BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

RESEARCHING LEXICAL THRESHOLDS AND LEXICAL PROFILES ACROSS THE COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES (CEFR) LEVELS ASSESSED IN THE APTIS TEST

AR-G/2021/1

Dr Nathaniel Owen, Dr Prithvi Shrestha and Professor Stephen Bax
Open University

ARAGs RESEARCH REPORTS ONLINE
SERIES EDITORS: RICHARD SPIBY AND CAROLYN WESTBROOK
ISSN 2057-5203

DEDICATION

Professor Stephen Bax
1960–2017

We dedicate this work to Professor Stephen Bax,
who continues to inspire colleagues and the students he worked with
during the course of his career.

ABSTRACT

This project uses automated analysis software (www.textinspector.com) to research the lexical and metadiscourse thresholds, and lexical and metadiscourse profiles, of test-takers' writing in the British Council's Aptis Writing test, benchmarked to the Common European Framework of Reference for Languages (CEFR). Large quantities of Aptis writing responses (n=6,407), representing 65 countries, together with their score data, were analysed in terms of their use of lexis and metadiscourse. Measures and datasets used in the analysis include standard readability measures, the British National Corpus, the Corpus of Contemporary American English, English Vocabulary profile, the Academic Word List, and a bespoke corpus of metadiscourse markers. The purpose of the research is to enhance the validation argument for the Aptis test through large-scale profiling of candidates' writing performance.

The findings reveal that the Aptis writing test provides evidence that lexical complexity changes systematically as the CEFR level of learners increases. Of the 110 Text Inspector metrics used in the study, 26 metrics were significant across all CEFR boundaries, including measures of text length (sentence, token and type count), and metrics of lexical sophistication (syllable count and number of words with more than two syllables). Fourteen of the 26 metrics represent vocabulary use. One metric of text complexity (voc-d) was also significant across all thresholds.

The study also explores the utility of these metrics for use in an automated scoring engine. Twenty metrics were used to build an ordinal logistic regression which was trained on a stratified subset of the data. This model was then used to predict the CEFR band of a testing subset which held nationality data constant. The data revealed that lexical use metrics from the Cambridge Learner Corpus (CLC) were the most successful at identifying CEFR level, and the model was most successful in identifying A1 and C-level responses. However, the model failed to accurately differentiate A2, B1 and B2 responses, suggesting that other, organisational variables play a significant role in human judgements, which are not accounted for in this study. The paper concludes with recommendations for rater training on the basis of the findings.

Authors

Nathaniel Owen is an Associate member and former Research Associate at the Open University. He holds a PhD in language testing from the University of Leicester and has published articles in peer-reviewed journals and book chapters in edited volumes on subjects including language testing, research methods and widening participation in higher education. He has experience of teaching English as a foreign language in Spain, the UK and Australia, with expertise in teaching EAP and exam preparation courses. He has previously worked for the examination board, Cambridge Assessment and has managed international-funded research projects with Educational Testing Service, in addition to the British Council. He joined Oxford University Press in February 2020 as a Senior Research and Validation Manager.

Prithvi N. Shrestha, an award-winning author (British Council ELTons finalist 2019), is Senior Lecturer in English Language at the Open University, UK. He has led or co-led a number of funded research projects. He has published over 40 research outputs, including one research monograph (*Dynamic Assessment of Students' Academic Writing*, Springer, 2020) and an edited volume, covering academic writing assessment in distance education, language assessment, English language education in developing countries, English medium instruction and mobile learning. His research is informed by Systemic Functional Linguistics and sociocultural theory.

Professor Stephen Bax was a professor of Modern Languages and Linguistics, and Director of Research Excellence in the School of Languages and Applied Linguistics, The Open University (UK). His research included work on the application of new technologies in language education, and he was awarded the 2014 TESOL Distinguished Researcher Award for his research using eye tracking technology to investigate reading. He was responsible for developing Text Inspector, an online tool for analysing lexis in text, which was shortlisted for the international British Council ELTons awards for Digital Innovation 2017. Stephen was the PI of this project until he sadly passed away in 2017. We express our condolences to his friends and family. We all miss him terribly.

ACKNOWLEDGEMENTS

We wish to thank the anonymous referees, Richard Spiby and Carolyn Westbrook for their many useful suggestions, and Dana Therova for her assistance in data management in the preparation of this manuscript.

CONTENTS

List of tables

List of figures

1. BACKGROUND

This study falls under the Aptis 2016 Call for research proposals, in the category of *Test Development and Validation*, specifically:

> Studies investigating the usefulness of applying automated analysis techniques to investigate lexical thresholds and lexical profiles across the Common European Framework of Reference for Languages (CEFR) levels assessed in Aptis.

In line with this category, this report details a large-scale investigation of the value of automated analyses, using the advanced TextInspector.com tool together with a concordancing tool, to research the lexical and metadiscourse thresholds, and lexical and metadiscourse profiles, of test-takers' writing in the British Council's Aptis writing test, benchmarked to the Common European Framework of Reference for Languages (CEFR).

1.1 Rationale

Major language exam boards have carried out, or funded, extensive research into the lexical thresholds and profiles of parts of their tests. For example, Read and Nation (2002), and Seedhouse, Harris, Naeb and Üstünel (2014) looked at aspects of the lexical profiles of the IELTS speaking test. Khalifa and Schmitt (2010) examined lexis in Cambridge Main Suite reading papers. O'Loughlin (2013) examined aspects of the lexical profiles of candidates in three different written production tasks within the Pearson Test of Academic English. Weir (2014) produced a lengthy analysis of a large set of lexical measures used by candidates in the Eiken TEAP test in Japan. Bax (2015) successfully used the Text Inspector tool to analyse lexical profiles in Cambridge Main Suite reading exams. The reason for this focus on lexis is that lexical proficiency, and the lexical profile of a candidate, is acknowledged to play a particularly significant role in comprehension – indeed some have claimed that vocabulary is "the major factor" in reading comprehension (Laufer and Ravenhorst-Kalovski 2010, p. 26).

However, lexical profiling in learner writing remains an under-addressed area. This paper builds upon these earlier studies by expanding the scope of investigation to a large corpus of learner written data, and it considers a much greater number of possible metrics which might contribute towards establishing criterial features across the CEFR. Exam boards consider it essential when constructing a validity argument to develop a solid and large-scale research base of evidence concerning the developing use of lexis by candidates in the productive skills, and to benchmark these as far as possible to external criteria such as the CEFR. Research into the area of lexical thresholds and profiles in Aptis (both reading and writing) is currently limited, so it is timely at this stage to seek to fill that research gap. Additionally, this report also considers which lexical measures might be beneficial for developing an automated assessment tool using machine learning techniques.

1.2 Aims

This report offers a contribution to the Aptis research base with respect to candidates' developing lexical profiles across the levels examined by the Aptis writing test. A large number of Aptis writing responses from a multinational test-taker cohort, together with their score data, were analysed in terms of their use of lexis and metadiscourse, with a view to enhancing the validation argument for the Aptis test through large-scale profiling of candidates' lexical use in the Aptis writing test (Parts 2, 3 and 4).

The research examines candidates' lexical diversity, range and lexical sophistication using measures and datasets which include standard readability measures, the British National Corpus, the Corpus of Contemporary American English, English Vocabulary Profile (EVP), the Academic Word List (AWL), and a selected list of metadiscourse markers. The report also considers how this data might be employed in the development of an automated assessment tool.

2. REVIEW OF LITERATURE

2.1 Lexical and metadiscourse thresholds and profiles

This project is concerned with investigating *lexical profiles* (LPs) and *lexical thresholds* (LTs) in test-taker writing which can potentially assist in improving the context validity (Weir, 2005) of the British Council Aptis test. LPs describe language users in terms of their lexical deployment at a particular stage of developing language proficiency expressed in terms of particular, predefined levels, such as those expressed in the *Common European Framework of Reference for Languages* (CEFR) (Council of Europe, 2001) and the new CEFR companion volume (Council of Europe, 2018). LPs contain four key characteristics: they incorporate quantitative metrics of lexical deployment; are empirically derived; benchmarked to an established framework; and strive to be comprehensive. Lexical deployment refers to lexical features which the language users in question can command, expressed in quantifiable terms. LPs should have as strong an empirical research base as possible, deriving from the analysis of appropriate large datasets of linguistic evidence. They need to be benchmarked to an established and widely-recognised framework of language knowledge and use (e.g., the CEFR), to cohere with other dimensions of language proficiency and use, and to ensure utility. A comprehensive LP strives to incorporate all metrics which show sufficient sensitivity to differences across the levels of the framework. A complete LP would incorporate writing, speaking, reading and listening. However, this report concerns efforts to develop LP for learner written English in the Aptis test.

Lexical thresholds describe the boundaries between levels of the framework. As LPs describe language deployment at particular stages of development, the LTs describe the transition from one level to another. Like LPs, LTs also contain the four key characteristics of being described in terms of: quantitative metrics of lexical deployment; empirically-derived from a large dataset of learner language; benchmarked to an established framework; and strive to be comprehensive by describing all metrics for which clear boundaries emerge between levels.

2.2 Investigating text using automated analytical tools

In recent years, analytical tools to investigate text have been used to identify parameters of texts used in English language reading tests (Green et al, 2010; Green et al, 2013; Khalifa & Schmitt, 2010; O'Sullivan, 2015a; 2015b), language produced by test-takers during speaking tests (Read and Nation, 2002; Seedhouse, Harris, Naeb & Üstünel, 2014) and, importantly for this report, learner language produced during writing tests (Laufer & Nation, 1995; O'Loughlin, 2013; Weir, 2014). Each of these studies is marked by the development and analysis of large datasets. Responses are grouped according to language level, e.g., CEFR band they have been awarded. These groups are then compared using statistical hypothesis testing to determine whether there are any significant differences across score boundaries. The findings are used to provide evidence that newly-developed or existing rating scales are functioning as intended by demonstrating that raters are responding to observable linguistic differences across language learners at different stages of proficiency.

Automated text indices which have shown to consistently discriminate across score boundaries have been used to develop automated-marking software for use in large-scale testing (Chapelle & Chung, 2010; Enright & Quinlan, 2010; Xi, 2010; Weigle, 2010). These indices have become more available to researchers due to the emergence of free online analytical tools to investigate texts. These have been used for a variety of research purposes, such as investigating parameters of reading texts used in tests targeting different CEFR levels (Green et al, 2010; Green et al, 2013; Khalifa & Schmitt, 2010; O'Sullivan, 2015a; 2015b), language produced by test-takers during speaking tests (Read & Nation, 2002; Seedhouse, Harris, Naeb & Üstünel, 2014) and language produced during writing tests (Laufer & Nation, 1995; O'Loughlin, 2013; Weir, 2014). Each of these studies is marked by the production and analysis of large datasets. Text data is analysed against predefined metrics to produce numerical matrices of cases against variables such as sentence length (number of words), number of sentences, number of paragraphs or more complex metrics of textual complexity like Flesch-Kincaid reading ease (a readability formula combining word and sentence length). Responses are grouped depending on the grade or band they have been awarded. These groups may then be compared using descriptive data (e.g. O'Loughlin, 2013), analysis of variance (e.g. Laufer & Nation, 1995), or regression analysis techniques (e.g. Weir, 2014). In the latter case, regression analysis can be used to develop automated assessment engines which can score responses based on lexical metrics extracted from learner writing (Xi, 2010).

2.3 Automated analysis of learner written data

Laufer and Nation (1995) present one of the earliest efforts to use an automated text analysis tool as a means of measuring the quality of student written work. The authors used *VocabProfile* (Cobb, 2019; Heatley, Nation & Coxhead, 2002) to compare learner written work (rather than exam texts) against corpus data (e.g., BNC and COCA) to establish the proportion of the most frequent 2,000 words in English and the proportion of lexis from the University Word List (UWL) (Xue & Nation, 1984) used by learners of English. Results indicated that the percentage of word families from the first and second 1,000 most frequent word lists was highest for the least proficient learners, while the percentage of word families used from the UWL was highest for the most proficient learners. The authors concluded that quantitative metrics are sufficiently sensitive to the quality of student writing to establish lexical profiles for learners at different stages of proficiency. Since then, studies have sought to expand upon this early work by investigating test data and exploring other metrics which may discriminate between stages of proficiency.

From a testing point of view, O'Loughlin (2013) examined data from candidates in three different written production tasks within the Pearson Test of Academic English (PTE). Two of the tasks require test-takers to explicitly use material presented to test-takers (aural and text input). The study investigated the extent to which respondents at different levels of proficiency incorporated lexis from the input into their responses. Additionally, percentages of academic vocabulary were compared across score bands. Although the percentage of academic vocabulary increased with proficiency, the number of tokens from the input varied considerably across proficiency level. One caveat of analysis of lexical use is that the data is decontextualised and does not consider how lexis has been used to form a coherent task response.

Some automated analysis tools have attempted to address this perceived weakness by incorporating metrics of text difficulty. For example, *Coh-Metrix* (Graesser et al, 2004) is a computer program developed at the University of Memphis which "analyses over 200 measures of cohesion, language and readability… [using] part-of-speech classifiers, syntactic parsers, templates, corpora, latent semantic analysis and other measures that are widely used in computational linguistics" (Graesser et al., 2004, p. 193). Weir (2014) used Coh-Metrix to compare learner written responses to tasks in the Eiken TEAP Writing Test. Responses were compared across bands of the rating scale which were aligned to bands A2–B1 on the CEFR. Statistical analysis indicated that there were significant differences between the scripts in adjacent band levels for a range of metrics.

These included number of words, average syllables per word, sentence length, number of modifiers per noun phrase, number of words before the main verb, Flesch-Kincaid reading ease and frequency of content words (adverbs, adjectives, main verbs). Additionally, Weir identified a difference in cohesion between A2 and B1 in terms of stem overlap, a measure of cohesion which examines the proportion of adjacent sentences which share one or more-word stems. Coh-Metrix is notable for including metrics of textual cohesion in addition to more descriptive metrics. For example, Latent Semantic Analysis (LSA) is a series of metrics which measures lexical co-occurrence across sentences and paragraphs. The underlying assumption is that lexis which co-occurs frequently will have greater conceptual closeness than lexis which does not frequently co-occur. Texts which have greater conceptual closeness will therefore be easier to parse than texts which contain greater conceptual diversity (Vigliocco & Vinson, 2007). However, Coh-Metrix is undermined by the opaqueness of its output, with only 54 of around 600 algorithms publicly available (Weir, 2014) and thus not suitable for inclusion in item-writer guidelines or test specifications due to uncertainty about the meaning of the output and the surfeit of accountability this would create.

Despite the successes of identifying metrics which discriminate across levels of learner writing proficiency, alternative voices have suggested that lexical metrics may lack discriminatory power, as lexical metrics do not account for crucial components of the construct of learner writing proficiency. For example, Albrechtsen, Haastrup and Henriksen (2008) note in the context of their study of Danish students' writing in L1 and L2 that other factors are of significance, reporting that

> "students have to generate ideas, organise these ideas with a view to their audience and express their ideas in the L2. For the latter, again, more is at stake than lexical knowledge; they also need to be able to produce correct sentences and, thus, have to draw on other aspects of their linguistic competence." (Albrechtsen, Haastrup and Henriksen, 2008, p. 172)

The implication is that although lexical knowledge and use is of great importance in determining L2 writing proficiency, other variables impact the writing process which are not accounted for in lexical metrics, with the result that studies which only identify boundaries using lexical metrics do not explore differences which cannot be transformed easily into numerical data, or if they are, cannot easily be related to observable features of text. As such, these metrics may be of limited value to item writers or raters in identifying differences between texts.

