previous research into integrated reading and writing has investigated the effect of source use on test takers, these studies have generally used corpora composed of texts written for pre-entry tests of language proficiency (Guo, Crossley, & McNamara, 2013). This is an important difference because the tasks performed in pre-entry tests of language proficiency are timed, and may not represent source use in academic discourse as authentically as the British Academic Written English corpus.

The corpus has been used in research to identify differences in language use between genres, with the corpus holdings being divided into 13 genre families (Gardner & Nesi, 2013). Essays and reviews of literature were selected to represent student work which makes use of sources. Case studies and explanations were selected as genres representative of student work which is less likely to make use of sources.

8.3 Sample

The sample used to answer the first research question was composed of student writing written in L2 English (n = 396). The sample of student work representing the physical and life sciences (n = 101) was roughly one third the size of the sample representing the social sciences, arts,

and humanities (n = 295). There was a large difference in the size of the sample of source-based (n = 249) and other (n = 46) academic writing for the social sciences, arts, and humanities. Levene's test of homoskedasticity was used to determine whether equal variances were to be assumed in the t-test.

8.4 Selection Of Indices

Three variables were selected to measure the construct of coherence. The construct was operationalised as consisting of inter-sentence cohesion, use of formulaic language and syntax complexity.

The first variable, relating to sentence level cohesion, was selected to apply latent semantic analysis to measure the semantic similarity between all pairs of adjacent sentences. Greater semantic relatedness of two sentences is contingent on the ratio of new information in each sentence. A higher score in this sense, suggests that there is less new information introduced in each successive sentence. This relates to local coherence as operationalised by Wang and Xie (2022). This method of semantic analysis was preferred because of the robust body of research investigating its psycholinguistic validity (Kintsch, 2018; Landauer & Dumais, 1997).

The second variable selected for use in answering the first research question relates to the use of formulaic language. The selected variable for measuring this element of the construct was the mutual information score of word pairs. This method of calculating word association was first proposed by Church and Hanks (1990). The preference for using bi-grams for this study, is due to the exploratory nature of the research. While longer sequences would doubtless provide a more qualitatively descriptive picture of the use of formulaic language, this was less appropriate for the present research which aims only to establish a difference in the use of formulaic language between source-based and other academic writing. Using a shorter sequence of words increases the likelihood that relationships can be found.

Mutual information scores were selected instead of raw scores because of their use in previous research on collocational priming (Durrant & Doherty, 2010). Other measures under-emphasise the strength of a relationship between the two words in a bi-gram, and consequently are less appropriate for the present method. To reiterate a previously used example, “taste arbiters” would receive a higher score than “taste for” despite the latter occurring more frequently because “taste” occurs with a higher than expected frequency in a position adjacent to “arbiters” (Durrant & Doherty, 2010).

The final variable selected for this research question intends to measure the complexity of the writers’ syntax. Because of the measures ability to differentiate between dependent and independent clauses, mean length of t-unit was most appropriate for the exploratory nature of the present research. Kyle (2016) defines a t-unit, in accordance with much earlier research, as an independent clause and any dependent clauses. Recent research has demonstrated that corelation between mean length of t-unit and human ratings of essay quality reached statistical significance in both independent and integrated writing tasks (Kim & Crossley, 2018).

8.5 Data Analysis

Data was analysed using three natural language processing tools to measure the selected variables in the sample. TAACO 2.0 was used to measure inter-sentence cohesion using Latent Semantic Analysis to determine the level of semantic overlap between adjacent sentences (Crossley, Kyle, & Dascalu, 2019).

The use of formulaic language was measured using TAALES 2.0 (Kyle, Crossley, & Berger, 2018). The software was used to measure the strength of association of words in bigrams which also appear in the Corpus of Contemporary American, Academic Subcorpus (Davies, 2009). Mutual Information scores were used to measure the strength of relationship between the words in the bi-grams.

The complexity of the students' syntax was measured using TAASC (Kyle, 2016). The software was used to measure mean length of t-unit (Lu, 2010).

The resulting data was processed using SPSS. Independent samples t-tests were conducted comparing the means of the sourced and unsourced writing. Levene's test of homoskedasticity was used to determine whether equal variances should be assumed in the sample. Gpower was used to calculate effect sizes for differences which reached statistical significance.

9 RQ2: Is There a Difference in Cohesion-Based Reading Ease Measures Between Texts Which Have Been Abbreviated Using Latent Semantic Analysis, Word2vec, or Latent Dirichlet Allocation?

9.1 Research Design

The second research question in this study relates to how source texts might be modified as an intervention to aid coherence in student writing. Recent research into diagnostic and integrated language testing has highlighted the role of thinking skills in post entry diagnostic tests, and the role that source text cohesion plays in strategies used by test takers (Bilki & Plakans, 2022; Cai & Chen, 2022). The effectiveness of a diagnostic test is in part contingent on the application of appropriate interventions based on this test(Fenton-Smith & Humphreys, 2015; Phakiti & Isaacs, 2021). Furthermore, in order for the intervention to be effective, research into social constructivist paradigms of pedagogical development emphasizes the relationship of communication and content knowledge (J. Xi & Lantolf, 2021). This view is consistent with some cognitive accounts of comprehension (Kintsch, 2018).

J. Xi and Lantolf (2021) take issue with the contention that the development of a learner can be separated from that learner's social context. Rather than seeing the qualities of fairness and authenticity as trade-offs (as might be inferred from the architectural metaphor of scaffolding), social-constructivist paradigms of development assert the necessity of both for learning. J. Xi and Lantolf (2021) contrast the architectural metaphor of scaffolding, where support is traded for authenticity, with the agricultural metaphor used by Vygotsky to explain the concept of the zone of

proximal development. Of relevance for the present work is the way in which the agricultural metaphor situates the learner in their context as a seed in fertile earth. Recent research in language instruction has increasingly emphasized L2 learners as a part of, rather than separate from the communicative contexts in which they are invited to participate (Dewaele, 2018).

The second research question aims to apply what is known about the cognitive processes involved in text comprehension in order to establish an effective intervention for students following the diagnostic test. The intervention is not intended to make a text “easier.” Instead it aims to make use of known cognitive processes to encourage syllogistic inferencing in readers in order for that reader to form a more robust mental representation of the content of the text. Previous research has demonstrated the effects of text coherence and prior knowledge on comprehension (Kintsch, 2018; McNamara, 2001). Furthermore, more recent research in to L2 comprehension processes has demonstrated both the interwoven nature of receptive and productive processes as well as the role of inferencing and predicting (Ito et al., 2018; Kim et al., 2022).

Consequently, the second research question aims to evaluate 4 methods of text abbreviation as suggested by Hoey (1991). Hoey (1991) proposes that by abridging a text to only include sentences that are semantically linked to other sentences, it will be possible for a second language reader to infer a mental model of the information contained in that text. In Hoey (1991) the sentences which interact with more semantic chains of reference are considered more central. Clarke and Lapata (2010) applied a similar model of centering demonstrating positive results with a ranking exercise of coherence using human raters. In this way, it will be possible to aid readers by directing their attention to the way in which semantic chains interact, as recommended by M. Chen and Cui (2022). To illustrate, fig. 1 shows a network of such interactions as presented in Hoey (1991). Each node in the network represents a sentence, and each edge represents a shared semantic reference between the sentences.

Student ID : 21048826

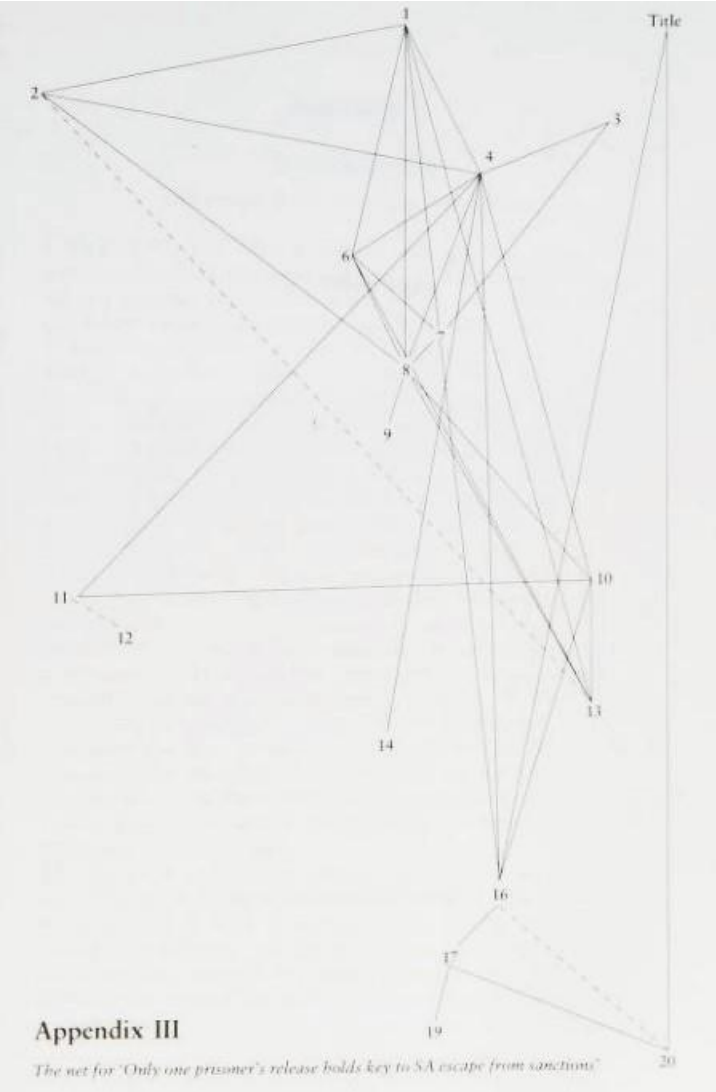
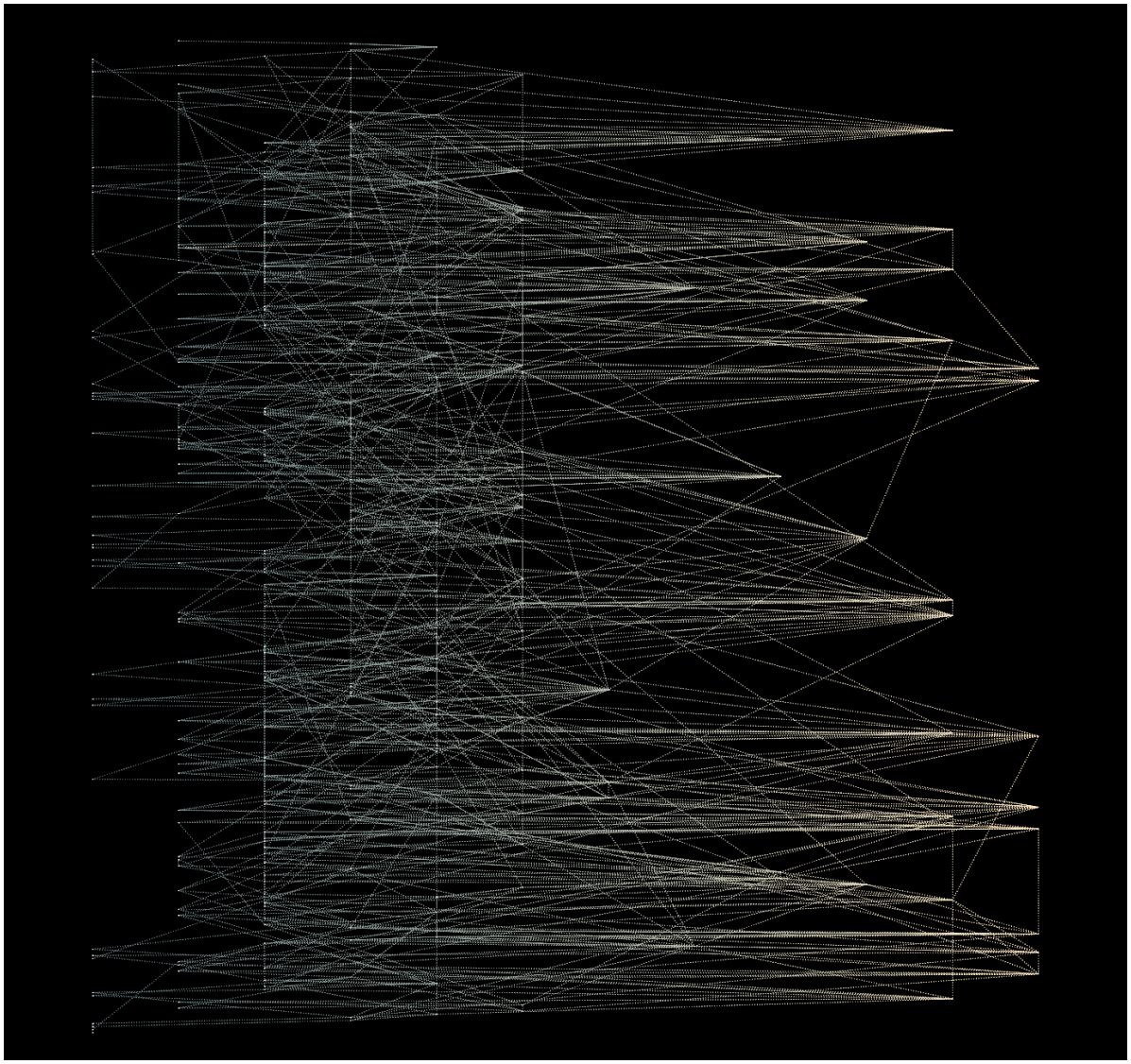