Additionally, although the use of test-taker data has contributed to understanding the development of L2 writing proficiency, there remain no common standards of learner written proficiency and therefore no agreement in the field regarding whether (and which) metrics ought to be included in test specifications or item writer guidelines. Learner samples used in Laufer and Nation (1995), O'Loughlin (2013) and Weir (2014) were from different tests which have different tasks and task requirements. As a result, outcomes are not directly comparable. Therefore, there is a wider necessity within the literature for research to contain sufficient detail of analytical procedures to ensure replicability and a movement towards more robust lexical profiles to match the widely-used English Vocabulary Profile (EVP) and English Grammar Profile (EGP).

Research has also focused on using machine learning (ML) and natural language processing (NLP) to align texts to levels of the CEFR. These studies examine the relationship between passage difficulty and linguistic features such as text length, average word length, frequency of negations, and rhetorical organisation using regression techniques on a training dataset to predict CEFR levels for texts in a separate test dataset. For example, Xia et al. (2016) used lexical, discourse and syntax features to predict CEFR levels for reading passages from a suite of Cambridge English exams, targeting levels A2–C2. Text difficulty is an important element of task difficulty (albeit this is also dependent on task design). However, these features can be used as part of validity arguments for tasks and tests (Freedle & Kostin, 1993, 1999) and inform test development (Nissan et al., 1995; Kostin, 2004). In particular, average word length and sentence length have been shown to correlate ($r = .91$) with comprehension scores in reading tests (DuBay, 2007), improving the prospect of using these lexical and discourse features as part of a model of automated assessment of student writing.

2.4 Researching metadiscourse use in learner written data

As part of the development of lexical profiles in learner written data, this report also focuses on the use of metadiscourse at different stages of learner writing proficiency. Metadiscourse refers to language used to manage the progression of text, such as organisation, and to overtly state the attitude of the writer (Burneikaite, 2008, p. 39). Examples include ‘therefore’ (logical connective); ‘essential’ (emphatic); ‘so’ (topic shift); and ‘finally’ (frame marker). Metadiscourse is an example of explicit organisational structure of a text. It can also indicate a writer’s stance towards the text’s content or towards the reader (Hyland, 2004, p. 109). It is important in more academic writing styles to guide readers through conceptually dense text. Hyland (2004) identifies 13 categories of metadiscourse marker as shown in Table 1.

Table 1: Hyland’s categories of metadiscourse markers (Hyland, 2004, pp. 109–111)

	Category analysed		Function	Examples
Textual metadiscourse	Logical connectives		Express semantic relation between main clauses	In addition / but / thus / and
	Frame markers:	Sequencing	Explicitly refer to discourse acts or text stages	Finally / to repeat / here, we try to
		Label stages		
		Announce goals		
		Topic shift		
	Code glosses		Help readers grasp meanings of ideational material	Namely / such as / e.g. / i.e.
	Endophoric markers		Refer to information in other parts of the text	Noted above / see figure X
	Evidentials		Refer to source of information from other texts	According to X, … / 1990 / X argues that…
Interpersonal metadiscourse	Attitude markers		Expressing opinion of propositional content	I agree that… / X claims that…
	Hedges		Withhold writer’s full commitment to statements	Might / perhaps / possible
	Relational markers		Explicitly refer to or build relationship with reader	Frankly / note that / as you can see…
	Person markers		Explicit reference to author	I / we / mine / our
	Emphatics		Emphasise force or certainty in message	Definitely / in fact / it is certain that…

The CEFR states that at the B2+ level and beyond there is “a new focus on discourse skills… [in which] the learner can arrange sentences in sequence so as to produce coherent stretches of language” (Council of Europe, 2001, p. 35, p. 123). This implies that for candidates writing at these levels there should be an increased awareness of macro-features such as genre, audience and text purpose and of how micro-features such as organisation and discourse markers contribute to the target genre.

Studies to date have provided conflicting data regarding the use of metadiscourse in learner writing. Burneikaite (2008), Hawkey and Barker (2004) and Carlsen (2010) argue that higher-level writing exhibits proportionally fewer simple logical connectives than lower level writing due to a wider range of metadiscourse and lexis employed. Burneikaite (2008) argues that higher level writing will exhibit significantly higher use of endophoric markers (markers referring to information in other parts of the text, i.e., cross-referencing) and significantly higher use of evaluative markers, particularly emphatics (e.g., definitely). In contrast, Sanford (2012) argues that higher-level writers will use proportionally more metadiscourse markers overall, regardless of communicative function. Bax et al (2012) noted that although the use of metadiscourse proportionally decreased from B1 to C2 levels of the CEFR, different metadiscourse markers displayed different trends. The authors found that the use of endophoric and evidential markers increased with proficiency, while use of emphatic, hedge, person and relational markers all decreased. Exploring further, Bax et al (2019) concluded that the *range* of metadiscourse is equally important to discriminate between proficiency levels rather than just categories of metadiscourse. Of the 13 metadiscourse categories, 10 were significantly different across CEFR thresholds. Announce goals, Label stage and Sequencing were not significant. In eight of the 10 significant categories (attitude markers, code glosses, emphatics, endophorics, evidentials, hedges, logical connectives, topic shifts), higher-level writers used a greater variety of metadiscourse markers than lower level writers (2019, p. 9).

These mixed findings are indicative of the cognitive complexity of L2 writing for academic purposes and demonstrate that analysing only categories of metadiscourse may conceal significant variation in use of lexical items, and that use of individual words is dependent on a variety of intersecting variables including complexity, concreteness and frequency.

2.5 Aligning the Aptis test to the CEFR

Part of the definition of learner lexical profiles is to align learner progression to a widely-used framework of language proficiency. As this research focuses on Aptis candidates' writing, this section looks at the alignment evidence of the Aptis writing test against the CEFR. Language for the Aptis scale descriptors is purposefully based on language from CEFR descriptors (O'Sullivan, 2015a, p. 60) and each level of the scale is designed to align to a level of the CEFR. The Aptis test has undergone alignment to the CEFR following the five-stage alignment process recommended by the Council of Europe (2003; 2009): familiarisation, specification, standardisation training and benchmarking, standard setting, and validation. Each of the components (speaking, reading, writing and listening) was aligned separately. The alignment process followed an analytical judgement method in which samples of candidate writing were rated against CEFR descriptors. The developers do not claim that the outcome of the research is a definitive link to the CEFR but rather a provisional standard-setting project providing evidence of alignment (O'Sullivan, 2015, p. 40).

Descriptions of lexical use in the specifications and scale descriptors consist of words such as 'limitations', 'sufficient' or 'not sufficient' based on the communicative goal of the task. From a lexical point of view, these words are ambiguous and do not easily provide scope for comparing lexical use across CEFR thresholds. However, the test specifications for each of the tasks outline 'features of the expected response' (Appendices 4 to 6) which provide information on lexis required to successfully complete each task. The lexical levels of the expected responses are specified using the BNC-20 lists derived from the British National Corpus by Nation (2006) and adapted by Tom Cobb (http://www.lextutor.ca/freq/eng/).

The British National Corpus (BNC) is based on texts from a variety of sources from written and spoken English and covers a wide variety of genres, totalling more than 100-million words (Burnard & Aston, 1998). Text Inspector also provides data derived from The Corpus of Contemporary American English (COCA), which is composed of more than 560 million words from 220,225 texts (Davies, 2009) and The English Vocabulary Profile (EVP), which describes what lexis are typically learned at each level of the CEFR, based on the Cambridge Learner Corpus (CLC), developed by Cambridge ESOL, a compilation of several hundred thousand Cambridge examination scripts from multiple regions and countries. The lists comprise 20 levels, each with 1,000-word families. K1 refers to the most frequent 1,000-word families, K2 the next most frequent 1,000-word families, etc. (O'Sullivan et al., 2020, p. 59).

The British Council provides information regarding the layout, content and characteristics of the expected response of the writing tasks in freely available test specifications (O'Sullivan et al., 2020, pp. 84–88). These specifications include claims for lexical content expected by test-takers in the different tasks. Part 1 of the test requires test-takers to respond to five text messages. There is no extended writing in this part. Test-takers are required to respond using only individual words or phrases. Spelling, grammar and punctuation are not explicitly considered in this part. The focus is on meaningful communication (British Council, 2020)[1]. Part 2 (see Appendix 1) requires a short, constructed response of 20 to 30 words, control of A2-level grammar and K1-K2 lexis, written with complete sentences. A2 responses use simple grammatical structures, complete sentences and some evidence of using punctuation and spelling conventions although mistakes will be common. A2 responses are described as using 'mostly sufficient' vocabulary to complete the task and will show 'some evidence' of using simple connectors whereas A1 responses will show no evidence of cohesive devices. The rating scale is a 6-point scale from 0–5. An A2-level performance is required to achieve score bands 3–4. A score of 5 is awarded for performances beyond A2 level.

Part 3 covers bands A2 and B1 and requires candidates to produce three short constructed responses (30 to 40 words per response). Each response needs to be structured as sentences, and the candidate must respond adequately to at least two questions to receive a rating of 3 or more (out of 5), which equates to a CEFR band of B1. In the scale descriptors, the number of relevant, on-topic responses answered is the first entry in each of the levels of the scale descriptor, suggesting that this is the primary consideration within each CEFR band (see Appendix 2). There is some overlap between the scale descriptors which contain the same CEFR bands, for example, A2 responses for Part 3 will contain 'simple grammatical structures' and complete sentences. However, there is also a divergence. At A2 level for Part 3, responses will *not* contain sufficient vocabulary to respond to the task effectively, suggesting that Part 3 is designed to elicit a wider range of lexis and grammatical constructions than Part 2, and the complexity of Part 3 will stretch A2 candidates to attempt language beyond their ability, causing breakdowns in communication. As a result, candidates will make 'inappropriate lexical choices' and make 'errors with simple structures' which 'sometimes impede understanding'. B1 responses for Part 3 demonstrate control of grammatical structures, sufficient vocabulary to respond to the questions and basic cohesive devices which link information as a linear sequence of events.

Part 4 requires candidates to write two separate emails, one in an informal register, one in a formal register. The first email is approximately 50 and the second 120 to 150 words. Successful completion of this task requires use of K4–K5 level lexis. B2 responses should therefore employ responses beyond the most common 4,000 words, and C-level responses will likely demonstrate command of lexis beyond the most common 5,000 words. Part 4 covers bands B1 to C2. As before, there is overlap between scale descriptors which contain references to the same CEFR levels. For this task, B1 candidates experience limitations in vocabulary which make it difficult to deal with the demands of the task. B2 responses contain sufficient vocabulary, but sometimes make 'inappropriate' lexical choices.

[1] N.B. The dataset used in this study was based on a previous version of Task 1. The previous version was not scored using a rating scale, and so was not included in this analysis.

B1 responses are characterised as only using ‘simple cohesive devices’ and that ‘ideas between sentences may not be indicated’, consistent with the B1 descriptor for cohesive devices in Part 3. B2 responses are characterised as using limited numbers of cohesive devices to link ideas. C-level candidates use ‘a *range* of cohesive devices’ (emphasis added) to indicate links between sentences (Appendix 3).

Returning to the CEFR for additional evidence regarding writing proficiency at each level, the written assessment grid (Council of Europe, 2003, p. 187) provides information on lexical differences across thresholds, particularly with reference to cohesive devices. A1 responses contain a basic repertoire of words and simple phrases employing connecters such as ‘and’ and ‘then’. A2 responses use basic sentence patterns with memorised phrases and can group words together using basic connecters such as ‘and’, ‘but’ and ‘because’. The B1 descriptor is consistent with those of the British Council, with responses containing ‘sufficient’ vocabulary to express themselves with some circumlocutions on basic topics and can link text together in a linear fashion. B2 level responses show evidence of an ability to construct some complex sentences, although at this level, language lacks expressiveness, idiomaticity and is stereotypic. Responses at this level use a number of cohesive devices to link sentences into coherent text, although text contains some ‘jumpiness’. C-level responses go beyond B2 level responses by deploying idiomatic formulaic sequences and controlled use of organisational patterns.

3. RESEARCH QUESTIONS

From the outlined research agenda and consideration of the literature above, the following research questions were devised.

1. **Research question 1:** What are the lexical thresholds and lexical profiles of candidates taking the Aptis writing test across the Common European Framework of Reference for Languages (CEFR) levels?

2. **Research question 2:** To what extent and in what ways are the metrics identified in RQ1 of value, or deficient, for the purposes of automated assessment of learner writing?

The following section details how the research questions were addressed, detailing data collection and analytical procedures.

4. RESEARCH DESIGN AND METHODOLOGY

4.1 Materials

The dataset for the study consists of a corpus of 6,407 scripts of Aptis candidates' writing. Characteristics of the sample, including scores awarded by country, may be viewed in Appendix 4. The corpus represents 65 countries. The most represented country was Spain, with 829 samples. Mexico, Egypt, Colombia and Saudi Arabia all contributed more than 400 samples. Conversely, 24 countries all contributed less than 10 samples each, with one from Iraq, two each from Belgium, Ethiopia, Slovenia, Afghanistan and Tanzania, and three each from France, Greece, Libya, Russia and South Africa. The sample is roughly normally distributed in terms of score, as shown in Table 2, with most samples being awarded a score of B1, with fewest samples being awarded bands A0 and C, although the sample under-represents A2 candidates.

Table 2: Score distribution of sample

CEFR band	A0	A1	A2	B1	B2	C	Grand total
Total	237	723	688	2546	1948	265	6,407

Scores were not evenly distributed across countries. Although Senegal contributed 121 samples, none of these were awarded a C band. Conversely, Nigeria contributed 54 samples, of which 21 were awarded a C band. The average length of the transcripts is 273.76 words and the total corpus size is 1,753,431 words. The data used in this study does not distinguish between C1 and C2 levels (Zheng & Berry, 2015, p. 4).

The Aptis writing test is comprised of four parts built around a common topic, e.g., becoming a sports club member. Raters mark learner writing by individual task and not by the whole test, although the marks for each are aggregated and transformed into a global score which is also assigned a CEFR band. Different parts fulfil different rhetorical purposes, and writing strategies by candidates mean that different parts of a text exhibit different lexical diversity. For the purposes of this research, the responses to the different parts were amalgamated into a single text for each candidate.

4.2 Data cleaning

The dataset was cleaned prior to analysis with Text Inspector. The three parts for each candidate were amalgamated into a single response for analytical purposes. Three research assistants were recruited to visually inspect the dataset for errors which would hinder automated analysis. Visual inspection revealed extraneous punctuation to be removed (Table 3). Extra punctuation marks risked undermining sentence counts, as they could be misread by the software as sentence boundaries. In order for the software to not mis-read this data, it was cleaned in the following ways, as shown in Table 3.

Table 3: Data cleaning procedure

Sample	Action
Hi Maria,	All names and other personal identifying information were removed.
Hi Maria,!the next week, I don't have cooking classes.	Exclamation marks removed to ensure that a single sentence is not divided into two.
you will go to the street.!Don't worry my friend.!	Exclamation mark removed to ensure that additional sentence is not inserted. Full stop retained.
I'm waitting your replay...see you	Ellipses removed, although left as one sentence to retain original meaning of author. Spelling errors retained.
The teacher is in a Per£, he went go to the voluntier for de NGO, only one week	Special symbols removed, although spelling mistakes were retained.
Dear Dear Sir Dear Managers Hi! Yours faithfully, XXX.	Salutations removed due to potential negative impact on number of words per sentence metric.

There was a tendency for lower-level candidates to include personally identifying information to make up for a lack of content in their responses. For example, they would write their names, addresses and emails in full, which were removed for analytical purposes in this study.