Fig 1. A network diagram representing sentence bonding in Hoey (1991: 253).

Unfortunately, the example used in Hoey (1991) is hand coded and relates to a short, newspaper passage. To demonstrate how centrality might be represented in a full academic text, a series of 80 vectors was derived to represent lemmas which appear frequently in an example text. The lemmas were selected on the basis that they were content lemmas, and that their dispersion across sentences was higher than the modal average for all content lemmas. Consequently, only the lemmas which appeared in the most sentences were represented.

A part-of-speech tagger included in TAACO 2.0 based on Stanford Core Natural Language Processing algorithms was used to determine the vectors. Each vector was composed of the number of the sentence as it appeared in the sequence of the original text. If a lemma appeared in the first,

fifth, and sixth sentences in the text, these numbers would be what that lemma's vector consisted of. These vectors were plotted on a graph using Microsoft Excel, with the vertical axis representing the sequence of the sentences with the first sentence at the top of the graph. The sentences, represented as nodes were plotted along the horizontal axis according to number of content lemmas contained in the sentences. The sentences which contained the largest amount of content lemmas were represented as central. Sentences which contained fewer content lemmas than the modal average were plotted to the left of the central sentences. Sentences which contained more of the content lemmas than the modal average, and would therefore be selected for the abridged texts were plotted to the right of the central sentences. In both directions, the sentences on the margins of the network diagram contain the lowest amount of content lemmas.



Student ID : 21048826

Fig. 2. A network diagram representing sentence bonding in an academic text used in the corpus for this study.

This research set out to find out if reading ease measures used in previous research, would be positively affected by the method of abbreviation proposed in Hoey (1991). Hoey (1991) makes the case for categorising cohesive references into three groups: simple lexical repetition (typified by graphological or phonological identity of lemma forms), complex lexical repetition (typified by identity of distributional frequency), and paraphrase (typified by hypernymy, hyponymy or coreference). By elaborating a method of categorisation using a flow chart, Hoey (1991) argues that each sentence in a cohesive text might be thought of as a text in and of itself, implying a fractal relationship. In this way, he demonstrates how strings are formed between sentences, and these strings might be further elaborated into networks of inter-sentence referentiality.

To tackle the problem of moving beyond simple lexical repetition, TAACO 2.0 was used to generate a score of semantic similarity between each individual sentence in a text and all the content words in that text. TAACO 2.0 offers 3 methods of measuring semantic similarity based on machine learning algorithms trained on the Corpus of Contemporary American (Davies, 2009). To evaluate each method, a corpus of 41 texts taken from the reading lists of the 7 faculties of University College London were abbreviated using simple lexical repetition, or complex lexical repetition based on the algorithms available in TAACO 2.0.

Simple Lexical Repetition. To summarise the texts using simple lexical repetition, the 41 texts in the corpus were lemmatised using TAACO 2.0. Subsequently, the lemmas of content words were ranked according to the frequency of their appearance in the text. Content words which occurred in the text with a frequency above the modal average were selected for coding. Each sentence was given a score based on how many of these words appear in that sentence. Sentences with a score above the modal average were selected for inclusion in the abbreviated text. While this method is relatively fast to compute, it only accounts for part of Hoey's (1991) model, with

semantically related lemmas, which are not orthographically identical remaining uncounted in the method.

Latent Semantic Analysis. LSA is a method of emulating lexical processing using Singular Value Decomposition (Kintsch, 1988; Landauer & Dumais, 1997). Words are represented as vectors in a matrix. A process of matrix factorisation computes each vector based on the probability of words co-occurring within a specified window in a large corpus. A generated sparse matrix then ranks vectors according to their similarity and reduces the dimensionality of these vectors. One differentiating feature between LSA and other vector based word representation algorithms is that optimisation has demonstrated that the algorithm begins to lose accuracy when the vector space is reduced below 500 dimensions (X. Chen, Qi, Bai, Lin, & Carbonell, 2011). In this way, it differs from methods such as Word2Vec and Skipgram which use more computationally efficient algorithms, but that also have less research validating their application in modelling cognitive processes (Goldberg & Levy, 2014; Günther, Rinaldi, & Marelli, 2019).

Latent Dirichlet Allocation. LDA is a probabilistic modelling approach to discovering latent connections within collections of discreet data (Blei, Ng, & Jordan, 2003). The generative process layers a hierarchical Bayesian model on three levels to find an underlying set of latent topics. If words are related, this is indicated in the probability of their co-occurrence. In the model, relationships between words are represented as latent variables and Dirichlet distributions. For the purpose of this application, a topic can be understood as a latent shared meaning between two synonyms. Whilst LDA does not capture correlations between discovered latent topics, as a supervised machine learning method it may still be a sufficiently effective way of measuring complex lexical repetition and paraphrase (Li & McCallum, 2006).

Word2Vec. Word2Vec uses a technique comparable to LSA to encode vectors as representations of words, with more similar vectors indicating similar words (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Whilst LSA applies

singular value decomposition, generating two orthonormal, transposed matrices of singular vectors, and a sparse, diagonalised matrix of singular values, Word2Vec applies a two-layer neural-network where one layer is represented by a matrix which consists of vectors representing word meaning and the second matrix consists of representations of words which are likely to appear in a context window around the target word. Whilst this method uses a less computationally expensive method of processing latent word meanings, the reasons behind its performance is less well understood than earlier models of word meaning representation.

9.2 Corpus

A corpus of 41 texts was selected from seven representative faculties of University College, London. The corpus totalled 336,677 words of writing from a range of genres.

9.3 Selection of Indices

Research into reading ease tends to classify features according to the complexity of lexis, syntax and discourse cohesion (Martinc, Pollak, Robnik-Šikonja, 2021; De Clercq and Hoste, 2016). Due to the need to maintain authenticity in the representation of academic discourse and technical vocabulary to learners, only cohesion is of interest as a variable in the evaluation of the potential model of text summarisation (BIlki and Plakans, 2022).

To measure the sentence level cohesion, the present work used the same variable which was reported in Crossley, Greenfield and McNamara (2008). The lemmas of each sentence were compared with the lemmas of the two adjacent sentences around it, with a higher overlap of lemmas indicating a more cohesive text. To determine if there was any difference in the likelihood that ambiguity may arise in the abbreviations, two indices were chosen to consider pronoun: noun ratio and givenness measured by frequency of unattended demonstratives. Means were compared against the unabbreviated texts. Matched samples t-tests were performed on all texts using data generated from the aforementioned indices. Means, standard deviations, t-values and effect sizes are reported in the results.

Student ID : 21048826

9.4 Data Analysis

The corpus was processed using TAACO 2.0 software. Results were analysed using SPSS to calculate descriptive statistics and T-tests. Effect sizes were calculated using Gpower software.

10 RQ3: Is There a Difference in the Coherence of L2 Student's Writing in a Pilot Trial of a Classroom-Based Test of Coherence?

10.1 Research Design

The review of literature revealed that the effects of cultural and social differences on the use of academic language is an area of interest for research into diagnostic language testing (Cai & Chen, 2022; Duan & Shi, 2021). Following on from a corpus based investigation into coherence and text modification, the final research question aims to pilot an application of the research. Due to time constraints and a small sample size, this research question is intended only as a proof of concept and pilot study for future research. For this purpose, an integrated reading writing task was developed and used with four participants. Participants responded to a source text which had been abbreviated using the methods used for the second research question in the present work. Latent Semantic Analysis was the selected method of abbreviation due to the results demonstrating greater sentence level cohesion. Furthermore, the method of abbreviation accounted for complex lexical repetition and has a strong body of evidence supporting its cognitive validity (Kintsch, 2018).

Four participants attempted the integrated, reading-writing tasks with both a full academic text and an abbreviated academic text. The sequence of the texts was counterbalanced. A questionnaire was adapted and completed by participants in order to identify potential areas of interest which might be investigated with larger samples, relating to the participants' language use and proficiency. The participants responses were analysed using the same variables as the texts which were analysed for the first research question of the present work. Following writing their response, participants responded to a questionnaire to investigate confounds and moderating factors. The sources from which the questionnaire will be adapted are one proposed by Ehrman

(1996) and a second questionnaire used in Conklin, Alotaibi, Pellicer-Sánchez, and Vilkaitė-Lozdienė (2020). A copy of this questionnaire may be found in Appendix 2.

The test question was developed based on principles articulated in Plakans (2021). Two raters from the target population of test takers were asked to consider the clarity and appropriateness of the task representation. Both raters were L2 users of English enrolled at a UK university studying varying academic disciplines. A short questionnaire was written to evaluate the task representation. This questionnaire is available to view in Appendix 3 of the present study.