4.3 Manual analysis

For metadiscourse markers, a sample of the responses (n = 200) were selected at random from the overall sample. These were uploaded to *Simple Concordance Program* (SCP 4.0) in plain text format. Searches were conducted for each of the metadiscourse markers included in the analysis. The researchers then independently reviewed the occurrence of metadiscourse markers against Text Inspector data to ensure the markers had been correctly labelled. This is an important step in analysing metadiscourse data, as context is important to understand the function of individual words. Text Inspector outputs for the entire dataset were manually adjusted to mitigate any biases in the computer analysis. Therefore, the data for the remaining 6,167 texts were based on estimates from the initial analysis of these 200 texts.

4.4 Data analysis

The transcripts were entered carefully into the automated online tool, TextInspector.com, in batches by each of the three research assistants. Although the tool is automated, the texts have to be entered in batches due to limitations on the number of texts that can be analysed simultaneously. Output data was then visually checked against the input text to identify any obvious errors. Results for each batch were then stored in Excel documents. Once the dataset had been processed via the online analysis tools, and checked, the resulting Excel spreadsheets were amalgamated into one spreadsheet containing data for all 6,407 texts. As data collection and management was conducted by three independent research assistants over a period of several months, inconsistencies were noticed in the dataset. For 283 of the texts, no metadiscourse data was recorded. This was coded in the spreadsheet as missing data. Data for metadiscourse markers is therefore based on a revised n-size of 6,124. The Excel spreadsheet was then imported into SPSS v.24 for statistical analysis.

An overview of the metrics used in the study can be found in Appendix 5. As can be seen from that list, they include standard statistics such as Type-Token ratio and average sentence length, classic *readability* measures such as Flesch Kincaid reading ease, *lexical diversity* measures (voc-d and MTLD), and the *Academic Word List* (AWL). In addition, a range of measures unique to Text Inspector were used, including measures using the English Vocabulary Profile tool (in association with Cambridge University Press), which is of particular importance to this study since it classifies lexis according to CEFR level. Additional measures of lexical frequency include the British National Corpus (BNC), which is also important for this study as features of the expected response in task specifications are described in relation to BNC lexis, and the Corpus of Contemporary American English (COCA). Finally, measures of the incidence and frequency of 13 categories of *metadiscourse marker* (Hyland, 2005, modified by Bax, Waller & Nakatsuhara 2014) are also included. The full list of metadiscourse markers analysed can be seen in Appendix 6.

4.5 Research question 1

To address research question 1, a series of Kruskal-Wallis tests for independent samples were performed on the data, using CEFR level as grouping variable. This is a non-parametric test and was preferred to parametric statistical approaches as Shapiro-Wilk tests of normality indicated that data for all of the metrics was non-normal and assumptions of equal variance could not be met across CEFR group distributions. The Kruskal-Wallis tests indicate whether a metric discriminates across any CEFR thresholds, and so were followed up with a series of Mann-Whitney U tests to determine which lexical thresholds were significant. This procedure was followed for all metrics used in the study, including those describing metadiscourse use.

4.6 Research question 2

The second research question explores the replicability of the findings by re-analysing a stratified holdout set of the writing samples used to address RQ1. It is possible that statistically significant indices which emerge in relation to RQ1 may not be generalisable beyond the data used in this study, due to differences in sample characteristics such as learner demographics, learner level and test task. Indices which are shown to be generalisable beyond the dataset will be of value in moving towards automated assessment.

To address this, a subset of the data was created by controlling for CEFR level and nationality. This proved to be impossible for the A0 band due to insufficient numbers of test-takers from Spanish-speaking countries, which was therefore not used in this analysis. No Sudanese test-takers were awarded a C band, so this test-taker was replaced at this level by a test-taker from another Arabic-speaking country (Libya). Additional Spanish speaking test-takers were included from Mexico to address the shortfall in Spanish test-takers at A1. The composition of the stratified sample is outlined in Table 4.

Table 4: Stratified sample for replication analysis

Countries	A1	A2	B1	B2	C	Total
China	4	4	4	4	4	20
Colombia	4	4	4	4	4	20
Egypt	13	13	13	13	13	65
India	20	20	20	20	20	100
Mexico	39	22	12	12	12	97
Saudi Arabia	6	6	6	6	6	30
Spain	13	30	40	40	40	163
Sudan	1	1	1	1	0	4
Taiwan	2	2	2	2	2	10
Libya	0	0	0	0	1	1
Total	**102**	**102**	**102**	**102**	**102**	**510**

The data was analysed using the proportional odds logistic regression (polr) model within the MASS package (Venables & Ripley, 2002) in RStudio v3.6.0. Ordinal logistic regression (OLR) is used to predict the dependent variable with ‘ordered’ multiple categories and independent variables. In other words, it is used to facilitate the interaction of dependent variables (having multiple ordered levels) with one or more independent variables. Unlike multinomial logistic regression (MLR), which produces multiple sets of regression coefficients with associated significance tests, OLR produces single regression coefficients to estimate the relationship between predictor and dependent variables, thus reducing the prospect of Type I error. OLR requires additional tests of the proportional odds assumption, which states that the relationship between the dependent and predictor variables is constant across groups of the dependent variable (Osborne, 2015).

The proportional odds assumption is based on the cumulative odds of achieving each level of the dependent variable. The odds of achieving each level can be calculated as the proportion of test-takers who achieve that level out of the total number of test-takers. Cumulative odds can be calculated using the formula *p / (1-p).* This can be used to calculate the odds of achieving a specific level or above of the dependent variable (e.g. CEFR level B1 or C). This creates a series of probabilistic thresholds. A key assumption of ordinal regression is that the effects of the independent variables are consistent or proportional across these thresholds. For example, if we observe that females outperform males in test performance, we expect to see proportionally more females than males in the higher categories and proportionally more males in the lower categories. Any discrepancy between categories would violate the proportional odds assumption and necessitate the use of a series of binary logistic regressions at each score boundary. The Brant test in *R* is used to test the proportional odds assumption for ordinal logistic regression models generated using the polr() function from the MASS package (Brant, 1990). Goodness-of-fit measures are constructed to test the assumption of proportional odds across boundaries. If the p-value is >0.5, then the dataset satisfies the proportional odds assumption.

The objective of the analysis is to predict CEFR level based on the metrics and to evaluate the utility of the metrics by comparing the CEFR group assigned by the regression model to the score awarded by Aptis raters. The polr() function allows for partitioning of data into training and testing sets. Assignment into either set is random. For this analysis, 75% of the sample were used to train the regression model, which was then used to predict the CEFR levels of the remaining 25% of the sample.

Before running the model, it is essential to consider the exploratory data analysis conducted in relation to RQ1. Specifically, multicollinearity needs to be evaluated by examining correlations among the independent variables. In this study, discriminating indices which correlate above 0.9 were flagged and then removed if they were thematically related. For example, there are multiple indices of text length which correlate highly with one another. Additionally, missing values in the metadiscourse data meant that these were eliminated from inclusion. Instead, the overall percentage of metadiscourse was included in the model. Ultimately, 16 variables were included in the regression model.

5. FINDINGS AND DISCUSSION

5.1 Research question 1

What are the lexical thresholds and lexical profiles of candidates taking the Aptis writing test across the Common European Framework of Reference for Languages (CEFR) levels?

Independent-samples Kruskal-Wallis tests, using CEFR level as grouping variable were performed on the Text Inspector data. Post-hoc Mann-Whitney U tests were also performed to identify which CEFR thresholds were significant for those metrics which recorded significant Kruskal-Wallis results. Those metrics which discriminated across all CEFR thresholds are reproduced in Table 5. Descriptive data detailing the lexical thresholds for these metrics is available in Appendix 7. This data provides the basis for claims regarding lexical thresholds and profiles. Appendix 7 provides descriptive data detailing the lexical profiles of each CEFR level for each significant metric. This includes the mean, standard deviation, confidence intervals, minimum and maximum values, medians and inter-quartile ranges for each CEFR level for each metric. Minimum values for each metric can be interpreted as lexical thresholds for that CEFR band. The full results of significance testing are available in Appendix 8.

Due to multiple significant testing, three significance levels are reported. Although there is a high probability of type I error due to multiple testing, results which meet a significant threshold of $p < .05$ are reported so that future studies can compare findings to those presented here.

Table 5: Metrics contributing to lexical profiles in Aptis writing responses

	Text Inspector Metric	CEFR threshold				
		A0–A1	A1–A2	A2–B1	B1–B2	B2–C
Basic statistics	Sentence count	***	***	***	***	**
	Token count	***	***	***	***	*
	Type count	**	***	***	***	***
	Syllable count	***	***	***	***	***
	Words with > 2 syllables	***	***	***	***	***
Lexical diversity	Lexical diversity (VOCD)	**	***	***	***	***
Lexical profile	EVP A1 type %	***	*	***	***	***
	EVP B2 type %	*	***	***	***	***
	EVP C1 type %	**	***	***	***	***
	EVP A1 token %	***	***	***	***	***
	EVP B2 token %	**	***	***	***	***
	EVP C1 token %	**	***	***	***	***
	BNC type percent (2K–3K)	***	***	***	***	***
	BNC type percent (4K–5K)	***	***	***	***	*
	BNC type percent (6K–7K)	***	***	***	***	***
	BNC token percent (4K–5K)	***	***	**	***	***
	BNC token percent (6K–7K)	***	***	***	***	***
	COCA type percent (5K–6K)	***	***	*	***	***
	COCA type percent (7K–8K)	***	***	***	***	***
	COCA token percent (7K–8K)	***	**	***	***	**
Academic lexis	AWL all types	**	***	***	***	***
	AWL all tokens	**	***	***	***	***
Metadiscourse	Hedge (% of tokens)	**	***	***	***	***
	Emphatic (% of tokens)	**	***	***	***	*
	Emphatic (% of types)	***	***	***	***	*
	Person marker (% of types)	***	***	***	***	**

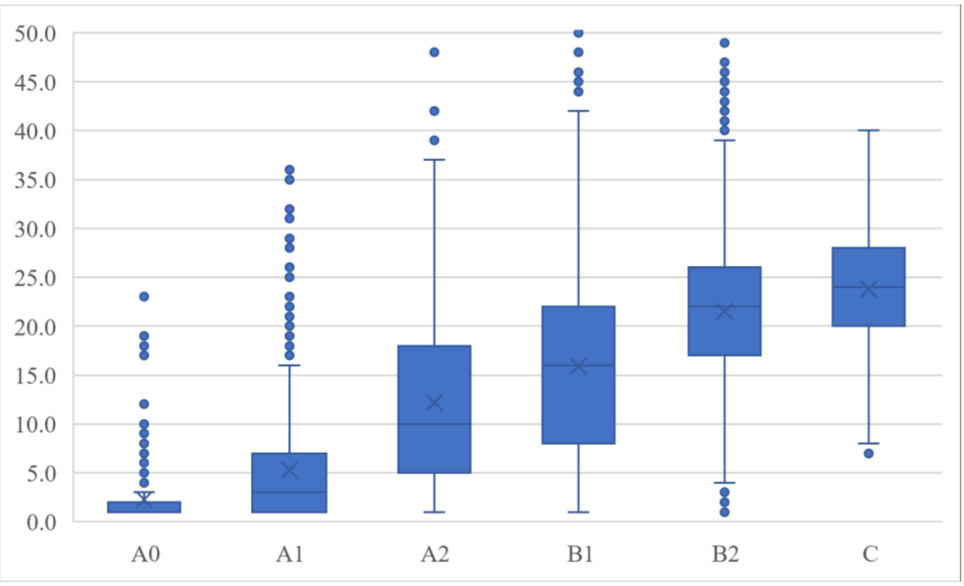
Each column tests the null hypothesis that the sample distributions are the same across CEFR thresholds.
*** = significant at *p* < .001; ** = significant at *p* < .01; * = significant at *p* < .05.

Of the 110 metrics submitted for testing, 26 metrics discriminated across all CEFR thresholds. Each column in Table 5 represents a CEFR threshold and tests the null hypothesis that the sample distributions are the same across thresholds for each of the metrics listed. The level of significance is produced for each threshold. The first five metrics are unweighted statistics relating to the amount of text produced by candidates. One metric of lexical diversity discriminated across all thresholds (voc-d). Weighted values are presented as percentages (i.e. amount of categorised text per 100 words). Fourteen weighted metrics cover vocabulary content represented by corpus data from the English vocabulary profile (EVP), the British National Corpus (BNC) and Corpus of Contemporary American English (COCA). Two further unweighted metrics discriminated in terms of the academic vocabulary used by the candidates. Finally, four weighted metrics of metadiscourse use were significant across all thresholds: the percentage of hedging tokens, the percentage of person marker types and the percentage of emphatic types and tokens. The findings for each of the categories and associated metrics will now be discussed in turn with reference to the literature alongside boxplot diagrams.

5.2 Basic statistics

Basic statistics include measures of text length (sentence, token and type count; number of verb and noun elements) and metrics of lexical sophistication (syllable count and number of words with more than two syllables). The *sentence count* refers to the average number of sentences produced in candidates' responses for each CEFR band. This metric showed an increase across all CEFR thresholds. The mean number of sentences for an A0 response was 2.25. This increased to 5.27 for A1 responses. The largest jump is from A1 to A2 responses, which recorded an average of 12.16 sentences. This then levelled off for B1 to C responses, which ranged from 15.92 to 23.81. B2 responses contain an average of 21.51 sentences.

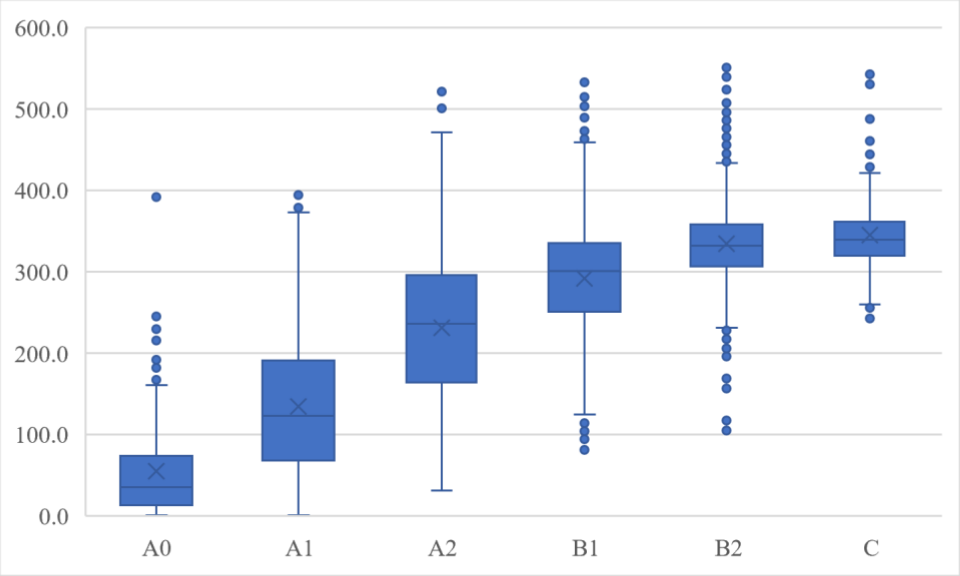
Figure 1: Number of sentences by CEFR level



Standard deviations were largest for A2 and B1 responses (8.65 and 8.98 respectively). Although there was no overlap in upper and lower confidence intervals for each band, the range was very large for all responses, with only C responses recording a minimum of more than one sentence. This finding is likely due to an artefact of Text Inspector responding to punctuation marks. If candidates omit punctuation marks from their responses, the software will record one sentence.

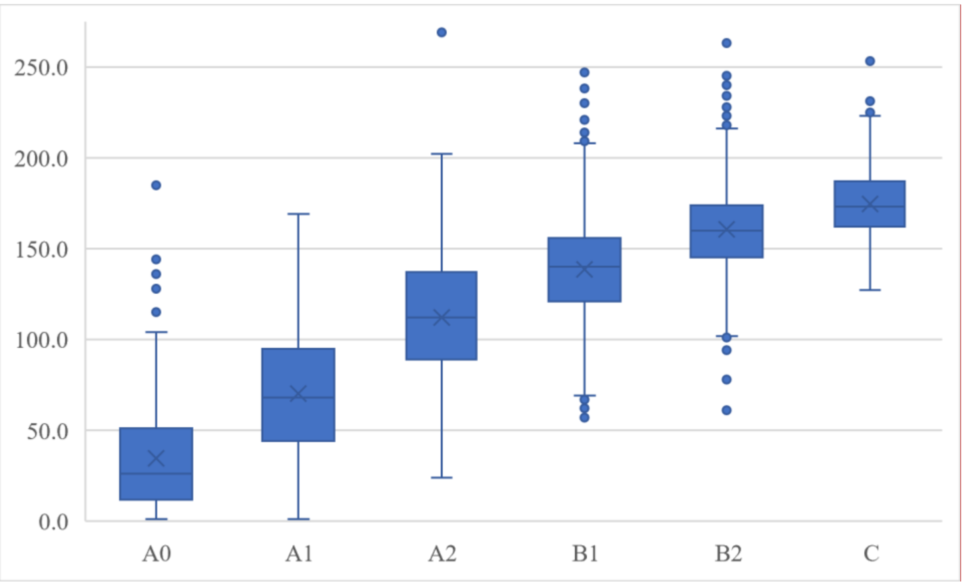
The *token count* is the total count of every occurrence of word types. This metric is therefore highly correlated with sentence count. It also followed a similar pattern. A0 responses recorded the fewest tokens, with an average of 54.95. A1 responses recorded an average of 134.18. The largest increase was from A1 to A2 responses, which contain an average of 231.05 tokens. This then plateaus, with B1 responses recording an average of 291.55 tokens, B2 responses 334.42 and C-level responses 345.40. Similarly, there was no overlap in 95% confidence intervals for all levels, although the ranges are more than 500 words for A2–B2 bands. B2 responses recorded the largest range (644 tokens). C-level responses recorded the smallest range (300 tokens).

Figure 2: Token count by CEFR level



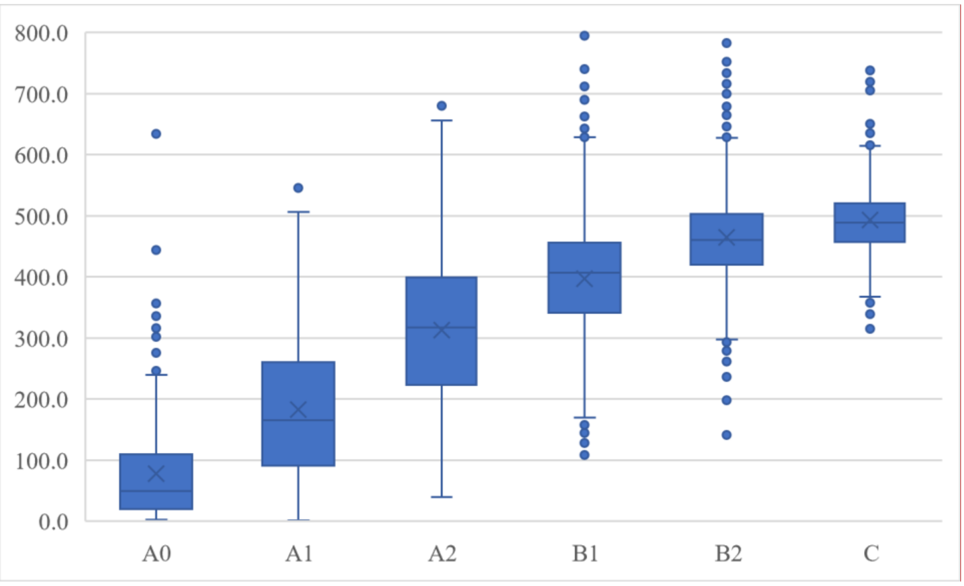
The *type count* refers to the number of unique tokens in each response. Repeated words only count as a single type. A larger number of types indicates greater lexical diversity. As expected, higher-rated responses contain greater diversity. A0 responses recorded an average of 34.67 unique tokens. A1 and A2 saw large step increases, to 70.27 and 112.19 respectively. The number of unique words then plateaued, to 138.5, 160.47 and 174.6 for B1, B2 and C-level responses, respectively. As with sentence and token counts, there was no overlap in 95% confidence intervals. There is a large overlap in range, although C-level responses contain a minimum of 127 different types, more than twice as many as the minimum B2 threshold (61 types).

Figure 3: Type count by CEFR level



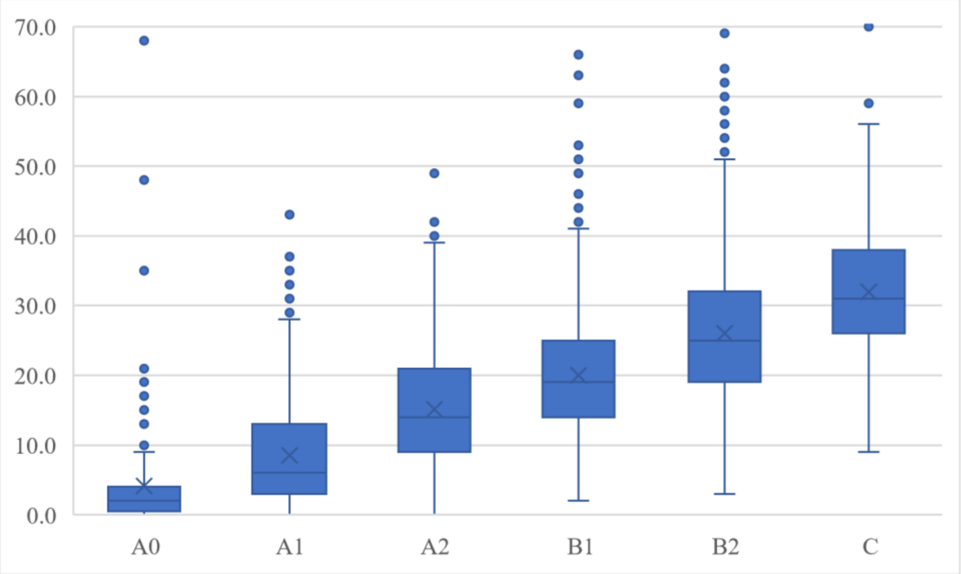
The s*yllable count* refers to the number of syllables produced in each response. Syllables are counted by comparing input against entries in the Carnegie Mellon dictionary. However, this metric is unable to account for regional variation. The number of words with more than two syllables provides an indication of text complexity and how many advanced words a test-taker at that level might be expected to produce. These metrics correlate highly with token count and sentence count.
The average number of syllables increases with CEFR band, from 77.87 at A0 to 493.39 at C-level. The largest increase is bands A1–A2, which sees a jump from 182.49 to 312.84. Minimum values for levels A0–A2 are low (2–40), which rise substantially to 108 at B1 level, 141 at B2 level and 315 at C level.

Figure 4: Syllable count by CEFR level



The second syllable-related metric which discriminates across all thresholds is the *average number of words with more than two syllables*. A0 responses contain an average of 4.13 words with more than two syllables, rising to 8.49 for A1 responses. C-level responses contain an average of 31.93 words with more than two syllables. Again, there is no overlap at the 95% confidence intervals.

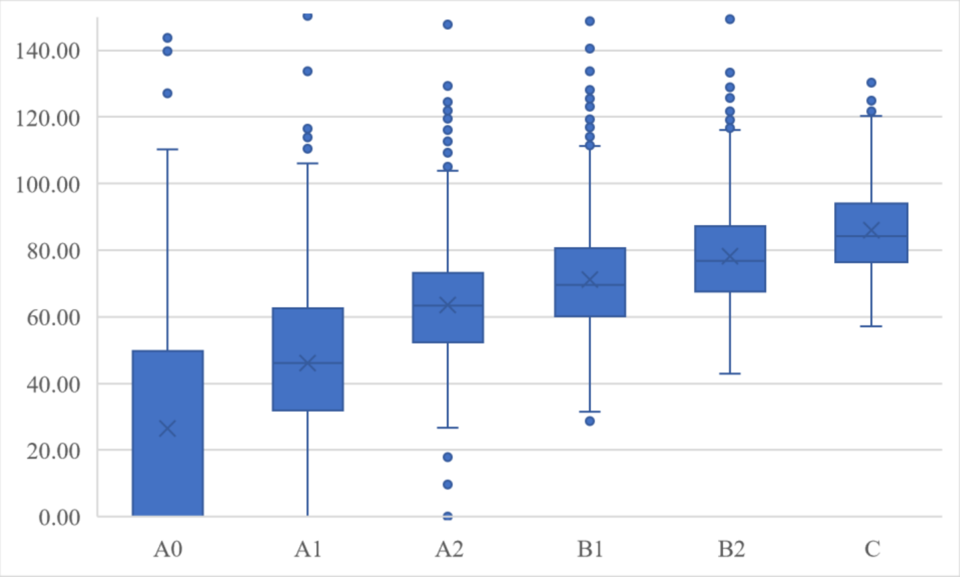
Figure 5: Average number of words >2 syllables by CEFR level



5.3 Lexical diversity

Lexical diversity (LD) refers to the range of different words used in a text, with a greater range indicating a higher diversity (McCarthy & Jarvis, 2010, p. 381). The previous section revealed increases in both tokens and types across CEFR thresholds, although we noted that the number of types plateaus as CEFR level increases. While the number of tokens in a text increases in a linear fashion (as the token count is simply a measure of the number of words), the overall number of types increases more slowly due to the necessity of repeating grammatical function words in longer texts. Therefore, the traditional measure of lexical diversity, the type-token ratio (TTR) becomes arbitrary when analysing texts of varying lengths. Voc-d is a measure of textual complexity which operationalises this understanding of text production. A voc-d score is established by taking 100 random samples of 35 tokens and calculating a type-token ratio (TTR; number of types as a proportion of number of tokens) for each. The mean TTR is stored. The same procedure is then repeated for samples from 36 to 50 tokens. The overall procedure is repeated three times and an empirical TTR curve is then created from the means of each of these samples (McCarthy & Jarvis, 2010, p. 383). This is a non-sequential approach to analysing lexical diversity, meaning that the order of words as they appear in a text does not contribute to the overall voc-d score a text receives. This approach avoids the pitfall of being influenced by localised clustering of content words but does not consider text structure which readers use to form a coherent mental picture of a text (Van Dijk & Kintsch, 1983). Values typically range from 10–100 (though can be greater), with higher values displaying greater lexical diversity.

Figure 6: Voc-d score by CEFR level



McCarthy and Jarvis (2007) demonstrated that the voc-d metric significantly varied as a function of text length. Evidence of the impact of text length on voc-d scores is visible in the findings, as shorter A0 and A1 responses record greater lexical diversity than A2 responses. The data reveals that the average voc-d score decreases from A0–A1 (from 67.9 to 55.93), before increasing to a maximum average value of 86.04 for C-level responses. A0 responses were also higher than A2 responses, which display an average voc-d value of 64. However, it should be noted that the median voc-d score for A2 responses is higher than the median for A0 responses (63.45 versus 55.82), also visible in Figure 6. This indicates that A0 responses are highly variable. The visual inspection of A0 responses shows that they are composed not only of error-strewn text, but also learned responses which are unrelated to the task and responses which simply repeat large amounts of task input material.

This makes voc-d scores for low-level responses unreliable, explaining the wide variability in values for A0 texts. However, statistically significant increases across B1, B2 and C-level responses, combined with smaller score ranges, suggest that voc-d is suitable for describing lexical diversity at higher CEFR levels and that Aptis tasks stimulate diverse language production by higher-ability candidates. There is no plateauing or saturation exhibited within the dataset to indicate tasks only stimulate a narrow range of language (Morse, 1995). However, the wide diversity of scores obtained by lower-scoring candidates indicates that this metric may not contribute successfully to an automated approach to scoring responses due to sensitivity to lower text length and test-wiseness strategies by lower-ability candidates such as reproducing task input material.

5.4 Lexical profiles

The Aptis writing test provides evidence that lexical complexity of candidate responses changes as the CEFR level of candidates increases. Lower-level responses demonstrate knowledge of fewer English words than advanced learners and higher levels demonstrate knowledge of less-frequently occurring vocabulary. This is demonstrated through 14 of the 28 statistically significant metrics which represent vocabulary use in candidate responses (Table 5). These metrics compile individual words into their respective frequency bands according to how frequently these words are used in the English language. These metrics represent three corpora of either native-speaker or learner language. This data provides empirical evidence of vocabulary deployed by learners at each CEFR band, although does not offer evidence of the appropriateness of vocabulary use. Three metrics from the COCA, five from the BNC and six representing the EVP were significant across all CEFR thresholds.

Six significant EVP metrics relate to type and token percentages of A1, B2 and C1 lexis. Tables 6 and 7 provide a summary and overall trend of EVP lexis used by test-takers at different levels of the CEFR as elicited by the Aptis test.

Table 6: Summary of EVP tokens used by test-takers at different CEFR levels of the Aptis writing test

CEFR band	Average EVP A1 token count %	Average EVP A2 token count %	Average EVP B1 token count %	Average EVP B2 token count %	Average EVP C1 token count %	Average EVP C2 token count %	Total % EVP coverage
A0	64.771	6.479	2.678	0.918	0.400	0.092	75.338
A1	72.446	7.173	2.391	0.771	0.400	0.081	83.263
A2	75.487	9.202	3.581	1.087	0.422	0.095	89.874
B1	74.280	10.838	4.477	1.494	0.519	0.100	91.708
B2	72.483	12.060	5.649	2.163	0.694	0.130	93.178
C	69.517	12.822	7.059	3.253	1.018	0.226	93.895

Table 7: Summary of EVP types used by candidates at different CEFR levels of the Aptis writing test

CEFR band	Average EVP A1 type count %	Average EVP A2 type count %	Average EVP B1 type count %	Average EVP B2 type count %	Average EVP C1 type count %	Average EVP C2 type count %	Total % EVP coverage
A0	54.83	7.61	3.55	1.21	0.53	0.11	67.84
A1	61.68	8.96	3.44	1.12	0.58	0.13	75.90
A2	63.57	12.51	5.67	1.77	0.68	0.17	84.36
B1	61.03	14.89	7.35	2.49	0.84	0.18	86.77
B2	58.36	16.50	9.44	3.63	1.12	0.23	89.28
C	54.27	17.47	11.48	5.32	1.65	0.39	90.58

The quantity of A1 lexis peaks in A2 responses before falling slowly to B2 level, then dropping by more than three percentage points for C-level responses. The proportion of A2 and B1 lexis rises steadily across CEFR bands. B2 lexis remains low for A0 to B1 candidates, then rises by 0.61% to B2 and by more than a percentage point for C-level responses. The proportion of C1 and C2 lexis is highest in C-level responses, as expected, but remains surprisingly low at 1% for C1 lexis and 0.22% for C2 lexis. The proportion of off-list vocabulary remains at 7–10% for B2 and C-level responses. Off-list vocabulary predominantly consisted of proper nouns which are not included in the EVP. It is noteworthy that overall EVP coverage is lower for A0 and A1 responses than any other CEFR level, contrary to what might be expected if lower ability candidate responses are primarily composed of fewer numbers of high frequency lexis. These percentage figures are accounted for by the greater number of errors made by these students. Words with spelling errors are counted as 'off-list', resulting in lower overall coverage.