Raters agreed that the instructions explaining the purpose of the text which they were being asked to write was explained adequately clearly. This was also the case for the raters' judgements regarding how clearly the instructions explain the intended audience for whom they were being asked to write. Following research into test taker strategy, an important element of validity related to how test takers approach the task. The construct of coherence, as conceptualised by the present work involves higher level inferential processing in which test takers are being asked to apply their long-term working memory (Ericsson & Kintsch, 1995). Consequently, raters were asked to consider how much time they were likely to spend focusing on understanding unfamiliar words when reading the text. Agreement was reached that the text used sufficiently accessible lexis. Similarly, raters agreed that they were both likely to draw on background knowledge when writing their answer.

Raters reported disagreement regarding the frequency with which they expected to monitor for comprehension when reading the text. It is unclear whether this is a result of individual differences between raters, or whether the question was not adequately clear. It is therefore a limitation of the present work that further trialling was not conducted with follow up interviews and with a larger sample of raters. Both raters agreed that the majority of their time would not be spent on low-level decoding and re-reading the text and felt capable summarising the texts' main points.

There was disagreement between the raters regarding how explicitly the purpose of the passage was explained. Consequently, this portion of information was re-worded and reformatted. Raters agreed

that the assessment criteria, and instructions regarding what a test taker should do if they want to directly copy phrases from the text was explained adequately clearly.

The original length of the test was based on the time allocated for integrated writing in the TOEFL iBT (20 minutes). The expected essay length of 300 words was based on research into L2 English for academic purposes (Y. Chen & Baker, 2016). Nevertheless, raters agreed that the length and relative complexity of the text meant that more time was necessary to elicit a response of this length, consequently, the task was lengthened to 40 minutes.

Both raters offered feedback regarding the assumed knowledge of test takers. Topic familiarity is a complex variable to control for. Contextualised in the target domain of the present work, the intention of this intervention is for instructors who are conferring highly specialised information to proficient L2 users of English. The interaction between background knowledge and inferential skills is therefore intimately tied to the construct of coherence as this investigation has conceptualised. Literature investigating genre, discipline and topic familiarity has emphasized the complexity of controlling for this variable (Tabari & Wang, 2022; Yoon, 2021).

With this feedback in mind, a test question was developed and administered to participants. Participants were asked to write a response to an abbreviated and an unabbreviated source text. The texts were not related by topic. The sequence of texts was counterbalanced between the four participants.

10.2 Participants

The participants for this part of the study were all second language English speakers enrolled at a UK university. Convenience sampling was used due to this research being a proof of concept pilot study.

Student ID : 21048826

10.3 Data Analysis

Data was analysed using three natural language processing tools to measure the selected variables in the sample. TAACO 2.0 was used to measure inter-sentence cohesion using Latent Semantic Analysis to determine the level of semantic overlap between adjacent sentences (Crossley et al., 2019).

The use of formulaic language was measured using TAALES 2.0 (Kyle et al., 2018). The software was used to measure the strength of association of words in bigrams which also appear in the Corpus of Contemporary American, Academic Subcorpus (Davies, 2009). Mutual Information scores were used to measure the strength of relationship between the words in the bi-grams.

The complexity of the students' syntax was measured using TAASC (Kyle, 2016). The software was used to measure mean length of t-unit (Lu, 2010).

11 Ethics

A relationship of fidelity and responsibility between stake holders, including test takers, instructors and test developers is central not only for the consequence implication of the proposed test to be valid, but also for trust to be maintained between institutions and the general public. Institutional oversight for this research has been sought and been approved. Related to this is the question of integrity, which applied linguists might understand in terms articulated by Labov (1982) as the principle of error correction and the principle of obligation. Commitment to respect for people's rights and dignity is addressed through explicitly seeking participants' consent, and the relatively low risk of adverse impact on test takers is maximised through the use of methods that do not require face to face meeting, and do not require any covert research or deception. A consent form explicitly seeking active consent and clearly articulating that participation is voluntary and may be revoked at any point without warning or explanation was shared with all participants. Data was anonymised through the use of pseudonyms and stored on a password protected online area according to IoE guidelines should the need arise for post-analysis dissemination. Questionnaire data

was stored through the online service Google Forms, but participants will use a pseudonym for data protection and to be able to match the experiment results with data from the questionnaire.

Treatment of participants is an important element of researchers' commitment to ethical investigation. The effects of cognitive fatigue are well documented in relevant research (Hagger, Wood, Stiff, & Chatzisarantis, 2010; Xu, Zhang, & Gaffney, 2022). Minimising the effect of fatigue has the methodological advantage of maintaining the function of cognitive processes which are affected by the construct being investigated and have been described in recent research (Kintsch, 2018; Plakans, Liao, & Wang, 2019). There are further ethical considerations relating to the depleting effect of cognitive fatigue on participants' positive affect (Parke, Seo, & Sherf, 2015). Consequently, the information sheet will seek to make clear that the construct being measured relates to the modified text and should not be interpreted as a reflection of the participants' intelligence or proficiency in order to minimise the possibility that their performance in the study might negatively impact their self-concept. A de-brief was offered to participants following the procedure to make sure they are doing well. Participants were told of the results of the study.

Student ID : 21048826

12 Results and Discussion

This chapter will present the main results of the study with reference to the research questions expressed in the methodology chapter. Following the presentation of the results, this chapter will discuss the results and implications of the findings with reference to supporting theory and previous research. An outline of some of the limitations of this study will follow, along with suggestions for future investigations. Finally, this study will conclude with a summary of the research.

13 Results

13.1 RQ1: Is There a Difference in the Coherence of L2 English Students' Writing in Genres Which Make Reference to External Sources Compared to Genres Which Do Not, As Represented in the British Academic Written English Corpus?

Variable 1: Inter-Sentence Cohesion Measured by Latent Semantic Analysis. The 51 examples of explanations and case studies from the physical and life sciences (M = .4, SD = .06) compared to 50 examples of essays and literature reviews from the same disciplines (M = .4, SD = .1) did not demonstrate significantly stronger inter-sentence cohesion, t(87.43) = .55, p = .29. Similarly, there was no significant effect for social sciences, arts and humanities, t(293) = 1.61, p = .054, despite explanations and case studies (M = .4, SD = .08) showing the use of more cohesive sentences (M = .4, SD = .07). These results are summarised in tables 1 – 3.

Table 1:

Descriptive Statistics for Cohesion in Physical and Life Science Texts

Genre	N	Mean	SD
Explanations and Case Studies	51	.4	.06
Essays and Literature Reviews	50	.4	.1

Student ID : 21048826

Table 2:

Descriptive Statistics for Cohesion in Social Sciences, Arts and Humanities Texts

Genre	N	Mean	SD
Explanations and Case Studies	46	.4	.08
Essays and Literature Reviews	249	.4	.07

Table 3:

Independent Samples T-Tests for Cohesion in Sourced and Unsourced Genres

Discipline	T	Df	One-Sided P
Physical and Life Sciences	.55	87.43	.29
Social Sciences, Arts and Humanities	1.61	293	.054

Variable 2: Association Strength of Bigrams from the Corpus of Contemporary American, Academic Writing Sub-Corpus. The 51 examples of explanations and case studies from the physical and life sciences (M = 1.5, SD = .2) compared to 50 examples of essays and literature reviews from the physical and life sciences (M = 1.6, SD = .1) demonstrated significantly weaker association strength between words in bigrams which appear in the COCA academic writing sub-corpus t(99) = 1.85, p = .03. There was also a significant effect for social sciences, arts and humanities, t(293) = 2.65, p = .004, with explanations and case studies (M = 1.6, SD = .1) showing stronger bigram associations than essays and reviews of literature (M = 1.6, SD = .1). These results are summarised in tables 4 - 6.

Student ID : 21048826

Table 4:

Descriptive Statistics for Bigrams in Physical and Life Science Texts

Genre	N	Mean	SD
Explanations and Case Studies	51	1.5	.2
Essays and Literature Reviews	50	1.6	.1

Table 5:

Descriptive Statistics for Bigrams in for Social Sciences, Arts and Humanities Texts

Genre	N	Mean	SD
Explanations and Case Studies	46	1.6	.1
Essays and Literature Reviews	249	1.6	.1

Table 6:

Independent Samples T-Tests for Bigrams in Sourced and Unsourced Genres

Discipline	T	Df	One-Sided P	Cohen's D
Physical and Life Sciences	1.85	99	.03	.37
Social Sciences, Arts and Humanities	2.65	293	.004	.44

Variable 3: Mean Length of T-Unit. The 51 examples of explanations and case studies from the physical and life sciences (M = 21.1, SD = 4.9) compared to 50 examples of essays and literature reviews from the physical and life sciences (M = 21.1, SD = 4.5) demonstrated significantly shorter t-unit length t(99) = 2.5, p = .007. There was also a significant effect for social sciences, arts and humanities, t(293) = 1.84, p = .03, with explanations and case studies (M = 23.9, SD = 4.9) showing shorter average length of t-unit compared with essays and reviews of literature (M = 25.6, SD = 5.4). These results are summarised in tables 7 – 9.

Student ID : 21048826

Table 7:

Descriptive Statistics for T-Units in Physical and Life Science Texts

Genre	N	Mean	SD
Explanations and Case Studies	51	21.1	4.9
Essays and Literature Reviews	50	23.4	4.5

Table 8:

Descriptive Statistics for T-Units in Social Sciences, Arts and Humanities Texts

Genre	N	Mean	SD
Explanations and Case Studies	46	23.9	4.9
Essays and Literature Reviews	249	25.6	5.4

Table 9:

Independent Samples T-Tests for t-units in Sourced and Unsourced Genres

Discipline	T	Df	One-Sided P	Cohen's D
Physical and Life Sciences	2.5	99	.007	.49
Social Sciences, Arts and Humanities	1.84	293	.03	.3

13.2 RQ2: Is There a Difference In Cohesion-Based Reading Ease Measures Between Texts Which Have Been Abbreviated Using Latent Semantic Analysis, Word2Vec, or Latent Dirichlet Allocation?

Variable 1: Adjacent Two-Sentence Overlap Of All Lemmas. The results from the unabbreviated texts (M = 6.5, SD = 1.6) and texts abbreviated using Latent Semantic Analysis (M = 8.6, SD = 2.4) indicate that sentences in a moving 2 sentence window are more cohesively tied in the abbreviated text, t(40) = 5 , p = <.001. The results of the text which had been abbreviated using only simple lexical repetition (M= 10.3, SD= 2.9) also suggested that the sentences in the abbreviated text were more cohesively tied to the subsequent 2 sentences in the text, t(40) = 7.68, p = <0.001. The

other methods of abbreviation did not show significant effects compared to the full texts. These results are summarised in tables 10 – 11.