Five BNC metrics were statistically significant across all thresholds: Type percent (2K–3K; 4K–5K; 6K–7K) and Token percent (4K–5K; 6K–7K). These are lexical frequency metrics which state the percentage of words occurring among the second and third, fourth and fifth and sixth and seventh most frequent 1000 words based on corpus data. The data is reproduced in Table 8.

Table 8: Summary of statistically significant BNC tokens and types used by candidates at different CEFR levels of the Aptis writing test

CEFR band	Average BNC Type count %			Average BNC Token count %	
	2K–3K	4K–5K	6K–7K	4K–5K	6K–7K
A0	3.15	0.70	0.34	0.56	0.23
A1	3.62	0.80	0.45	0.55	0.33
A2	4.26	1.17	0.74	0.70	0.48
B1	4.78	1.34	0.88	0.78	0.53
B2	5.41	1.64	1.05	0.95	0.61
C	5.95	1.87	1.30	1.15	0.77

Lower frequency bands (2–3K and 4–5K) account for approximately 4–8% of responses in Aptis (the majority of lexis in all responses is from 0–2K bands). The data demonstrates that higher-ability candidates use less-frequently occurring lexis in their responses. Three COCA metrics were also sensitive at less frequent levels:

Table 9: Summary of statistically significant COCA tokens and types used by candidates at different CEFR levels of the Aptis writing test

CEFR band	Average COCA Type count %		Average COCA Token count %
	5K–6K	7K–8K	7K–8K
A0	0.58	0.47	0.38
A1	0.81	0.56	0.49
A2	1.06	0.63	0.51
B1	1.20	0.88	0.80
B2	1.41	1.04	0.92
C	1.73	1.30	1.11

That 14 of the 28 metrics identified here relate directly to vocabulary suggest that vocabulary metrics are the most consistent discriminators across levels of the CEFR. This finding is consistent with those of Stæhr (2008, p. 148), who argues that writing proficiency correlates significantly with vocabulary size (0.73)… [and that] "more than half of the variance in the ability to perform above average in [a] writing test was explained by vocabulary size". As a result, automated essay scoring engines incorporate measures of both expected, prompt-specific vocabulary and lexical sophistication based on a large corpus (Weigle, 2010) as part of a wider range of essay characteristics including organisation, mechanics, style, grammar and language use (including errors).

These findings call into question the use of frequency lists in test specification response attributes. For example, the response attribute for Task 4 (level B2) indicates that "K4–K5 lexis will be sufficient to complete both emails adequately" (O'Sullivan et al., 2015, p. 87). Statements about lexis are not reproduced in marking criteria for raters, as it is not easy to discern quickly or efficiently which frequency bands individual lexical items appear. Multiplying the percentage data from Table 7 by the average number of types and tokens for each CEFR band indicates that K4–K5 lexical items appear very infrequently, even in higher-level responses. C-level responses contain an average of 5.18 4–5K tokens and 3.26 4–5K types. The data from Appendix 7 also contains the ranges for statistically significant frequency metrics, which indicate that responses may be awarded bands B2 or C without containing any 4–5K lexis. Additionally, the appearance of K4–K5 lexis offers no indication of whether this lexis has been used appropriately in context. However, the appearance of infrequent lexis in conjunction with specific structures may be of utility in the development of automated scoring algorithms (see Research Question 2).

5.5 Academic lexis

Text Inspector also examines the proportion of academic lexis contained within a text based on the word list developed by Coxhead (2000). It is described on Victoria University's website as follows:

> The Academic Word List (AWL) was developed by Averil Coxhead as her MA thesis at the School of Linguistics and Applied Language Studies at Victoria University of Wellington, New Zealand. The list contains 570 word families…The list does not include words that are in the most frequent 2000 words of English. (See Coxhead, 2000 for further information). (https://www.victoria.ac.nz/lals/resources/academicwordlist/information)

Table 10: Summary of statistically significant AWL tokens and types used by candidates at different CEFR levels of the Aptis writing test

	Academic word list (AWL)	
CEFR band	**Average of all types %**	**Average of all tokens %**
A0	1.28	1.06
A1	1.30	0.93
A2	1.68	1.08
B1	2.43	1.53
B2	3.51	2.23
C	4.96	3.19

Although the Aptis test is not designed to be a test of academic English, the proportion of lexis from the AWL used by candidates proved to be a good indicator of writing proficiency. The contents of the AWL are semi-technical, function words likely to be encountered in academic texts, although they are not subject specific. Type and token percentages generally increased with CEFR band, with larger jumps at CEFR levels B1–C. Although, as Coxhead (2000) found that academic textbooks are composed of about 10% lexis from the AWL, the findings here are consistent with a test designed to assess general English proficiency.

5.6 Metadiscourse profiles

Twenty-eight metrics covering the percentage of metadiscourse types and tokens were analysed using a series of Kruskal-Wallis tests for independent samples, using CEFR level as grouping variable (as this data was also non-normal and assumptions of equal variance could not be met across CEFR groups). As higher-level candidates will typically deploy more metadiscourse markers simply by virtue of producing more words in total, percentage statistics were used which measure the proportion of metadiscourse as a percentage of the total text in each group. A complex picture of metadiscourse use emerged from the data. Tables 11 and 12 display the statistically significant CEFR thresholds for metadiscourse use (type and token percentages), following the convention of Table 5.

Table 11: Significant differences in metadiscourse tokens used across CEFR thresholds

Percentage metric (token)	CEFR threshold				
	A0–A1	A1–A2	A2–B1	B1–B2	B2–C
Metadiscourse	***	**	-	***	***
Announce goals	**	***	***	***	-
Attitude marker	***	***	***	***	-
Code gloss	-	***	-	-	-
Emphatic	**	***	***	***	*
Endophoric	**	*	-	-	-
Evidential	-	**	***	***	***
Hedge	**	***	***	***	***
Label stage	-	-	-	***	-
Logical connective	***	-	-	***	***
Person marker	***	-	***	***	***
Relational marker	-	***	***	-	-
Sequencing	-	***	-	-	-
Topic shift	-	***	***	***	-

*** = significant at p < .001; ** = significant at p < .01; * = significant at p < .05.

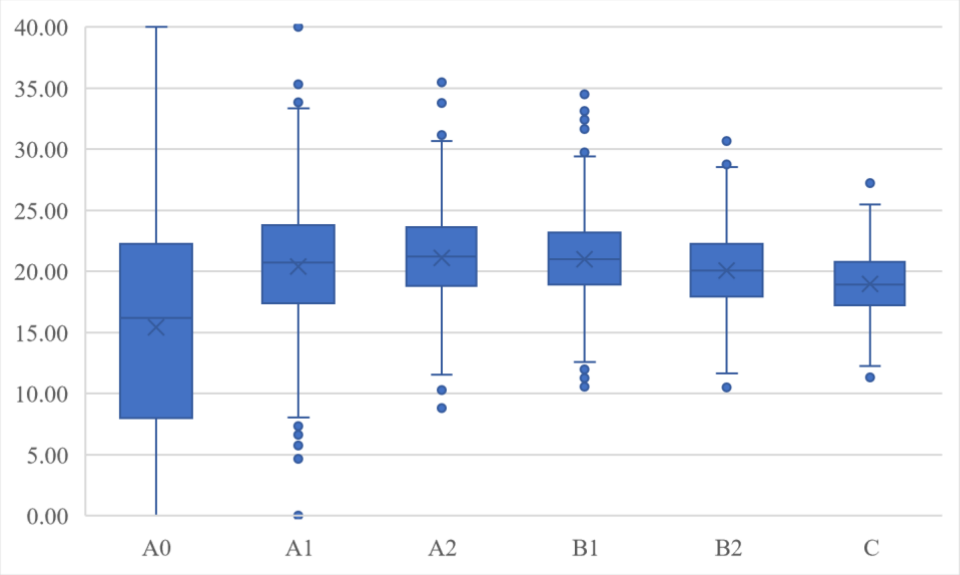
Table 12: Significant differences in metadiscourse types used across CEFR thresholds

Percentage metric (type)	CEFR threshold				
	A0–A1	A1–A2	A2–B1	B1–B2	B2–C
Metadiscourse	-	*	**	-	-
Announce goals	***	***	***	***	-
Attitude marker	***	***	***	**	-
Code gloss	-	***	-	-	-
Emphatic	***	***	***	***	*
Endophoric	**	*	-	-	-
Evidential	-	***	***	***	***
Hedge	**	***	***	***	-
Label stage	-	-	-	***	-
Logical connective	***	-	-	*	-
Person marker	***	***	***	***	**
Relational marker	-	**	-	-	-
Sequencing	*	***	-	-	-
Topic shift	-	***	**	***	-

*** = significant at p < .001; ** = significant at p < .01; * = significant at p < .05.

Each metadiscourse marker category discriminates across at least one CEFR boundary. The overall deployment of metadiscourse (percentage of tokens) changes significantly in transitioning from A0 to A2 levels and from B1 to C levels (see Figure 7). The average proportion of metadiscourse varies from a low of 17.22% of tokens in A0 responses to 21.49% in B1 responses. The proportion remains static from A2 to B1 levels and falls to 19.5% in C-level responses. C-level responses contain the second-lowest proportion of metadiscourse, after A0 responses. In terms of metadiscourse types, the number rises across CEFR thresholds from a low of 11.06% at A0 to a high of 13.54% in C-level responses but is only significant across A1–A2 and A2–B1 thresholds.

Figure 7: Percentage of metadiscourse by CEFR level



The results contrast with the claim of Sanford (2012) who argued that higher-level writers will employ proportionally more metadiscourse markers than lower-level writers. The results are broadly consistent with Burneikaite (2008), who argued that there is little difference in the overall quantity of meta-discourse markers used by lower and higher-level L2 writers (from 17.22% to 21.49%). However, this conclusion ignores significant changes in discourse competence which occur as learners develop their writing ability.

All metrics were significant across at least one CEFR threshold, although only four of the 28 investigated were significant across all thresholds. These were emphatics (type and token percent), hedge (token percent) and person marker (type percent). The percentage of emphatic tokens increases across all CEFR thresholds, from a low of 0.51% in A0 responses to a high of 1.12 % in C-level responses. Simultaneously, the percentage of emphatic types increases from a low of 0.62% at A0 to a high of 1.5% in C-level responses (see Appendix 9 for all metadiscourse data). Burneikaite's (2008) claim that higher-level writing will show significantly higher use of emphatics (e.g. definitely) is therefore strongly supported in the dataset. The use of hedges (e.g., could, possibly, may) also increases as a proportion of overall text from 0.3% for A0 responses to 1.18% for C-level responses. In contrast, the use of person markers (e.g., I, you, she, him etc.) decreased as a proportion of overall tokens. A1 and A2 responses are composed of more than 10% person markers, reducing to 9.6% for B1 responses, 8.61% for B2 and 7.76% for C-level responses. A0 responses were outliers, containing only 7.06%. Although Lee and Deakin (2016) indicated that stronger L2 writers used great amounts of interpersonal markers, their categorisation of interpersonal markers included emphatics and hedges, which we have shown increase significantly across CEFR level.

Regarding the remaining metadiscourse categories, the claim by Burneikaite (2008), Hawkey and Barker (2004) and Carlsen (2010) that higher-level writing will show significantly lower use of common logical connectives is also supported. The proportion of logical connectives is highest in A1 and A2 responses at 5.46%, falling to 4.55% for C-level responses. However, the claim that higher-level writing will show significantly higher use of endophoric markers (referring to information in other parts of the text, e.g., 'noted above') is not supported, as the proportion of tokens falls from a high of 0.18% at A1 level to a low of 0.07% in C-level responses. This may be due to an artefact of task design, which is composed of three individual tasks which do not prompt within-text referencing coupled with limited text length produced by candidates. Code glosses and relational markers are more stable across CEFR thresholds, only significant at the A1–A2 threshold.

The Aptis scale descriptors and specifications contain four significant claims related to metadiscourse use (see Appendices 1–3). These claims are as follows.

1. A2+ responses will show 'some evidence' of using simple connectors whereas A1 responses will show no evidence of cohesive devices.
2. B1 responses (for Part 3) will use basic cohesive devices which link information, such as a linear sequence of events.
3. B2 responses are characterised as using limited numbers of cohesive devices to link ideas.
4. C-level candidates use 'a *range* of cohesive devices' (emphasis added) to indicate links between sentences.

No explicit claims are made for A0 responses, although data suggests that candidates at this low level are also capable of using simple connecters such as 'and' and 'but' to link individual words. As A0 responses are so limited, the inclusion of these connecters raises the overall percentage of metadiscourse in these responses to a level consistent with the proportion of metadiscourse in C-level responses. We argue the claim that A1 responses will show 'no evidence of cohesive devices' is not supported here as the minimum percentage of metadiscourse recorded for A1 responses was 4.63%. Only the A0 band recorded a minimum percentage of metadiscourse of zero.

The proportion of *sequencing* metadiscourse is less than 1% for all CEFR bands. However, the proportion is highest for B1 responses (0.91%). Therefore, the second claim above is partially supported. Lower levels show some evidence of deploying sequencing markers, but the percentage of sequencing is low for all CEFR bands. The third claim is also partially supported. The proportion of logical connectives and evidentials changes from B1 to B2 levels and from B2 to C at significant levels. However, they change in opposite directions. The proportion of logical connectives falls from B1 to C levels (from 5.35% to 4.55%), but the proportion of evidentials rises (from 0.12% to 0.26%). The final claim (C-level candidates use 'a *range* of cohesive devices') is also partially supported. The proportion of announcing goal, emphatic, evidential, hedge, label stage and topic shift metadiscourse types is highest in C-level responses. However, the overall percentage of types is highest for B2 responses (13.58% versus 13.54%). B2 responses contain higher proportions of person marker, relational marker, sequencing and logical connective types.

Therefore, we argue that more precise claims for metadiscourse use can be made from this data relating to specific categories of use, and that recognition of metadiscourse use can be incorporated into rater training.

More general claims can also be made, which the British Council should consider integrating into test specifications. No responses composed of less than 10% metadiscourse received a band above A2 (see Appendix 9), indicating that effective metadiscourse use is a crucial component of higher-ability L2 writing in the Aptis test, with a minimum of 10.82% metadiscourse required for a B1 level score, although the use of specific metadiscourse categories varies widely across CEFR bands. The findings call into question the assertion by the Council of Europe that "a new focus on discourse skills" occurs from Level B2 upwards (Council of Europe, 2001, p. 35). Observable changes in use of metadiscourse occur across all CEFR thresholds.

The evidence suggests that the claim for Aptis should be ‘a new focus on *controlled* discourse’ as the overall proportion of metadiscourse actually decreases across the B1–B2 boundary, as does the standard deviation and range. A focus on the B1–B2 boundary ignores the more significant changes that occur across the B2–C threshold, across which the average proportion of metadiscourse and standard deviation drops by more than one percentage point and the range by five percentage points (Appendix 9). One major caveat for the findings is that minimum values for 11 of the 13 metadiscourse categories are zero for all CEFR levels. That is, they cannot be used to make claims about metadiscourse use across CEFR levels for the Aptis test because they are not essential to effective task completion. Only logical connectives (e.g., and, but) and person markers (I, you, me etc.) contain minimum threshold values other than zero. The findings therefore provide strong evidence for a claim that use of metadiscourse may be analysed separately from vocabulary as part of rhetorical organisational competence (Bachman & Palmer, 2010).