Table 10:

Descriptive Statistics for Inter-Sentence Cohesion

Method of Abbreviation	N	Mean	SD
Full Text (FT)	41	6.5	1.6
Latent Dirichlet Allocation (LDA)	41	5.9	1.9
Latent Semantic Analysis (LSA)	41	8.6	2.4
Simple Lexical Repetition (SLR)	41	10.3	2.9
Word2Vec (W2V)	41	7.4	2.8

Table 11:

Paired Samples T-Tests for Inter-Sentence Cohesion

Pair	T	Df	One-Sided p	Cohen's D
FT - LDA	1.4	40	.84	-
FT - LSA	5	40	<.001	0.78
FT - SLR	7.68	40	<.001	1.2
FT - W2V	1.78	40	.041	-

Variable 2: Ratio Of Pronouns To Nouns. The results from the unabbreviated texts (M = .1, SD = .06) and texts abbreviated using Latent Semantic Analysis (M = .1, SD = .08) indicate that the ratio of pronouns to nouns was not significantly different, t(40) = .95 , p = .35. The results of the text which had been abbreviated using simple lexical repetition (M= .1, SD= .06) also indicated that the sentences in the abbreviated text contained a ratio of pronouns to nouns which was not significantly different from the unabbreviated texts, t(40) = 1.4, p = .2. Similarly, the other methods of

abbreviation did not show significant effects compared to the full texts. These results are summarised in tables 7 – 9.

Variable 3: Givenness Measured by Unattended Demonstratives Divided by Number of Words. The results from the unabbreviated texts (M = .1, SD = .004) and texts abbreviated using Latent Semantic Analysis (M = .2, SD = .007) indicate that givenness measured by unattended demonstratives was not significantly different between the texts, t(40) = .1.01 , p = .33. The results of the text which had been abbreviated using simple lexical repetition (M= .1, SD= .005) also indicated that givenness was not significantly different between the abbreviated and unabbreviated texts, t(40) = .58, p = .56. Similarly, the other methods of abbreviation did not show significant effects compared to the full texts.

13.3 RQ3: Is There a Difference in the Coherence of L2 Student's Writing in a Pilot Trial of a Classroom-Based Test of Coherence When Responding to a Text Abbreviated Using Latent Semantic Analysis?

Three of the four participants who took part in the pilot study used more complex syntax measured by mean length of t-unit when writing a response to the source text which had been abbreviated using latent semantic analysis. Results for the test takers' use of formulaic language and sentence level cohesion was less clear. Half of the participants used more cohesively tied sentences with the abbreviated text. The amount of semantic overlap in the work of the participants who used less semantically cohesive sentences with the shorter text declined less than the increase in cohesiveness in the work of the other participants. This means that there was an aggregate increase in cohesion measured by semantic overlap between adjacent sentences. The strength of association of bigrams which appeared in the academic sub-corpus also sightly increased on aggregate.

Three of the four participants reported feeling less familiar with the topic of the abbreviated text. All participants reported first exposure to English between the age of 9 and 15. Participants were more varied in the age that they reported they had become fluent in English, ranging from 15

to 28. Another significant difference between the participants was the amount of time which they reported living in the UK, ranging from 10 months to 30 years. Participants all reported that talking with family was the least helpful task they undertook in their acquisition of the language, compared to reading which all participants reported had a much stronger beneficial effect on their acquisition. Results were mixed in the reported effectiveness of digital media, audio lessons, talking with friends and watching television. Most participants spent the majority of their reading time, reading in L2.

All participants reported frequently talking to friends and work colleagues in English and frequently listening to songs in English. Participants all also reported frequently reading professional emails and academic texts in English. Most participants did not report frequently watching television or films in English. Nor did most participants report frequently reading fiction or news and magazines in English. Two of the participants reported frequently watching digital media online and listening to podcasts. Participant results are summarised in tables 12 – 14.

Table 12:

Participant results for Inter-Sentence Cohesion

Participant	Cohesion full text	Cohesion abbreviated
Participant 1	0.3	0.2
Participant 2	0.2	0.5
Participant 3	0.4	0.4
Participant 4	0.2	0.4

Student ID : 21048826

Table 13:

Participant Mean Length of T-Unit

Participant	MLT full text	MLT abbreviated
Participant 1	12.6	18.9
Participant 2	18.1	27.8
Participant 3	29.0	19.8
Participant 4	15.4	22.4

Table 14:

Association Strength of Academic Bigrams in Participants' Responses

Participant	bigram MI full text	bigram MI abbreviated
Participant 1	1.5	1.6
Participant 2	1.8	1.6
Participant 3	1.3	1.8
Participant 4	1.5	1.3

14 Discussion

14.1 RQ1: Is There a Difference In The Coherence of L2 English Students' Writing in Genres Which Make Reference to External Sources Compared to Genres Which Do Not, As Represented in the British Academic Written English Corpus?

The first question in this study sought to determine if the use of sources affected variables which previous research had indicated might constitute a construct of coherence. It was hypothesised that participants who are engaged in inferential processing will make use of information stored in long-term memory and subsequently form a more coherent mental representation of the information they are reading. Consequently, it was hypothesized that this coherence in the students' mental representation would be reflected in their writing. Results

suggested that two out of the three variables operationalised to measure coherence showed significant differences when measured.

For the purpose of this study, the construct of coherence was operationalised as consisting of four elements: Sentence level cohesion of test taker writing; use of register-appropriate, formulaic language; syntax complexity; inferring from academic writing. Each of these elements will be discussed presently with reference to the results of this study, theoretical frameworks and previous, relevant research.

Prior studies have noted the importance of both reading and writing as constituent elements of the construct of coherence. With respect to reading, in the context of English medium instruction in higher education, this places very specific demands on test developers. Psycholinguistic theories in language comprehension suggest potential threats to validity may arise from underrepresenting the role of prior knowledge, or the differences in cognitive processes involved in text comprehension which are specific to L2 English students. Kintsch (2018) notes the importance of prior knowledge in text comprehension as a basis for constraint satisfaction when inferring the sense of a lexical item. Insufficient prior knowledge is therefore likely to result in less coherent mental representation of the information contained in a text. Research has also investigated the role prior knowledge plays in the inferential processing involved in language users' abilities to form coherent mental representations from texts altered to reduce their coherence (McNamara, 2001). However, research into the interwoven productive-receptive cognitive processing involved in L2 comprehension has demonstrated that L2 language users differ from L1 users, suggesting higher cognitive load placed on language users whose comprehending is less automatised (Ito, Pickering and Corley, 2018). This theory is consistent with recent eye-tracking research (Kuperman et al., 2022; Nisbet et al., 2022).

While the nature of the corpus precluded the ability to measure any parametric data from the sources which the examples of student essays reference, the corpus was organised according to the genre of the student writing. Comparing the student essays between genres which are likely or

unlikely to make reference to sources may therefore offer some insight into the relationship of source use and L2 student output.

The results of this study indicate that source use does indeed affect student output. This finding is largely consistent with previous research. Abrams (2019) reported increased syntactic complexity, grammatical accuracy, fluency, lexical accuracy, choice and richness in writing produced by intermediate L2 German learners at a US university. These participants were of a comparable age. The corpus used in the present study might have represented a broader range of L1s, though this is not reported in Abrams (2019). All students included in the British Academic Written English corpus are assumed to be above intermediate proficiency since they are studying content in L2. Consequently, one possible implication of this might be that these effects are relatively consistent across proficiency levels. This account must be approached with some caution because it is unclear whether these results are caused by cross linguistic influence, or cognitive load which might exist across L1 and L2 users.

The present study found significant effects on the mean length of t-unit when comparing source based writing (essays and reviews of literature) to explanations and case studies, though effect sizes were modest. Kim and Crossley (2018) measured similar indices using a structural equation modelling approach on a corpus of responses to both integrated reading-writing tasks, as well as independent writing tasks. The corpus used for the present study differed from that used in Kim and Crossley (2018) inasmuch as the examples of student writing included in the present study were not written under timed conditions. Timed conditions may emphasize the role of automaticity in productive fluency, therefore comparisons between the two studies must remain tentative. Nevertheless, Kim and Crossley (2018) found that mean length of clause correlated more strongly with human rater judgements than compared to mean length of t-unit. Further research is required for the appropriateness of this variable as a measure of coherence in a classroom based test.

Student ID : 21048826

Contrary to expectations, this study did not find a significant difference between the sentence level cohesion of students' writing between the selected genres. It has been suggested that sentence to sentence cohesion is an important indicator of student performance because it correlates strongly with human rater judgements of essay quality (Guo et al., 2013). That sentence level cohesion correlates with human judgements, but is not quantitatively different between genres may indicate that other variables are more closely related to the construct of coherence. Interestingly, Guo et al. (2013) found that the same variable used in this study (semantic similarity between sentences measured by Latent Semantic Analysis) was a significant predictors of human judgements of essay quality only in integrated writing tasks. There are two potential explanations for this inconsistency. It is possible that the sample used in the present study had a smaller range in the proficiency of language users, with less proficient language users being under-represented in the corpus. However, Gebril and Plakans (2013) found that cohesion played a critical role in predicting human judgements of the writing of more proficient university students in integrated, content responsive tasks. Another possible explanation for the inconsistency may be the effect of timed conditions, or that other variables would reveal a qualitative difference in how cohesion is expressed.

These results might suggest that the inferential processing involved in interpreting content written in an academic register will result in the sample of source based writing showing stronger association between words in bigrams which appear in the COCA academic written English sub-corpus. These results are consistent with previous research the effects of task complexity on source based writing (Golparvar & Rashidi, 2021). Longitudinal research indicates that the use of formulaic language in specialised registers is an important correlate of proficiency (Duan & Shi, 2021). The results of the present study might suggest that genre is a further variable which affects the use of formulaic sequences in L2. It is unclear from the present study whether the use of these formulaic sequences predicts human judgements of writing quality.

Student ID : 21048826

14.2 RQ2 Is There a Difference In Cohesion-Based Reading Ease Measures Between Texts Which Have Been Abbreviated Using Latent Semantic Analysis, Word2Vec, or Latent Dirichlet Allocation?

While the first research question in the present study aimed to establish if the use of sources in academic writing had an effect on variables which have been theorised to constitute the construct of coherence, it was not clear from the data available how the cohesion of source texts themselves might affect the inferential processes which a test of coherence may wish to elicit. Previous research has emphasized the role of sentence level cohesion of sources on task complexity in integrated reading writing tasks (Abrams, 2019; Bilki & Plakans, 2022; McNamara, 2001). Consequently, the second research question set out to investigate if it would be possible to maintain content specificity, syntactic complexity and formulaic sequences in the source text while making the text more cohesive.

A method of text abbreviation proposed in Hoey (1991) was applied to a corpus of academic texts. 3 machine learning algorithms, and one lemma based method was applied to measure the relatedness between all sentences in the text. The sentences which were more semantically tied to the other sentences in the text were selected for inclusion in the abbreviated texts. Finally, to measure the relative readability of the abbreviated texts, they were measured for three variables which research has suggested correlate with human ratings of reading ease (Crossley et al., 2008). The selected variables measured cohesion between every two adjacent sentences, and potential ambiguity which might arise from pronoun density and givenness.