5.7 Summary of findings

The data provides strong evidence for a validation argument that the Aptis writing test discriminates across five CEFR thresholds (A0–C). Twenty-eight metrics are sensitive to variations in candidates’ responses. Seven are basic statistics which relate to the amount of text produced by candidates; one is a metric of lexical diversity; 16 relate to vocabulary content represented by corpus data from the English vocabulary profile (EVP), the British National Corpus (BNC) and Corpus of Contemporary American English (COCA), and four metrics represent metadiscourse use. Higher-ability candidates produce more text than lower-ability candidates. Their responses also exhibit greater lexical diversity. A-level candidates do not have sufficient knowledge of metadiscourse to deploy markers and create coherent linked sentences. In contrast, B-level candidates rely on metadiscourse to create coherent responses. Higher-ability candidates use a greater variety of lexical resources to establish textual coherence, evidenced by declining proportions of metadiscourse in their responses. The findings strongly suggest that the most powerful means of discrimination across CEFR levels in the Aptis test is vocabulary use. The relative importance of observable differences in the dataset is now explored in depth for research question 2.

5.8 Research question 2

To what extent and in what ways are the metrics identified in RQ1 of value, or deficient, for the purposes of automated assessment of learner writing?

5.8.1 Constructing the model

The second research question was addressed using an ordinal logistic regression model. The objective of the analysis is to determine the level of agreement between human raters and classification based on the statistically significant metrics which emerged in relation to RQ1. The dependent variable is CEFR level and the independent variables were 20 variables identified in relation to RQ1. Prior to conducting the analysis, the data was explored for multicollinearity. Corpus type and token figures were all highly intercorrelated, as were metrics associated with text length. The model therefore included metrics associated with a single corpus (Cambridge Learner Corpus) and only one unweighted metric of text length (sentence count). The majority of the metrics are associated with lexis and lexical diversity. The final list of 20 metrics can be viewed in Table 13.

A stratified sample of 510 scripts were used in the analysis. The data was partitioned into training and testing sets. The regression model was built using the training set and evaluated using the data in the test set. The partitioning of the data into training and test sets was random. Of the samples, 75% were used to build the regression model, with the remaining 25% classified using the regression model. Table 13 shows the outcome of the analysis of the training set.

Table 13: Model parameter estimates and intercepts

	OLR					Brant test	
Metric	**Value**	**SE**	***OR***	***t***	***p***	X^2	***p***
Sentence count	0.07	0.02	1.07	3.98	0.00	1.75	0.63
Type-token ratio	-8.49	2.23	0.00	-3.80	0.00	4.48	0.21
Words with more than 2 syllables (%)	-0.10	0.09	0.91	-1.10	0.27	2.39	0.5
Average syllables per sentence	-0.01	0.01	0.99	-0.91	0.36	5.73	0.13
Average syllables per word	0.52	3.08	1.68	0.17	0.87	1.62	0.65
Lexical diversity VOCD	0.01	0.01	1.01	0.85	0.40	3.89	0.27
Lexical diversity MTLD	0.03	0.01	1.03	2.81	0.01	1.68	0.64
Verbal elements per sentence	0.10	0.07	1.10	1.29	0.20	0.02	1
Noun elements per sentence	0.01	0.06	1.01	0.18	0.86	9.48	0.02
A1 token	0.11	0.03	1.11	3.08	0.00	2.11	0.55
A2 token	0.28	0.05	1.33	5.98	0.00	2.13	0.55
B1 token	0.27	0.07	1.31	4.05	0.00	1.03	0.79
B2 token	0.55	0.11	1.74	4.81	0.00	8.11	0.04
C1 token	0.16	0.23	1.17	0.67	0.50	5.95	0.11
C2 token	-0.19	0.44	0.83	-0.43	0.67	4.37	0.22
Metadiscourse (%)	0.01	0.03	1.01	0.21	0.84	1.78	0.62
A1–A2 boundary	9.98	4.92	N/A	2.03	0.04		
A2–B1 boundary	12.24	4.94	N/A	2.48	0.01		
B1–B2 boundary	13.93	4.95	N/A	2.81	0.01		
B2–C boundary	15.67	4.96	N/A	3.16	0.00		

Residual Deviance: 834.6345
AIC: 874.6345

The table displays the value of coefficients, intercepts (the final four rows of the table), the corresponding standard errors and *t* values and the residual deviance (overall distance between data points and the best-fitting line). The coefficients are displayed in log odds units. Taking the exponential of the coefficients gives the odds ratio for each variable. The log odds assign weighting to the values of each sample and added to the intercept to obtain a predicted outcome for the dependent variable. The intercepts indicate where the latent variable is cut to make the five CEFR bands in the dataset. *P* values indicate which of the independent variables are most discriminating across CEFR bands. The value of each coefficient (positive or negative) indicates whether more or less of this variable is associated with higher CEFR bands. Discriminating metrics were sentence count, type-token ratio, one metric of lexical diversity (MTLD) and four metrics of the Cambridge Learner Corpus detailing vocabulary use. The Brant test assesses the proportional odds assumption of OLR. *P* values of greater than .01 indicate that each metric meets the assumption of proportional odds.

5.8.2 Evaluating the model

The regression model was evaluated using the test sample (n = 128). The regression model developed in the previous section was employed to assign a CEFR band to these samples of writing. We then examined the extent of agreement between the original CEFR band awarded to the samples by human raters and the band awarded by the regression model. The outcome is presented in Table 14.

Table 14: Comparison of ratings between human raters and the regression model

	CEFR group	A1	A2	B1	B2	C	Total samples	Proportion correctly assigned	Proportion within 1 CEFR group
Original CEFR group	**A1**	22	7	1	1	0	31	0.71	0.94
	A2	5	7	7	0	2	21	0.33	0.90
	B1	3	7	7	5	2	24	0.29	0.79
	B2	0	1	9	7	8	25	0.28	0.96
	C	0	0	1	10	16	27	0.59	0.96
							128	**0.44**	**0.91**

The regression model achieved an average agreement with human raters of 44% across the five CEFR bands. The model achieved greater success at the highest and lowest bands, with 59% of C band responses and 71% of A1 responses correctly classified. However, this fell substantially for A2–B2 bands, suggesting poor discriminatory power of the metrics used in the model at these levels. Nearly all samples were correctly classified within one CEFR band of that assigned by human raters, suggesting that additional metrics could be used to refine the model to achieve greater success.

6. DISCUSSION

The OLR model used only a single metric associated with text length (sentence count) to avoid over-dependence on text length as a predictor variable which may result in misclassification. The six other discriminating metrics (Table 13) are associated with lexical content of test-taker scripts. While lexis is moderately successful at discriminating the highest and lowest CEFR bands, this was not sufficient to discriminate between A2–B2 scripts. This finding suggests that vocabulary frequency measures are not sufficient to capture all variance between proficiency levels in second language writing, consistent with the findings of Read and Nation (2006), Schmitt (2005) and Albrechtsen, Haastrup and Henriksen (2008). It is also consistent with their claims that other aspects of writing ability such as planning, organisation, coherence of ideas and task completion account for approximately half of the total observable variance. The weaker discrimination between A2–B2 levels may be due to the non-linear relationship of some weighted percentage metrics with CEFR level (e.g. type-token ratio, total verbal elements and percentage of metadiscourse). As weaker responses are shorter, less frequent B1 and B2 lexis accounts for a higher proportion of overall language production in those responses.

The samples used in this study were analysed on the basis of the scores they received in the Aptis test and subsequent alignment to the CEFR (O'Sullivan, 2015). A working assumption of the study was that scoring of the provided samples was reliable and that the samples were representative of the different levels of the scoring rubric. This data is typically used as the basis for optimising computational models of text quality; correlation or level of agreement with computational model is the most widely used indicator of the quality of the predicted scores. Although Text Inspector metrics account for approximately half of the variance in test-taker responses, research suggests that raters' judgments are also based on "some complex and indefinable feeling about the text, rather than the scale content" (Lumley, 2002, p. 263). Raters' impressions of writing samples are also influenced by their previous experience of rating student writing from teaching experience (Xi & Weigle, 2010).

Examining which metrics are sensitive to changes in CEFR levels can be useful for rater training and mitigating the effects of rater bias. Further investigation of rater engagement with scoring rubrics detailing how raters score individual samples and which elements of the rubric they use to score essays is therefore justified. Greater information about rater focus during marking will provide greater context to comparative research of the performance of automated assessment models such as Random Forest (Breiman, 2001) and human raters, and avoid reproduction of rater bias in machine learning algorithms.

7. CONCLUSIONS AND RECOMMENDATION

This research presents a comprehensive picture of lexical and metadiscourse thresholds and profiles of candidates' writing in the British Council Aptis writing test. It adds to the research base underpinning the test by demonstrating that the Aptis test elicits written responses by candidates which can be differentiated across six levels of the CEFR. The research has also specifically considered claims from existing test specifications and task scale descriptors. The following basic metrics were able to discriminate across all levels of the CEFR, with minimum values other than zero associated with each CEFR level. The British Council should consider whether test/task specifications be amended to include guidelines on the following directly observable features:

- token/type count
- syllable count; number of words with more than two syllables
- lexical diversity (voc-d)
- total noun and verb elements.

These findings are probabilistic rather than deterministic, meaning there is a significant amount of overlap across CEFR thresholds within these metrics. However, they could be included in specifications as guidelines. For example, no C-level response contained fewer than nine multisyllabic words and 34 noun elements.

This data reveals that lexical features from the Cambridge Learner Corpus are the most predictive of a learner's level of attainment in the Aptis test. As the corpus was developed from learner writing, this finding is additional evidence that the Aptis test can elicit responses with a range of lexis across five CEFR bands. Response attribute descriptions of lexical frequency within the specifications require updating. The following metrics were statistically significant across all CEFR thresholds investigated in the study:

- proportion of lexis from EVP A1/B2/C1 (types/tokens)
- proportion of lexis from BNC 2K–3K (types); 4K–5K (types and tokens); 6K–7K (types and tokens)
- proportion of lexis from COCA 5K–6K (types); 7K–8K (types and tokens)
- proportion of lexis from the academic word list (types).

Despite changes in vocabulary use accounting for the majority of observable variance in candidate responses, statements such as A2 responses use *"mostly sufficient"* vocabulary to complete Task 2 and "*K4–K5 lexis will be sufficient to complete both emails adequately*" for Task 4 are not supported by the use of lexis in candidate data and do not readily map to CEFR bands.

Data from this project may be used to update existing statements regarding the proportions of lexis from the respective frequency bands which are produced by candidates at each level on average, but minimum values of zero for each CEFR band indicate that lexical use within CEFR bands may be too variable to be of practical benefit for raters. Alternatively, the British Council could consider removing them as they do not offer sufficient guidance regarding how they relate to CEFR bands or task completion.

The Aptis scale descriptors for metadiscourse use may also be updated based on the findings from the report. Specifically, the research revealed that no responses composed of less than 10% metadiscourse received a band of B1 or above (see Appendix 9), indicating that effective metadiscourse use is a crucial component of higher-ability L2 writing in the Aptis test. A minimum of 10.82% metadiscourse is required for a B1 level score.

The research has also shed light on the types of written metadiscourse used by language learners at different levels of the CEFR.

The study presents evidence that analysis of metadiscourse markers independently of vocabulary is justified. Four metrics of metadiscourse use were able to discriminate across all CEFR thresholds – two metrics of emphatic markers and one each of hedge and person markers:

- proportion of emphatic markers (types and tokens)
- proportion of hedge markers (tokens)
- proportion of person markers (types).

This empirical research study of the context validity of the Aptis writing test suggests that Text Inspector and similar other automated software programs may be useful for test developers in several ways. First, using automated tools to analyse candidate responses in the future would ensure greater consistency in grading responses based on observable characteristics of texts produced by candidates. This would ensure comparability across cohorts with regard to those contextual indices which are sufficiently sensitive to capture performance differences across CEFR thresholds. However, examination boards such as the British Council need to reflect upon the findings and determine which metrics they regard as suitable according to their conception of the domain of interest. Index data could be introduced into test task specifications as part of writing response attributes. This could also be used in scale descriptors in addition to detailed descriptors to assist raters. Second, information derived from text analysis tools may also prove to be of considerable value in rater training to help those involved better understand test task characteristics which represent different proficiency levels in the test. Information derived from text analysis tools may also be of value to inform teaching and test preparation practices by raising awareness of the differences in learner writing across levels of the CEFR.

However, developing an automated assessment model that does not require prompt or topic-specific training may therefore require a trade-off in automated assessment between a topic invariant approach versus optimal performance in terms of consistent agreement with human raters. Additionally, this study has shown that some metrics do not have a linear relationship with the level of performance, and that finely-grained measures of complexity may have a minimal effect on performance in test tasks of 200–250 words. Findings presented here may be susceptible to changes in lexical output across CEFR levels in response to different prompts or input material.

Further research should identify which of the metrics presented here are generalisable in further analysis of Aptis writing samples.

7.1 Limitations to the present study

While indices of lexis and metadiscourse provide a good general description of the proficiency levels elicited by the Aptis writing test, comparison across thresholds is based on averages and median values within each level. It is clear from examining the descriptive data in Appendix 9 that examining averages simplifies a complex picture of candidate responses. The large number of samples used in the study from 65 countries mean that the samples within each CEFR level exhibited highly variable data, observable in the large number of outliers present in the boxplot diagrams and large standard deviations and inter-quartile ranges (Appendix 9). Although observable differences are statistically significant, findings are at least partially an artefact of the very large sample size. Studies based on extremely large samples are more likely to produce highly significant outcomes but may have small effect sizes or account for very little overall variance (Meehl, 1990). In particular, step increase in overall percentages of different types of metadiscourse marker across CEFR thresholds is very small. Of the 13 types of metadiscourse marker analysed, 10 rarely exceeded more than 1% of the total response by candidates. Therefore, the practical significance of this finding may be limited as the difference across CEFR bands may be limited to only one or two tokens in candidate responses of 200–250 words. The overwhelming majority of metadiscourse markers used across all levels of the CEFR were logical connectives, emphatic and person markers. The step changes observed in these metadiscourse categories accounts for most of the overall observed variance in metadiscourse use.

Percentage figures, although weighted to account for word length, are not immune to disproportionate magnification of lexical or metadiscourse use at lower levels. The majority of the metrics investigated here were presented as percentages of the overall response. Percentage data is skewed by the limited production of language at A0 and A1 responses, meaning that occurrences of less-frequent (e.g. B1/B2) lexis and metadiscourse account for a disproportionate percentage of written production relative to higher level candidates who produce significantly longer responses.

Finally, the samples used in this study were analysed on the basis of the overall CEFR band they were awarded. The analysis presented here combined the responses by each test-taker from Aptis writing tasks 2 to 4 into a single response for analytical purposes. If metrics are to be included in future rating scales, as part of an automated assessment tool, or rater training, then individual lexical metrics may need to be associated with individual tasks depending on which elements of the writing construct each task is designed to measure. The research did not encompass other, crucial aspects of writing proficiency, such as task completion, or how individual responses cohere with the task input. Therefore, further research is required to investigate how raters engage with writing scale descriptors. For example, research investigating live marking with British Council raters would provide details about how raters score individual samples and which elements of the scale descriptors they use to score essays.

REFERENCES

Albrechtsen, D., Haastrup, K., and Henriksen, B. (2008). *Vocabulary and Writing in a First and Second Language: Processes and Development*. Basingstoke: Palgrave Macmillan.