Previous research has suggested that sentence level cohesion plays in important role in the processing strategies used by advanced second language users of English (Bilki & Plakans, 2022). Research has also indicated that the overlap of content words between two adjacent sentences more accurately predicts human judgements of readability for meaning construction when compared to traditional methods such as the Flesch reading ease score (Crossley et al., 2008). Consequently, the same variable which was used to measure cohesion for meaning construction in

Crossley et al. (2008) was used to evaluate machine learning methods' selection of "central" sentences.

The result of the present study show that Latent Semantic Analysis performed better than the other machine learning based methods of sentence selection. Though to my knowledge, no previous research has investigated this method of text abbreviation, Clarke and Lapata (2010) investigated text coherence finding that Latent Semantic Analysis was successful at modelling text coherence, but did not adequately account for sub-sentential cohesion. Consequently, the finding of this study add further evidence to prior research into applications of machine learning algorithms to investigate text coherence (Landauer & Dumais, 1997). The encouraging findings of this study must, nevertheless, be interpreted cautiously since they did not involve ratings of coherence by human raters.

14.3 RQ3: Is There a Difference in the Coherence of L2 Student's Writing in a Pilot Trial of a Classroom-Based Test of Coherence?

Finally, following an investigation into the constituent elements of how coherence might be theorised in the academic writing of L2 English students, and evaluating machine learning based methods of text abbreviation for sentence level cohesion, testing materials were developed for use with the target test taker population. To this end, two participants were recruited to take a test with the developed materials.

The results suggest that the participants used more complex grammatical structures when responding to an abbreviated text. This consistency between the participants ought to be interpreted cautiously due to the likelihood that the topic familiarity might have confounded these results. Furthermore, results for measures of cohesion and use of formulaic language were more ambiguous. Research on the use of formulaic language in L2 academic writing is inconsistent. Y. Chen and Baker (2016) propose that source-based integrated reading-writing tasks might inflate the likelihood that formulaic sequences are used by L2 writers. Conversely, Duan and Shi (2021) found

that strength of association between words in multiword sequences followed a u-shaped pattern over the course of two and a half years of learning in English majors studying in China. Evidently, further research is needed to investigate the relationships of the variables which affect this element of language use.

15 Limitations

The researcher acknowledges that, in spite of the use of rigorous methodology, there are still considerable limitations to this study. While the results presented in the present study go some way to developing a classroom based test of coherence in L2 writing, the time limitations associated with studying a course full time over the period of a year resulted in only a small sample size available to test the developed materials. Future studies may consider using a randomised controlled trial research design to measure the difference the abbreviated texts make used in a language testing context. Of particular interest would be how these abbreviation procedures affect performance in disciplines in the life and physical sciences. Whilst the method of abbreviation uses lexical information to establish connections between sentences in a text, how modalities such as mathematical formulas, tables and images are dealt with remains an important question to consider for language testing, and becomes especially pronounced in “hard” subjects according to the taxonomy proposed by Durrant (2017).

Furthermore, due to the increase in frequency of courses taught in English in higher education occurring predominantly in countries where English is spoken as a foreign language, future research using samples in these context would be especially valuable. The amount of input to which L2 English students enrolled in higher education courses in predominantly English speaking countries are exposed to is significantly larger when compared to students studying English as a foreign language. Consequently, samples ought to be taken from these populations to better reflect the performance of target test takers.

Student ID : 21048826

16 Conclusion

To conclude this dissertation, I will unpack a few pedagogical implications from the research and make recommendations for future research.

16.1 Pedagogical Implications

In the introduction of this dissertation, three areas of interest were articulated to held delineate the purpose of this research. To reiterate these areas of enquiry, they may be summarised as follows:

- The authenticity of tasks based on register, genre and discipline (Duan & Shi, 2021; Wang & Xie, 2022);
- The effect of source text cohesion on test taker performance (Bilki & Plakans, 2022; Cai & Chen, 2022);
- The effects of cultural and social differences on the use of academic language, (Cai & Chen, 2022; Duan & Shi, 2021).

Despite the findings of this study being largely inconclusive, some new information may be added to this research gap based on the insights found in the present research. There is now a stronger case to consider the use of formulaic language as an element of coherence, emphasising the pragmatic-communicative theoretical element of the construct. Nevertheless, the optimal variables through which this construct might be described by fine grained indices generated by automated essay scoring software remains an under-explored area for diagnostic language testing. This research might also tentatively suggest that using machine learning algorithms as a basis for determining which sentences ought to be included in an abridged text warrants further enquiry. Furthermore, building on what this study has suggested, how effective this theoretical approach to eliciting syllogistic inferencing from students remains inconclusive for samples where English is spoken as a foreign language, and with less topic knowledge.

Student ID : 21048826

The concurrent internationalisation of higher education and advances in automated essay scoring using natural language processing software present both non-trivial challenges arising from the newly emergent contexts, and new pedagogical opportunities to develop best practice. Content instructors responding to top-down policy changes are likely aware that their position within internationalising institutions is a position of considerable influence; it follows from this that their position in relation to students enrolled in courses where English is the medium of instruction is one of significant responsibility. It is possible that out of respect for these commitments, instructors who do not consider themselves to be experts in linguistics or language instruction consider delegating this responsibility to specialist teams as the optimal strategy to cater for the needs of learners who use English as a second language. Nevertheless, the intricately interwoven relationship of language knowledge and content knowledge suggests that close collaboration, and respect for the specificity of the conventions of discourse communities which make use of English in highly specialised registers is advisable on the part of applied linguists interested in addressing these increasingly common demands.

16.2 Summary

The main goal of the current study was to determine the feasibility of a subject specific diagnostic language test and a method of text abbreviation which might be applied as a intervention on the basis of diagnosis. Previous research suggested that an integrated test of reading and writing skills might be the most valid method of testing for this purpose

The dissertation was organised into three main research questions aimed at defining the construct of interest and validating its measurement through the use of a learner corpus, applying a novel method of text abbreviation on the basis of some insights from cognitive psycholinguistic theory, and piloting the application of the test and intervention.

The results of the first part of this investigation show that sentence level cohesion was not affected by source use in academic writing by the L2 English students represented in the British

Academic Written English corpus. Nevertheless, previous research suggests that this may be a significantly differentiating factor between learners, with less proficient learners being under-represented in the sample used for this study. Significant effects were found for syntax complexity and use of formulaic language between academic genres where source use is expected compared with genres where source use is less likely. These data suggest that coherence can be tested through integrated methods, though further research is necessary to establish which variables in student writing might provide the clearest representation of this construct.

In responding to the second research question, this study has shown that text abbreviation based on discourse features can be operationalised using the application of machine learning algorithms, accessible through freely available natural language processing software. The results of the study showed significant effects, with large effect sizes when modified texts were processed for reading ease measures based on sentence level cohesion. Variables relating to potential ambiguity and givenness did not return results which were significantly different from the unabbreviated texts. Though encouraging, these results must be interpreted carefully, as the quantitative measures used for this study do not exclude the possibility that qualitative features might affect student performance. For example, the amount of pronouns in an abbreviated text says little about how clear it is to what each of those pronouns refers. Nevertheless, the versatility of such an approach suggests it may be appropriate for content instruction in L2, where learners may depend on prior knowledge to develop robust mental representations of the content of the text as demonstrated in earlier research.

Finally, the study piloted testing materials on L2 English university students enrolled at a UK university. While the study found that both participants used more complex syntax when they made use of the abbreviated text, results for cohesion and formulaic language were less clear. Follow up questionnaire's suggested that language use and history had a significant influence on performance, consequently future research might consider the effects of reading habits of the test takers.

Student ID : 21048826

Although the current study is based on a small sample of participants, the findings lay the groundwork for future enquiry. Large randomised controlled trials could provide more definitive evidence regarding which variables are best suited for the purpose of a diagnostic test of coherence in L2 English student writing. A greater focus on the corpora used to train the machine learning algorithms could produce interesting findings that account more for the specificity of language use between academic disciplines. Further research also needs to examine more closely the relationship between learner proficiency and sentence level cohesion in their writing.

Student ID : 21048826

17 **References**

Abrams, Z. I. (2019). The effects of integrated writing on linguistic complexity in L2 writing and task-complexity. *System, 81*, 110-121.

Ahmed, A. H. (2010). Students' problems with cohesion and coherence in EFL essay writing in Egypt: Different perspectives. *Literacy Information and Computer Education Journal (LICEJ), 1*(4), 211-221.

Airey, J. (2020). The content lecturer and English-medium instruction (EMI): epilogue to the special issue on EMI in higher education. *International Journal of Bilingual Education and Bilingualism, 23*(3), 340-346.

Aizawa, I., & Rose, H. (2019). An analysis of Japan's English as medium of instruction initiatives within higher education: the gap between meso-level policy and micro-level practice. *Higher Education, 77*(6), 1125-1142.

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics, 36*(2), 236-260.

Alsop, S., & Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora, 4*(1), 71-83.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *tesol QUARTERLY, 16*(4), 449-465.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1): Oxford University Press.

Bilki, Z., & Plakans, L. (2022). Levels of textual cohesion and the meaning-construction process of advanced second language readers. *TESOL Journal*, e656.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Student ID : 21048826

Block, D., & Moncada-Comas, B. (2022). English-medium instruction in higher education and the ELT gaze: STEM lecturers' self-positioning as NOT English language teachers. *International Journal of Bilingual Education and Bilingualism, 25*(2), 401-417.

Bo, W. V., Fu, M., & Lim, W. Y. (2022). Revisiting English language proficiency and its impact on the academic performance of domestic university students in Singapore. *Language Testing*, 02655322211064629.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological review, 111*(4), 1061.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119-157.

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics, 32*(1), 13-47.

Cai, Y., & Chen, H. (2022). The fluctuating effect of thinking on language performance: New evidence for the Island Ridge Curve. *Language Assessment Quarterly*, 1-15.

Carrell, P. L. (1982). Cohesion is not coherence. *tesol QUARTERLY, 16*(4), 479-488.

Casal, J. E., Shirai, Y., & Lu, X. (2022). English verb-argument construction profiles in a specialized academic corpus: Variation by genre and discipline. *English for Specific Purposes, 66*, 94-107.

Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing, 33*(4), 453-472.

Chapelle, C. A. (2020a). *Argument-based validation in testing and assessment*: Sage Publications.

Chapelle, C. A. (2020b). Reflect, Revisit, Reimagine: Language Assessment in ARAL. *Annual Review of Applied Linguistics, 40*, 113-118.

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing, 32*(3), 385-405.

Chen, M., & Cui, Y. (2022). The effects of AWE and peer feedback on cohesion and coherence in continuation writing. *Journal of Second Language Writing*, 100915.

Student ID : 21048826

Chen, X., Qi, Y., Bai, B., Lin, Q., & Carbonell, J. G. (2011). *Sparse latent semantic analysis.* Paper presented at the Proceedings of the 2011 SIAM International Conference on Data Mining.

Chen, Y., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics, 37*(6), 849-880.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*(1), 22-29.

Clarke, J., & Lapata, M. (2010). Discourse constraints for document compression. *Computational Linguistics, 36*(3), 411-441.