Bax, S. (2015). *Text Inspector* and other online tools for analysing lexical loads in reading texts. Presentation at CRELLA Research Seminar, University of Bedfordshire (14 March 2015).

Bax, S., Waller, D., and Nakatsuhara, F. (2014). *Researching metadiscourse markers in candidates' writing at Cambridge FCE, CAE and CPE levels.* Cambridge ESOL Funded Research Programme Report, Cambridge ESOL.

Bax, S., Waller, D., and Nakatsuhara, F. (2019). Researching L2 writers' use of metadiscourse markers at intermediate and advanced levels, *System 83,* pp. 79–95. https://doi.org/10.1016/j.system.2019.02.010

Brant, R. (1990). Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. *Biometrics*, 46(4), pp. 1171–1178. https://dx.doi.org/10.2307/2532457

Breiman, L. (2001). Random Forests. *Machine Learning* 45(1), pp. 5–32. https://doi:10.1023/A:1010933404324.

British Council. (2013). Aptis for Teens. http://www.britishcouncil.org/sites/default/files/Aptis_for_teens_brochure_south_asia.pdf

British Council. (2018). Aptis Candidate Guide. https://www.britishcouncil.org/sites/default/files/Aptis_candidate_guide-web.pdf

British Council. (2020). Aptis Candidate Guide: April 2020. https://www.britishcouncil.org/sites/default/files/aptis_candidate_guide_2020_0.pdf

British Council (n.d.) Aptis English Language Test. http://www.britishcouncil.org/exams/aptis

The British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Burnard, L., and Aston, G. (1998). *The BNC Handbook: Exploring the British National Corpus*. Edinburgh: Edinburgh University Press.

Burneikaite, N. (2008). Metadiscourse in linguistics master theses in English L1 and L2. *Kalbotrya*, 59(3): pp. 38–47.

Carlsen, C. (2010). Discourse connectives across CEFR-levels: a corpus-based study. In I. Bartning, M. Martin, and I. Vedder, *Communication Proficiency and Language Development: Intersections between SLA and Language Testing Research*, pp. 191–210. European Second Language Association.

Chapelle, C. A., and Chung, Y-R. (2010). The Promise of NLP and speech processing technologies in language assessment. *Language Testing*, 27(3), pp. 301–315. https://doi.org/10.1177/0265532210364405

Cobb, T. (2020). Web Vocabprofile. An adaptation of Heatley, Nation and Coxhead's (2002) *Range*. www.lextutor.ca/vp/

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press. www.coe.int/en/web/portfolio/the-common-european-framework-of-reference-for-languages-learning-teaching-assessment-cefr

Council of Europe. (2003). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEF). Manual, preliminary pilot version*. Strasbourg: Council of Europe. https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEF). A Manual*. Strasbourg: Language Policy Division. https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr

Council of Europe. (2018). *Collated representative samples of descriptors of language competences developed for young learners (aged 7–10 and 11–15 years). Resource for educators.* Strasbourg: Council of Europe. https://rm.coe.int/collated-representative-samples-descriptors-young-learners-volume-1-ag/16808b1688

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, 34(2), pp. 213–238. https://doi.org/10.2307/3587951

Davies, M. (2009). The 385+ Million Word Corpus of Contemporary American English (1990–present). *International Journal of Corpus Linguistics* 14(2): pp. 159–190. https://doi.org/10.1075/ijcl.14.2.02dav

DuBay, W. H. (2007). *Smart Language: Readers, Readability, and the Grading of Text, Impact Information*. Costa Mesa, California. http://www.impact-information.com/impactinfo/newsletter/smartlanguage02.pdf

Dunlea, J. (2015) Integrating measures of text difficulty and reading ability with a descriptive framework of language proficiency. Paper presented at the 37th Annual Language Testing Research Colloquium (LTRC) Conference, Toronto, Canada. (19 March 2015).

Enright, M. K., and Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, 27(3), pp. 317–334. https://doi.org/10.1177/0265532210363144

Freedle, R., and Kostin, I. (1993). The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items. ETS Research Report, RR-93-13. http://dx.doi.org/10.1002/j.2333-8504.1993.tb01524.x

Freedle, R., and Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, 16 (1): pp. 2–32. https://doi.org/10.1177/026553229901600102

Graesser, A., McNamara, D., Louwerse, M., and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, 36, pp. 193–202. https://doi.org/10.3758/BF03195564

Green, A., and Hawkey, R. (2011). An empirical investigation of the process of writing Academic Reading test items for the international English Language Testing System. *IELTS Research Reports*, Vol. 11. IDP: IELTS Australia, Canberra and British Council, London. https://www.ielts.org/-/media/research-reports/ielts_rr_volume11_report5.ashx

Green, A., Khalifa, H., and Weir, C. J. (2013). Examining textual features of reading texts – a practical approach. Cambridge ESOL: *Research Notes*, 52, pp. 24–39.

Green, A., Unaldi, A., and Weir, C. J. (2010). Empiricism versus connoisseurship: establishing the appropriacy of texts for testing reading for academic purposes. *Language Testing* 27 (3): pp. 1–21 https://doi.org/10.1177/0265532209349471

Hawkey, R., and Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, 9, pp. 122–159. https://doi.org/10.1016/j.asw.2004.06.001

Heatley, A., Nation, I.S.P., and Coxhead, A. (2002). RANGE and FREQUENCY programs. http://www.victoria.ac.nz/lals/staff/paul-nation.aspx.

Hyland, K. (2005). *Metadiscourse*. London: Continuum.

Jones, G. (2015). A validation study of the British Council-EAQUALS core inventory for general English. British Council AR-A/2015/3. www.britishcouncil.org/sites/default/files/glyn_jones_layout.pdf

Khalifa, H., and Schmitt, N. (2010). A mixed-method approach towards investigating lexical progression in Main Suite Reading test papers. Cambridge ESOL: *Research Notes*, 41, pp. 19–25.

Kostin, I. (2004). Exploring item characteristics that are related to the difficulty of TOEFL dialogue items. *ETS Research Report,* 04-11. http://dx.doi.org/10.1002/j.2333-8504.2004.tb01938.x

Laufer, B., and G. Ravenhorst-Kalovski (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, April 2010, Volume 22, 1, pp. 15–30. https://www.jstor.org/stable/43267941

Laufer, B., and Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production, *Applied Linguistics*, 16, pp. 307–322. https://doi.org/10.1093/applin/16.3.307

McCarthy, P. M., and Jarvis, S. (2007). A theoretical and empirical evaluation of voc-d. *Language Testing*, 24, pp. 459–488. https://doi.org/10.1177/0265532207080767

McCarthy, P. M., and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42 (2), pp. 381–392. https://doi.org/10.3758/BRM.42.2.381

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry,* 1, pp. 108–141. https://doi.org/10.1207/s15327965pli0102_1

Morse, J. M. (1995). The significance of saturation. *Qualitative Health Research*, 5, pp. 147–149. https://doi.org/10.1177/104973239500500201

Nissan, S., DeVincenzi, F., and Tang, K. L. (1995). An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension. *ETS Research Report,* RR-95-37. https://doi.org/10.1002/J.2333-8504.1995.TB01671.X

O'Loughlin, K. (2013). Research Summary: Investigating lexical validity in the Pearson Test of English Academic. http://pearsonpte.com/wp-content/uploads/2014/07/OLoughlin_K_2014.pdf

O'Sullivan, B. (2015a). Linking the Aptis Reporting Scales to the CEFR. British Council. www.britishcouncil.org/sites/default/files/tech_003_barry_osullivan_linking_Aptis_v4_single_pages_0.pdf

O'Sullivan, B. (2015b). Formal Trials Feedback Report. British Council. www.britishcouncil.org/sites/default/files/tech_002_barry_osullivan_feedback_v4_0.pdf

O'Sullivan, B., Dunlea, J., Spiby, R., Westbrook, C., and Dunn, K., (2020). *Aptis Technical Manual* Version 2.2. Council. www.britishcouncil.org/sites/default/files/aptis_technical_manual_v_2.2_final.pdf

Osborne, J. W. (2015). *Best practices in logistic regression*. Sage Publications. https://dx.doi.org/10.4135/9781483399041

Read, J., and Nation, P., (2006). An investigation of the lexical dimension of the IELTS speaking test. *IELTS Research Reports*, Vol 6, pp. 207–231. IELTS Australia, Canberra and British Council, London. https://www.ielts.org/-/media/research-reports/ielts_rr_volume06_report7.ashx

Richards, B., and Malvern, D. (2000). *Measuring Vocabulary Richness in Teenage Learners of French.* Paper presented at the British Educational Research Association Conference, Cardiff University, 7–10 September 2000. http://www.leeds.ac.uk/educol/documents/00001541.htm

Sanford, S. G. (2012). *A Comparison of metadiscourse markers and writing quality in adolescent written narratives.* Unpublished PhD thesis, University of Montana.

Schmitt, N. (2005). *Lexical Resources in Main Suite Writing Examinations*. Cambridge ESOL consultant report.

Seedhouse, P., Harris, A., Naeb, R., and Üstünel, E., (2014). Relationship between speaking features and band descriptors: A mixed methods study. *IELTS Research Reports* online series, (2014/2). IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia. https://www.ielts.org/-/media/research-reports/ielts_online_rr_2014-2.ashx

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36 (2), pp. 139–152. https://doi.org/10.1080/09571730802389975

Van Dijk, T. A., and Kintsch, W. (1983). The notion of macrostructure. In T. A. van Dijk & W. Kintsch (Eds.) *Strategies of discourse comprehension*, pp. 189–223. New York: Academic Press.

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S.* 4th edition. New York: Springer.

Vigliocco, G., and Vinson, D. P. (2007). Semantic representation. In *The Oxford Handbook of Psycholinguistics*, pp. 195–215. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198568971.013.0012

Weigle, S. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, 27(3), pp. 335–353. https://doi.org/10.1177/0265532210364406

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Weir, C. J. (2014). A Research Report on the Development of the Test of English for Academic Purposes (TEAP) Writing Test for Japanese University Entrants Eiken and CRELLA, University of Bedfordshire. https://www.eiken.or.jp/teap/group/pdf/teap_writing_report.pdf

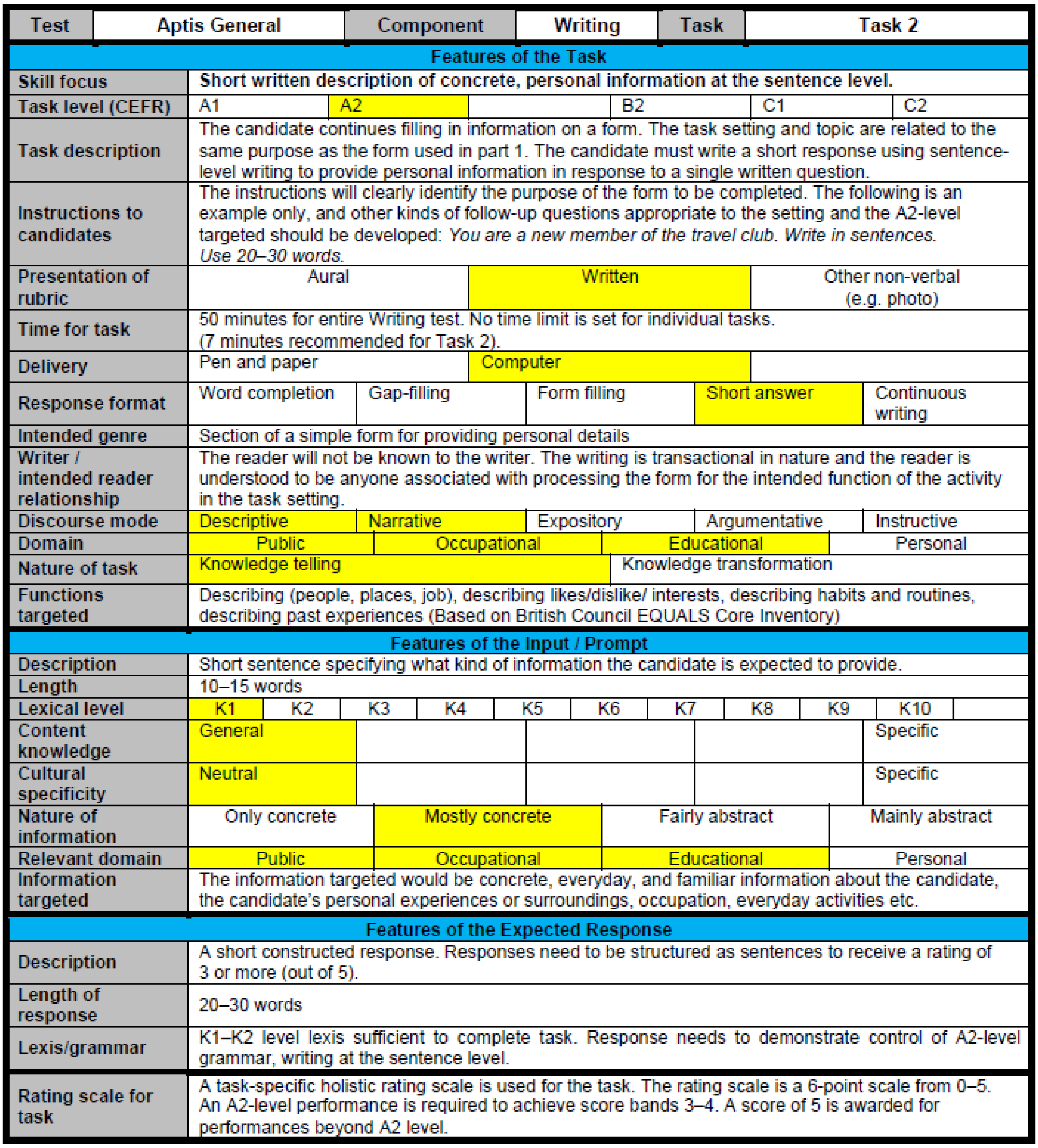
Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models*. Newbury Park NJ: Sage. (Chap. 11, pp. 256–293).