Conklin, K., Alotaibi, S., Pellicer-Sánchez, A., & Vilkaitė-Lozdienė, L. (2020). What eye-tracking tells us about reading-only and reading-while-listening in a first and second language. *Second Language Research, 36*(3), 257-276.

Council_of_Europe. (2020). *CEFR Descriptors*. Council of Europe Retrieved from https://rm.coe.int/cefr-descriptors-2020-/16809ed2c7

Cronbach, L. J. (1971). Test validation. *Educational measurement*.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *tesol QUARTERLY, 42*(3), 475-493.

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods, 51*(1), 14-27.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly, 10*(1), 1-8.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621-694): American Council on Education.

Dafouz, E., & Smit, U. (2021). English-medium education revisited. *European Journal of Language Policy, 13*(2), 141-160.

Student ID : 21048826

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics, 14*(2), 159-190.

del Mar Sánchez-Pérez, M. (2021). Predicting content proficiency through disciplinary-literacy variables in English-medium writing. *System, 97*, 102463.

Dewaele, J.-M. (2018). Why the dichotomy 'L1 versus LX user'is better than 'native versus non-native speaker'. *Applied Linguistics, 39*(2), 236-240.

Doiz, A., & Lasagabaster, D. (2021). An analysis of the use of cognitive discourse functions in English-medium history teaching at university. *English for Specific Purposes, 62*, 58-69.

Duan, S., & Shi, Z. (2021). A longitudinal study of formulaic sequence use in second language writing: Complex dynamic systems perspective. *Language Teaching Research*, 13621688211002942.

Durrant, P. (2017). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics, 38*(2), 165-193.

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming.

Educational_Testing_Service. (2021). Performance Descriptors for the TOEFL iBT® Test. In: Educational Testing Service.

Ehrman, M. E. (1996). *Understanding second language learning difficulties*: Sage.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological review, 102*(2), 211.

Fenton-Smith, B., & Humphreys, P. (2015). Language specialists' views on academic language and learning support mechanisms for EAL postgraduate coursework students: The case for adjunct tutorials. *Journal of English for academic purposes, 20*, 40-55.

Gardner, S., & Nesi, H. (2013). A classification of genre families in university student writing. *Applied Linguistics, 34*(1), 25-52.

Student ID : 21048826

Gebril, A., & Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly, 10*(1), 9-27.

Gobet, F. (2000). Some shortcomings of long-term working memory. *British Journal of Psychology, 91*(4), 551-570.

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Golparvar, S. E., & Rashidi, F. (2021). The effect of task complexity on integrated writing performance: The case of multiple-text source-based writing. *System, 99*, 102524.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science, 14*(6), 1006-1033.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18*(3), 218-238.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. (2010). Ego depletion and the strength model of self-control: a meta-analysis. *Psychological bulletin, 136*(4), 495.

Halliday, M. A. K., & Hasan, R. (2014). *Cohesion in english*: Routledge.

Hoey, M. (1991). *Patterns of lexis in text* (Vol. 299): Oxford University Press Oxford.

Hyland, K., & Jiang, F. K. (2018). “In this paper we suggest”: Changing patterns of disciplinary metadiscourse. *English for Specific Purposes, 51*, 18-30.

Ito, A., Corley, M., & Pickering, M. J. (2018). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Bilingualism: Language and Cognition, 21*(2), 251-264.

Joint_Committee_on_Testing_Practices. (2000). Code of fair testing practices in education. In.

Kane, M. (2010). Validity and fairness. *Language Testing, 27*(2), 177-182.

Student ID : 21048826

Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing, 37*, 39-56.

Kim, M., Nam, Y., & Crossley, S. A. (2022). Roles of working memory, syllogistic inferencing ability, and linguistic knowledge on second language listening comprehension for passages of different lengths. *Language Testing*, 02655322211060076.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review, 95*(2), 163.

Kintsch, W. (2018). Revisiting the construction—integration model of text comprehension and its Implications for Instruction. In *Theoretical models and processes of literacy* (pp. 178-203): Routledge.

Kuperman, V., Siegelman, N., Schroeder, S., Acartürk, C., Alexeeva, S., Amenta, S., . . . Chernova, D. (2022). Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus. *Studies in Second Language Acquisition*, 1-35.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication.* (Applied Linguistics and English as a Second Language Doctoral Dissertation). Department of Applied Linguistics and English as a Second Language, Atlanta, Georgia, USA.

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior research methods, 50*(3), 1030-1046.

Labov, W. (1982). Objectivity and commitment in linguistic science: The case of the Black English trial in Ann Arbor. *Language in society, 11*(2), 165-201.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review, 104*(2), 211.

Student ID : 21048826

Langer, J. A. (1983). Effects of Topic Knowledge on the Quality and Coherence of Informational Writing.

Lanvers, U., & Hultgren, A. K. (2018). The Englishization of European education: foreword. *European Journal of Language Policy, 10*(1), 1-11.

Li, W., & McCallum, A. (2006). *Pachinko allocation: DAG-structured mixture models of topic correlations.* Paper presented at the Proceedings of the 23rd international conference on Machine learning.

Lin, L. H., & Morrison, B. (2021). Challenges in academic writing: Perspectives of Engineering faculty and L2 postgraduate research students. *English for Specific Purposes, 63*, 59-70.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics, 15*(4), 474-496.

Macaro, E., Curle, S., Pun, J., An, J., & Dearden, J. (2018). A systematic review of English medium instruction in higher education. *Language Teaching, 51*(1), 36-76.

McGrath, L., Negretti, R., & Nicholls, K. (2019). Hidden expectations: Scaffolding subject specialists' genre knowledge of the assignments they set. *Higher Education, 78*(5), 835-853.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 55*(1), 51.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse processes, 22*(3), 247-288.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher, 18*(2), 5-11.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Student ID : 21048826

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems, 26*.

Murray, N. (2010). Considerations in the post-enrolment assessment of English language proficiency: Reflections from the Australian context. *Language Assessment Quarterly, 7*(4), 343-358.

Nisbet, K., Bertram, R., Erlinghagen, C., Pieczykolan, A., & Kuperman, V. (2022). Quantifying the difference in reading fluency between L1 and L2 readers of English. *Studies in Second Language Acquisition, 44*(2), 407-434.

Oller, J. W. (1979). Language tests at school.

Orman Quine, W. v. (1976). Two dogmas of empiricism. In *Can theories be refuted?* (pp. 41-64): Springer.

Parke, M. R., Seo, M.-G., & Sherf, E. N. (2015). Regulating and facilitating: the role of emotional intelligence in maintaining and using positive affect for creativity. *Journal of Applied Psychology, 100*(3), 917.

Pecorari, D., & Malmström, H. (2018). At the crossroads of TESOL and English medium instruction. *tesol QUARTERLY, 52*(3), 497-515.

Peng, J.-E., & Xie, X. S. (2021). English-medium instruction as a pedagogical strategy for the sustainable development of EFL learners in the Chinese context: A meta-analysis of its effectiveness. *Sustainability, 13*(10), 5637.

Phakiti, A., & Isaacs, T. (2021). Classroom assessment and validity: Psychometric and edumetric approaches. *European Journal of Applied Linguistics and TEFL, 10*(1), 3-24.

Plakans, L. (2021). Writing integrated tasks. In *The Routledge Handbook of Language Testing* (pp. 357-371): Routledge.

Plakans, L., Liao, J.-T., & Wang, F. (2019). “I should summarize this whole paragraph”: Shared processes of reading and writing in iterative integrated assessment tasks. *Assessing Writing, 40*, 14-26.

Student ID : 21048826

Read, J. (2008). Identifying academic language needs through diagnostic assessment. *Journal of English for academic purposes, 7*(3), 180-190.

Rose, H., Curle, S., Aizawa, I., & Thompson, G. (2020). What drives success in English medium taught courses? The interplay between language proficiency, academic skills, and motivation. *Studies in Higher Education, 45*(11), 2149-2161.

Sawaki, Y., Quinlan, T., & Lee, Y.-W. (2013). Understanding learner strengths and weaknesses: Assessing performance on an integrated writing task. *Language Assessment Quarterly, 10*(1), 73-95.

Shohamy, E. (1990). Discourse analysis in language testing. *Annual Review of Applied Linguistics, 11*, 115-131.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487-512.

Tabari, M. A., & Wang, Y. (2022). Assessing linguistic complexity features in L2 writing: Understanding effects of topic familiarity and strategic planning within the realm of task readiness. *Assessing Writing, 52*, 100605.

The British Council, I. I. A., the University of Cambridge ESOL Examinations (2022). WRITING TASK 1: Band Descriptors (public version). In: IELTS.

Wang, Y., & Xie, Q. (2022). Diagnosing EFL undergraduates' discourse competence in academic writing. *Assessing Writing, 53*, 100641.

Weigle, S. C., Yang, W., & Montee, M. (2013). Exploring reading processes in an academic reading test using short-answer questions. *Language Assessment Quarterly, 10*(1), 28-48.

Wolfersberger, M. (2013). Refining the construct of classroom-based writing-from-readings assessment: The role of task representation. *Language Assessment Quarterly, 10*(1), 49-72.

Xi, J., & Lantolf, J. P. (2021). Scaffolding and the zone of proximal development: A problematic relationship. *Journal for the Theory of Social Behaviour, 51*(1), 25-48.

Student ID : 21048826

Xi, X. (2017). What does corpus linguistics have to offer to language assessment? *Language Testing, 34*(4), 565-577.

Xie, Q., & Lei, Y. (2022). Diagnostic assessment of L2 academic writing product, process and self-regulatory strategy use with a comparative dimension. *Language Assessment Quarterly, 19*(3), 231-263.

Xu, T. S., Zhang, L. J., & Gaffney, J. S. (2022). EXAMINING THE RELATIVE EFFECTS OF TASK COMPLEXITY AND COGNITIVE DEMANDS ON STUDENTS'WRITING IN A SECOND LANGUAGE. *Studies in Second Language Acquisition, 44*(2), 483-506.

Yoon, H.-J. (2021). Challenging the connection between task perceptions and language use in L2 writing: Genre, cognitive task complexity, and linguistic complexity. *Journal of Second Language Writing, 54*, 100857.

Student ID : 21048826

18 Appendix 1: Original Dissertation Proposal

STUDENT NAME: Pasha Blanda

PROGRAMME: Applied Linguistics

Provisional Title of the Dissertation:

Sequencing input and automated essay scoring: local cohesion for domain specific coherence

Research area

Language testing and assessment

1. Research aims and rationale

This dissertation will examine second language (L2) teachers' application of automated essay scoring as assessment for learning to aid learners' development of mental models of coherence by sequencing input on a sentence level.

289 expert participants identify 'The effectiveness of multilingual support in regular school lessons' (Duarte et al., 2020, p. 1) as the most immediate research priority in a European context. The implications are complex in globalized, superdiverse conditions, rendering it increasingly difficult to establish traditional bilingual models (such as two-way immersion) for specific groups sharing a common family language or to establish pull-out programmes. (Duarte et al., 2020, p. 11). They continue that 'prioritising research in which didactic approaches are closely related to the ways in which languages are used by multilingual pupils' (Duarte et al., 2020, p. 11).

The use of content and language integrated learning (CLIL) has become increasingly common in the European context over the last decade (Goris, Denessen & Verhoeven, 2019). The effectiveness of learning a subject in L2, however, can result in a sacrifice in both language and content acquisition if administered poorly (Paran, 2013). Somers (2018) suggests CLIL has contributed to social inequality for immigrant and minority language families.

The construction-integration model, first proposed by Kintsch (1988), offers one direction of research for finding effective ways of supporting CLIL teachers and students. Two studies report a relationship between text cohesion and domain knowledge comprehension (McNamara &

Kintsch,1996; McNamara et al., 1996). Research in monolingual acquisition reports developmental trends in scrambled story recall (Fitzgerald & Spiegel, 1983), type of syntactic connection processed (McCutchen & Perfetti, 1982) and cohesion in text production (Yde & Spoelders, 1985). The separate nature of the concepts of cohesion and coherence, as emphasized by Carrell (1982), makes their interface all the more pertinent for this study.

Nevertheless, it is still unclear whether the sequencing of input in L2 will help learners build a more coherent mental model. Using an analytical system proposed by Hoey (1991) and adapted by Macmillan (2006), it may be possible to quantify the connectedness of individual sentences within a given text. Will sequencing input, with the most densely connected sentences being introduced to learners before less cohesive sentences, have an impact on the coherence of the mental model which the learners develop of that text?

'A good goal for the next decade would be to cull existing knowledge about commonly used configurations of test methods… this project would lay groundwork sorely needed for exploring the new test method characteristics made available through the use of technology' Chapelle (2020, p. 116). Indeed, the propositions made by Larsen-Freeman (1997) pose serious questions about applying assessment models based on linear regression to complex systems such as learner interlanguage. These questions are complicated further in classroom settings where teachers do not have the time or resources to develop elaborate edumetric methods.

Crossley, Kyle, & Dascalu (2019) offers one solution in the form of the Tool for the Automatic Analysis of Cohesion (TAACO 2.0) and Coh-Metrix. The latter has been applied to formative assessment (Wilson, Roscoe, & Ahmed, 2017) indicating the tools' suitability for this study. The proposed investigation is a small pilot study which would look for constructs in the learners' output as evidence of a more coherent mental model developed from the input received by the treatment group compared to a control.

2. *Provisional research questions*

Student ID : 21048826

Will participants exposed to input sequenced according to strength of local cohesion produce more globally cohesive texts compared to a control group?

Does a more coherent mental model mean that students will produce more cohesive texts?

3. *Brief literature review*

The four most pertinent studies for the presently proposed research are Wilson, Roscoe, & Ahmed (2017); Crossley, Kyle, & Dascalu (2019); Kintsch (1988); MacMillan (2006).

Firstly, to summarize the papers' contributions to the field of enquiry into coherence and cohesion, Kintsch (1988) elaborates a model of entirely bottom-up processing for written discourse. The model proposes a two-stage process consisting of construction from atomized textual elements and then being integrated into a knowledge network. The research is especially important for later applications due to the 'context-free process of activation of the closest neighbors of the original text-derived proposition in the general knowledge net' (Kintsch, 1988, p. 180). MacMillan (2006) investigates the domain validity of the TOEFL Reading Comprehension section by applying an adapted analytical system for measuring cohesion different from Kintch's (1988) model. The analytical framework is applied to examine inter-text cohesion between passages and the comprehension questions they are related to. The research then describes the type of lexical link between the passage and correct answer.

Secondly, I summarize the main findings of the two studies concerned with automated essay scoring. Wilson, Roscoe, & Ahmed, (2017) concerns the validity of the automated essay grading tool Coh-Metrix. They form a hypothesis from a tri-level model of writing competence (word, sentence, and discourse) and use multigroup confirmatory factor analysis to evaluate their prediction. Then, the researchers apply multigroup structural equation modelling to predict human ratings of writing competence for high-school essays submitted to a state wide test. Crossley, Kyle, & Dascalu's (2019) first experiment finds that word2vec is an important predictor of coherence ratings, having considered the effects of four factors on expert ratings of texts. The variables which the paper is interested in are: cohesion features, prompt, essay elaboration, and enhanced cohesion. The second

study reported on the extent to which source overlap between speaking samples and responses predicted human ratings of proficiency.

The selection of these papers was made for the following reasons. Kintsch (1988) is especially relevant to the application of automated essay scoring because algorithmically driven corpus analytical tools like latent semantic analysis (LSA) and word2vec analyses are a direct application of the implications of the construction-integration model. The relationship of cohesion and domain specific knowledge is also especially pertinent to CLIL settings because decision making regarding the processing of text to form mental representations of its content can be regulated for optimal impact by teachers. Wilson, Roscoe, & Ahmed, (2017) argue their research provides proof of concept for the validity of using automated essay scoring for formative writing assessment. They emphasize that formative assessment differs from summative assessment in that it aims 'to provide valid and nuanced information about distinct writing skills that are meaningfully related to outcomes of interest' (Wilson, Roscoe, & Ahmed, 2017, p. 31). Furthermore Coh-Metrix is a web-based tool and TAACO 2.0 works on the most commonly run operating systems, both are free to use making them especially accessible for teachers.

MacMillan's (2006) analytical model can be adapted to describe the internal coherence of texts. Each sentence can be measured in terms of how connected it is to other sentences in that text, and these sentences can be arranged so that more bonded sentences are presented to learners before less bonded ones. I hypothesizethat doing so will develop a more coherent mental model for learners and be visible in their output.

Whilst prompt topic and prompt genre tend to limit the generalizability of writing assessments , the former is mitigated by the number of prompts used by participants in Wilson, Roscoe, & Ahmed, (2017). Nevertheless, the presently proposed investigation aims to elicit discourse produced for informational as opposed to persuasive genres and use a more authentic approach providing an opportunity for participants to re-draft. Furthermore, while the paper argues for its applicability as a tool for formative assessments, its use of corpora as analytical input limits its

domain validity in a classroom context. Crossley, Kyle, & Dascalu (2019) report that the original version of TAACO has been applied in published studies relating to creativity, transcription disfluency, literary studies, formative writing assessment, predicting math performance, self-regulated learning, and medical discourse. Nevertheless, the compatibility of TAACO 2.0's use of Stanford Core natural language processing for word2vec analysis with versions of operating systems beyond Windows 10 means that if these issues are unresolvable, then Coh-Metrix is more suitable for teacher use as a web-based tool.

Proposed methodology

The presently proposed research would use a quasi-experimental, quantitative approach. The research would be conducted at the scale of a pilot study for a cross-sectional investigation. The methods for the research design are aimed at investigating associational relationships between participants, with two variables of particular interest. The independent variable would be the exposure of a treatment group to L2 input, adapted either from the British Academic Written English Corpus (BAWE) or McNamara et al., (1996) depending on the practicality and labor-intensity of pre-experiment input-coding. Each sentence in the input will be coded for internal cohesion, with the most bonded sentences being shared with participants in the treatment group before less cohesive sentences. Each sentence will be exposed to participants using a grammar dictation procedure first proposed by Wajnryb (1990). Participants in the control group would read an un-adapted version of the same text. Both groups would then have the opportunity to discuss and draw a map of their ideas before writing a short informational text explaining the domain specific knowledge present in the input. Any orthographic mistakes will be corrected in the output by the researcher.

The dependent variable is the score the output will receive when analysed for either TAACO 2.0's Word2Vec indices or Coh-Metrix' LSA indices, depending on TAACO 2.0's compatibility with Windows 11 and above. Finally, the participants will be rated for global coherence and proficiency by two expert raters using the framework used by Crossley, Kyle, & Dascalu (2019) to establish inter-rater reliability. Below is a visual representation of the proposed sequence of the study.

Student ID : 21048826

Group	Session 1			
Treatment	Domain knowledge battery	Language learning history questionnaire	Exposure to pre-test input	Pre-test Output
Control	Domain knowledge battery	Language learning history questionnaire	Exposure to pre-test input	Pre-test Output

Group	Session 2		
Treatment	Exposure to sequenced post-test input	drafting discussion	Post-test output
Control	Exposure to un-sequenced post-test input	drafting discussion	Post-test output

Student ID : 21048826

A validity argument for in class assessment ought to relate to what decisions are made on the basis of the elicited data. One potential application of automated essay scoring may be as an assessment tool revealing whether a student would benefit more from a less or more cohesive input. More cohesive input would be used for less advanced students to help them access domain specific knowledge, while using less cohesive input would help students with stronger domain knowledge by forcing higher level inferential processing as opposed to lower-level linguistic decoding. If the decision to be made by a teacher is whether the student ought to be helped through adapted input (assuming that student output is reflective of their constructed mental model of the input) then the location of that threshold would determine criterion validity. Nevertheless, the presently proposed study aims to be a step towards establishing where that criterion threshold (determining whether a student has an adequately developed mental representation of the input) might be. The small scope of the study precludes conclusive findings, but as a pilot it may yield fruitful directions for future enquiry.

To investigate possible confounds arising from the coherence-cohesion gap, a transcription of the participants interactions during the post-input discussion will be coded for language related episodes (LREs) according to the model proposed by Revesz (2011) by two expert raters to establish inter-rater reliability. Similarly, incidence of content words from the input will be calculated using LSA indices. This data will be compared with LSA or word2vec global cohesion indices to establish any correlative relationships.

In order to control for individual differences which might manifest moderating variables, a biographical questionnaire adapted from Ehrman (1996) will aim to elicit data on Language Background, Language Learning Experience, & Proficiency Level. Despite significantly limiting the generalizability of the findings, due to constrains on time, the participant sample will be small and the research will involve cluster sampling of already existing language classes in English language schools in London. Issues regarding moderating variables in the gap between input and participant output will be elaborated in more detail in the final study, but the validity of using such a tool as a

classroom-based test for formative assessment purposes arguably justifies some sacrifice in this regard.

4. *Data collection and analysis*

Data will be collected over Zoom from 8 participants, with 2 separate hour-long sessions being recorded for transcription and analysis. Participants will complete anonymized questionnaires through SurveyMonkey and submit their written work anonymously to a cloud storage space online. Participants will be given a code in order to identify the corresponding questionnaires to the participants' output. This method is appropriate because it facilitates both written and spoken data for collection in a format that will be easily convertible to be analysed by the essay scoring tools.

The analytic process of these tools has shown significant concurrent validity with human raters (Crossley, Kyle, & Dascalu, 2019, p. 20). The automatic essay scoring tool functions both in terms of quantifying semantic similarity features, and key word overlap features. The former is processed in two stages. The first of these stages comprises three unsupervised learning methods: latent semantic analysis, latent Drichlet allocation, and word2vec.

It is LSA and word2vec which are arguably of most importance to the presently proposed research. They are analogous to Kintsch's (1988) construction-integration model in that they also build from a bottom-up interpretive framework using proximity in the input before integrating semantic similarity, providing data on indices compatible with Kintsch's (1988) knowledge network. This similarity makes it especially interesting to consider in CLIL contexts where cohesion and high domain specific knowledge are interfaced for teachers and learners. While word2vec was the strongest predictor of human judgement reported by Crossley, Kyle, & Dascalu (2019), unsupervised computational learning methods in general have only been able to provide mixed results in this regard. The question of how these might be applied effectively in classroom contexts is arguably more salient than whether they ought to be applied at all.

5. *Ethics*

Student ID : 21048826

The considerations for the ethics of this study are taken from the 4th edition guidance document published by the British Association of Applied Linguistics (BAAL, 2021). The proposed study involves human participants. Consequently, there is a need to obtain informed consent and a duty to maintain confidentiality and anonymity. There is no problem posed by the participants knowing the purpose of the study and they will be made aware of this in emails eliciting participation. A consent form detailing that all data collected will be anonymized, and kept in online cloud storage until it passes to UCL will be signed by participants.

The participants will also be informed that their talk in the Zoom session will be transcribed and kept in cloud storage. The recordings of the Zoom sessions will be destroyed within the same week that the session occurs. The participants will be made aware that the data for this study will not be published and will only be shared with relevant tutors at UCL. Though it may limit the generalizability of the study's findings to the age of CLIL students, the participants will all be over the age of 16 in order to avoid ethical complications involving seeking the consent of parents.

Great care will be taken to accurately present the authority with which a researcher operates in the context in order not to overstate the researcher's status and power. Doing so will prevent risk of exploitation or personal disclosure due to undue deference.

6. *Time-line*

7/3/2022	submit re-drafted dissertation proposal
14/3/2022	Draft ethics approval form
21/3/2022	Submit ethics approval form
28/3/2022	Send emails to English language schools in London asking for participants
4/4/2022	Code selected input for local cohesion
11/4/2022	Administer Session 1 of the research
18/4/2022	Administer Session 2 of the research
25/4/2022	Code student out-put

Student ID : 21048826

9/5/2022	Calculate ANOVA of participant scores
16/5/2022	First draft of Introduction
23/5/2022	First draft of literature review
30/5/2022	First draft of method section
6/6/2022	First draft of results section
13/6/2022	First draft of Discussion and conclusion

Student ID : 21048826

7. *References in Proposal*

BAAL (2021). Recommendations on Good Practice in Applied Linguistics. 4th Edition. Available at www.baal.org.uk

Carrell, P. (1982). Cohesion is not coherence. *TESOL Quarterly, 16*(4). 479-488. https://doi.org/10.2307/3586466

Chapelle, C. A. (2020). Reflect, revisit, reimagine: Language assessment in ARAL*. Annual Review of Applied Linguistics 40*. Pp. 113–118. https://doi.org/10.1017/S0267190520000021

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods, 51*(1). Pp. 14-27. https://doi.org/10.3758/s13428-018-1142-4

Duarte, J., García-Jimenez, E., McMonagle, S., Hansenc, A., Gross, B., Szelei, N., Pinho A.S. (2020). Research priorities in the field of multilingualism and language education: A cross-national examination. *Journal of Multilingual and Multicultural Development*. Pp. 1-16. https://doi.org/10.1080/01434632.2020.1792475

Ehrman, M. E. (1996). *Understanding second language learning difficulties*. Thousand Oaks, CA: Sage. https://dx.doi.org/10.4135/9781452243436

Fitzgerald, J., & Spiegel, D. L. (1983). Enhancing children's reading comprehension through instruction in narrative structure. *Journal of Reading Behavior, 15*. Pp. 1–17. https://doi.org/10.1080/10862968309547480

Goris, J., Denessen, E., & Verhoeven, L. (2019). Effects of content and language integrated learning in Europe: A systematic review of longitudinal experimental studies. *European Educational Research Journal, 18*(6). Pp. 675-698. https://doi.org/10.1177/1474904119872426

Hoey, M. (1991). *Patterns of lexis in text*. Oxford: Oxford Univ. Press.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95.* Pp. 163-182.

Student ID : 21048826

Larsen-Freeman, D. (1997). Chaos/complexity science and second language acquisition. *Applied Linguistics, 18.* 141–165. https://doi-org.libproxy.ucl.ac.uk/10.1093/applin/18.2.141

MacMillan, F. M. (2006). Lexical patterns in the reading comprehension section of the TOEFL® test. *Revista do GEL 3.* Pp. 143–172.

McCutchen, D., & Perfetti, C. A. (1982). Coherence and connectedness in the development of discourse production. *Text—Interdisciplinary Journal for the Study of Discourse, 2*. Pp. 113–140. https://doi.org/10.1515/text.1.1982.2.1-3.113

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes, 22.* Pp. 247–288. https://doi.org/10.1080/01638539609544975

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14.* Pp. 1–43. https://doi-org.libproxy.ucl.ac.uk/10.1207/s1532690xci1401_1

Paran, A. (2013). Content and language integrated learning: Panacea or policy borrowing myth? *Applied Linguistics Review, 4*(2). Pp. 317-342. https://doi.org/10.1515/applirev-2013-0014

Revesz, A. (2011). Task complexity, focus on L2 constructions, and individual differences: A classroom-based study. *The Modern Language Journal, 95* (S1). Pp. 162-181.

Somers, T. (2018). Multilingualism for Europeans, monolingualism for immigrants? Towards policy-based inclusion of immigrant minority language students in content and language integrated learning (CLIL). *European Journal of Language Policy, 10*(2). Pp. 203-228. http://dx.doi.org/10.3828/ejlp.2018.12

Wajnryb, R. (1990). *Grammar Dictation*. Oxford: Oxford University Press.

Wilson, J., Roscoe, R., & Ahmed, Y. (2017). Automated formative writing assessment using a levels of language framework. *Assessing Writing, 34*. Pp. 16–36. https://doi.org/10.1016/j.asw.2017.08.002

Student ID : 21048826

Yde, P., & Spoelders, M. (1985). Text cohesion: An exploratory study with beginning writers. *Applied Psycholinguistics, 6*. Pp. 407–415. https://doi.org/10.1017/S0142716400006330

Student ID : 21048826

19 Appendix 2: Participant Language Use and Proficiency Questionnaire

How familiar were you with the topic discussed in the text which you read?

1 – 10

How often do you speak English with your family or friends?

- Every day
- A few times a week
- About once a week
- A few times a month
- Once a month
- Less than once a month

How often do you speak English at work/university?

- Every day
- A few times a week
- About once a week
- A few times a month
- Once a month
- Less than once a month

How often do you write in English to friends/on the web?

- Every day
- A few times a week
- About once a week
- A few times a month
- Once a month
- Less than once a month

Student ID : 21048826

How often do you watch TV programmes in English?

- Every day
- A few times a week
- About once a week
- A few times a month
- Once a month
- Less than once a month

How often do you watch films in English?

- Every day
- A few times a week
- About once a week
- A few times a month
- Once a month
- Less than once a month

How often do you watch videos on YouTube (or other similar sites) in English?

- Every day
- A few times a week
- About once a week
- A few times a month
- Once a month
- Less than once a month

How often do you listen to Podcasts/audiobooks/radio/etc in English?

- Every day
- A few times a week

Student ID : 21048826

- About once a week
- A few times a month
- Once a month
- Less than once a month

How often do you listen to songs in English?

- Every day
- A few times a week
- About once a week
- A few times a month
- Once a month
- Less than once a month

How often do you attend meetings/lectures/classes in English?

- Every day
- A few times a week
- About once a week
- A few times a month
- Once a month
- Less than once a month

How often do you read at work/university in English? (emails, academic articles, etc)

- Every day
- A few times a week
- About once a week
- A few times a month
- Once a month

Student ID : 21048826

- Less than once a month

How often do you read fiction or non-fiction books in English?

- Every day
- A few times a week
- About once a week
- A few times a month
- Once a month
- Less than once a month

How often do you read articles and material available on the web in English?

- Every day
- A few times a week
- About once a week
- A few times a month
- Once a month
- Less than once a month

How often do you read newspapers and magazines in English?

- Every day
- A few times a week
- About once a week
- A few times a month
- Once a month
- Less than once a month

What English language proficiency test did you take before beginning your studies?

What score did you receive on your English language proficiency test?

Student ID : 21048826

Please rate your proficiency in speaking English:

1 – 10

Please rate your proficiency in reading English:

1 – 10

Please rate your proficiency in understanding spoken English:

1 – 10

Please rate your proficiency in writing English:

1 – 10

If you have lived in the UK, how long have you lived in the UK? How many years and months?

At what age did you begin reading in English?

At what age did you become fluent in English?

When you were a secondary school student, what language was used at the school?

Talking with friends

1 – 10

Talking with family

1 – 10

Reading

1 – 10

Language CDs, Audio lessons etc.

1 – 10

Student ID : 21048826

Digital media (YouTube, Podcasts etc.)

1 – 10

Watching TV

1 – 10

Right now, what percentage of the time are you exposed to English?

1 – 10

When you read something, what percentage of the time do you read in English?

1 – 10

Student ID : 21048826

20 Appendix 3: Test Piloting Questionnaire

How familiar are you with the topic of the passage?

- Not at all familiar
- 1
- 2
- 3
- 4
- 5
- Very familiar

How clearly do the instructions explain the purpose of the text which you are being asked to write?

- Not clear at all
- 1
- 2
- 3
- 4
- 5
- very clear

How clearly do the instructions explain the intended audience for the text which you are being asked to write?

- Not clear at all
- 1
- 2
- 3
- 4

Student ID : 21048826

- 5
- Very clear

How much time are you likely to spend focusing on understanding unfamiliar words when reading the text?

- No time
- 1
- 2
- 3
- 4
- 5
- All of my time

How likely are you to draw on background knowledge when writing your answer?

- Not likely
- 1
- 2
- 3
- 4
- 5
- Very likely

How often will you monitor for comprehension when reading the text?

- Never
- 1
- 2
- 3

Student ID : 21048826

- 4
- 5
- More than once a sentence

How much time do you estimate you may devote to rereading the text?

- None of my time
- 1
- 2
- 3
- 4
- 5
- All of my time

How difficult do you think summarizing the main points from this text will be?

- Impossible
- 1
- 2
- 3
- 4
- 5
- Too Easy

How explicitly do the instructions explain the intended purpose of the reading texts?

- There is no explanation
- 1
- 2
- 3

Student ID : 21048826

- 4
- 5
- There is enough explanation

How clearly do the instructions explain what a student should do if they wish to directly copy phrases from the text?

- Not clearly at all
- 1
- 2
- 3
- 4
- 5
- Very clearly

How clearly do the instructions explain the criteria on which test takers will be evaluated?

- Not clearly at all
- 1
- 2
- 3
- 4
- 5
- Very clearly

How appropriate is the time frame (20 minutes) for reading and writing a 300 word response to the task?

- Unachievable
- 1

Student ID : 21048826

- 2
- 3
- 4
- 5
- Too easy

Do you have any further feedback regarding this draft of the test?