Xia, M., Kochmar, E., and Briscoe, T. (2016). Text readability assessment for second language learners. In *Proceedings of the Workshop on Building Educational Applications Using NLP*, pp. 12–22. https://doi.org/10.18653/v1/W16-0502

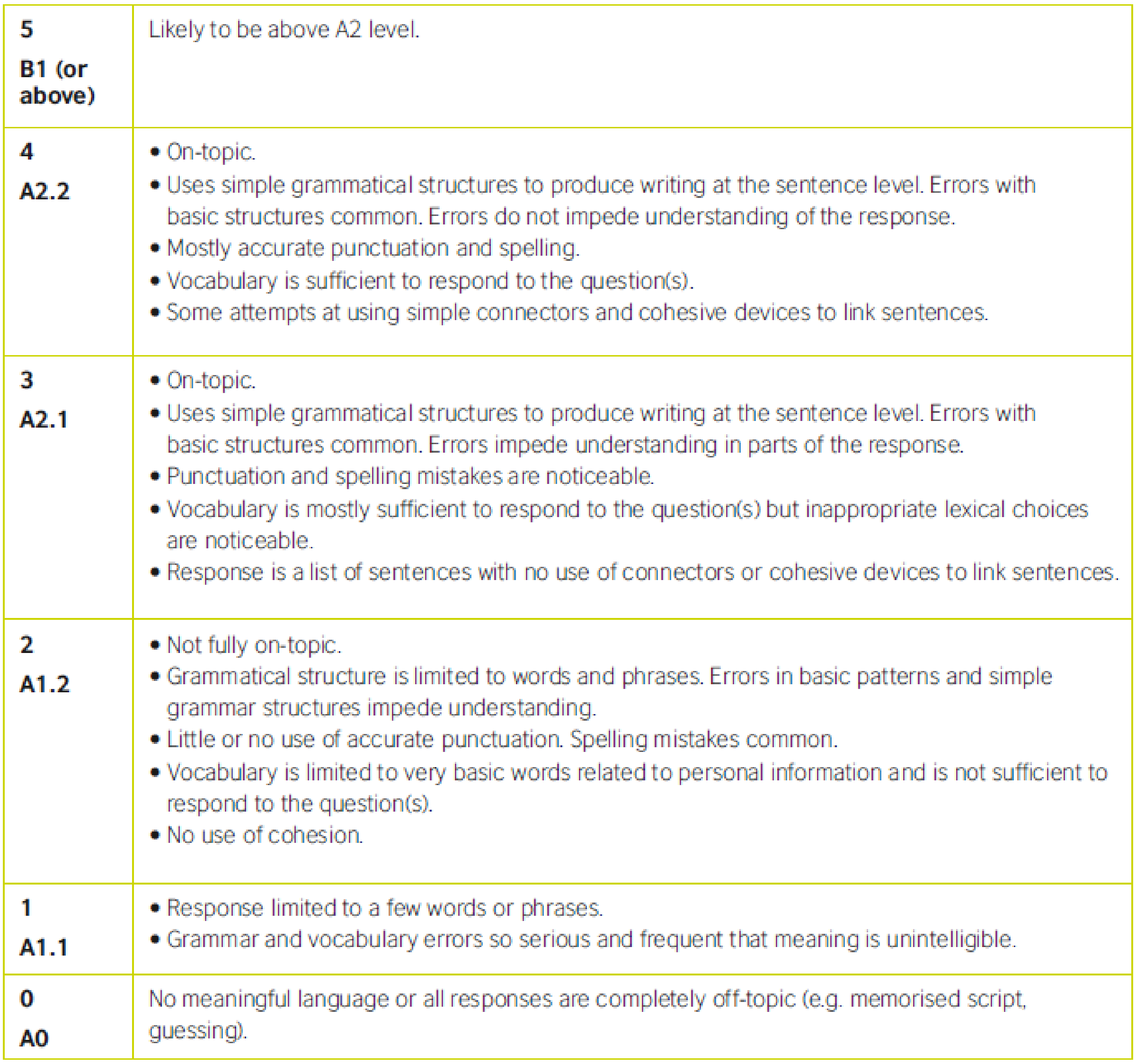
Xue, G., and Nation, I. S. P. (1984). A University Word List, *Language Learning and Communication*, 3 (2), pp. 215–229.

Zheng, Y., and Berry, V. (2015) Aptis for Teens: Analysis of pilot test data. London: British Council. http://www.britishcouncil.org/sites/default/files/aptis_for_teens_layout.pdf

APPENDIX 1:
Writing task 2 specifications and scale descriptors



Test	Aptis General	Component	Writing	Task	Task 2					
Features of the Task										
Skill focus	**Short written description of concrete, personal information at the sentence level.**									
Task level (CEFR)	A1	[✓ A2]		B2	C1	C2				
Task description	The candidate continues filling in information on a form. The task setting and topic are related to the same purpose as the form used in part 1. The candidate must write a short response using sentence-level writing to provide personal information in response to a single written question.									
Instructions to candidates	The instructions will clearly identify the purpose of the form to be completed. The following is an example only, and other kinds of follow-up questions appropriate to the setting and the A2-level targeted should be developed: *You are a new member of the travel club. Write in sentences. Use 20–30 words.*									
Presentation of rubric	Aural	[✓ Written]	Other non-verbal (e.g. photo)							
Time for task	50 minutes for entire Writing test. No time limit is set for individual tasks. (7 minutes recommended for Task 2).									
Delivery	Pen and paper	[✓ Computer]								
Response format	Word completion	Gap-filling	Form filling	[✓ Short answer]	Continuous writing					
Intended genre	Section of a simple form for providing personal details									
Writer / intended reader relationship	The reader will not be known to the writer. The writing is transactional in nature and the reader is understood to be anyone associated with processing the form for the intended function of the activity in the task setting.									
Discourse mode	[✓ Descriptive]	[✓ Narrative]	Expository	Argumentative	Instructive					
Domain	[✓ Public]	[✓ Occupational]	[✓ Educational]	Personal						
Nature of task	[✓ Knowledge telling]	Knowledge transformation								
Functions targeted	Describing (people, places, job), describing likes/dislike/ interests, describing habits and routines, describing past experiences (Based on British Council EQUALS Core Inventory)									
Features of the Input / Prompt										
Description	Short sentence specifying what kind of information the candidate is expected to provide.									
Length	10–15 words									
Lexical level	[✓ K1]	K2	K3	K4	K5	K6	K7	K8	K9	K10
Content knowledge	[✓ General]				Specific					
Cultural specificity	[✓ Neutral]				Specific					
Nature of information	Only concrete	[✓ Mostly concrete]	Fairly abstract	Mainly abstract						
Relevant domain	[✓ Public]	[✓ Occupational]	[✓ Educational]	Personal						
Information targeted	The information targeted would be concrete, everyday, and familiar information about the candidate, the candidate's personal experiences or surroundings, occupation, everyday activities etc.									
Features of the Expected Response										
Description	A short constructed response. Responses need to be structured as sentences to receive a rating of 3 or more (out of 5).									
Length of response	20–30 words									
Lexis/grammar	K1–K2 level lexis sufficient to complete task. Response needs to demonstrate control of A2-level grammar, writing at the sentence level.									
Rating scale for task	A task-specific holistic rating scale is used for the task. The rating scale is a 6-point scale from 0–5. An A2-level performance is required to achieve score bands 3–4. A score of 5 is awarded for performances beyond A2 level.									

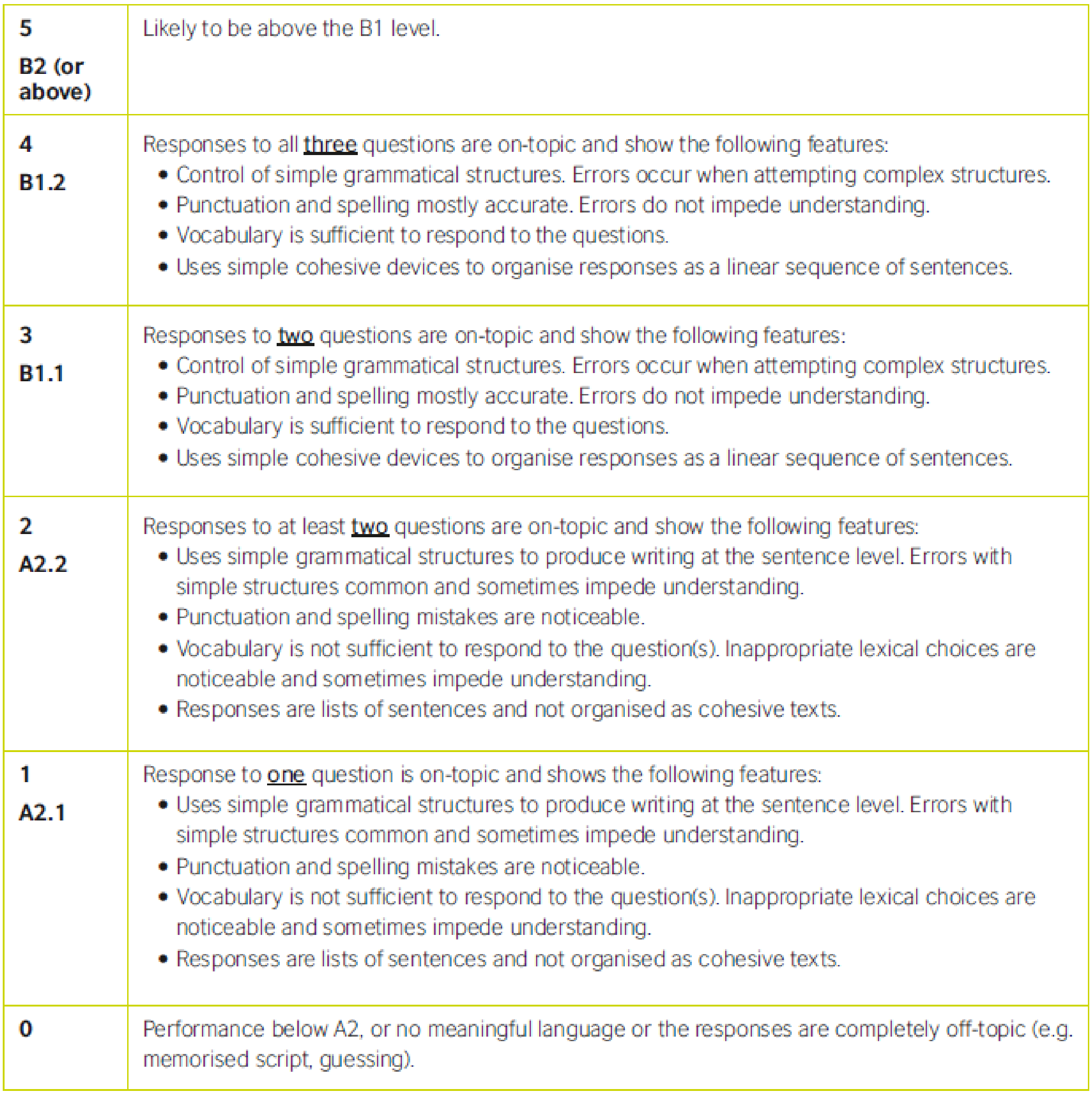


5 **B1 (or above)**	Likely to be above A2 level.
4 **A2.2**	• On-topic. • Uses simple grammatical structures to produce writing at the sentence level. Errors with basic structures common. Errors do not impede understanding of the response. • Mostly accurate punctuation and spelling. • Vocabulary is sufficient to respond to the question(s). • Some attempts at using simple connectors and cohesive devices to link sentences.
3 **A2.1**	• On-topic. • Uses simple grammatical structures to produce writing at the sentence level. Errors with basic structures common. Errors impede understanding in parts of the response. • Punctuation and spelling mistakes are noticeable. • Vocabulary is mostly sufficient to respond to the question(s) but inappropriate lexical choices are noticeable. • Response is a list of sentences with no use of connectors or cohesive devices to link sentences.
2 **A1.2**	• Not fully on-topic. • Grammatical structure is limited to words and phrases. Errors in basic patterns and simple grammar structures impede understanding. • Little or no use of accurate punctuation. Spelling mistakes common. • Vocabulary is limited to very basic words related to personal information and is not sufficient to respond to the question(s). • No use of cohesion.
1 **A1.1**	• Response limited to a few words or phrases. • Grammar and vocabulary errors so serious and frequent that meaning is unintelligible.
0 **A0**	No meaningful language or all responses are completely off-topic (e.g. memorised script, guessing).

APPENDIX 2:

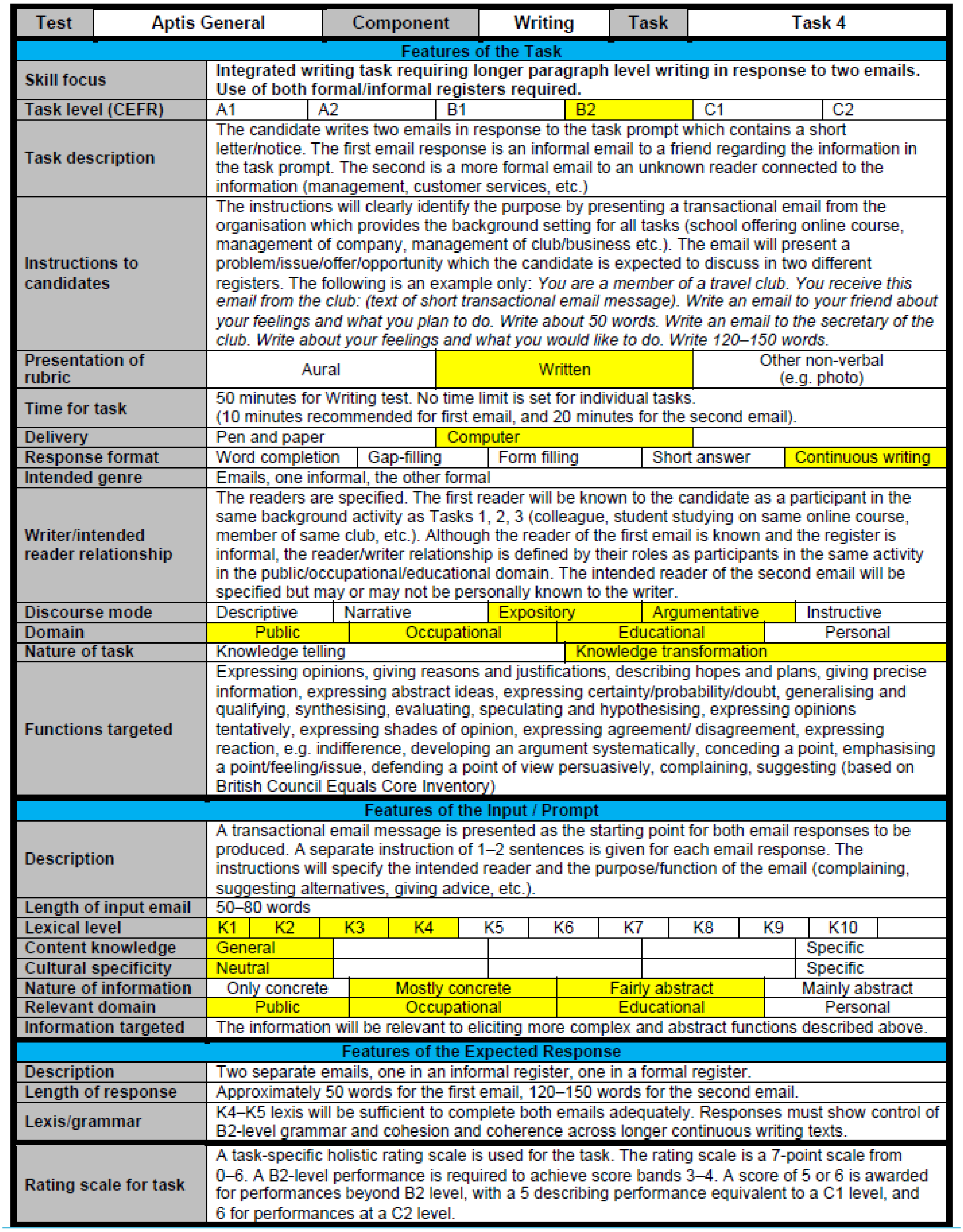
Writing task 3 specifications and scale descriptors

Test	Aptis General	Component	Writing	Task	Task 3					
Features of the Task										
Skill focus	**Interactive writing. Responding to a series of written questions with short paragraph-level responses.**									
Task level (CEFR)	A1	A2	[✓ B1]	B2	C1	C2				
Task description	The candidate responds interactively to three separate questions. Each response requires a short paragraph-level response. The questions are presented as if the candidate is writing on an internet forum or social network site. The task setting and topic are related to the same background activity used in parts 1 & 2.									
Instructions to candidates	The instructions identify the setting for the interaction and person or persons with whom the candidate is interacting. The following is an example only, and other kinds of follow-up questions appropriate to the setting and the B1-level targeted should be developed: *You are a member of a travel club. Talk to other members in the travel club chat room. Talk to them using sentences. Use 30–40 words per answer.*									
Presentation of rubric	Aural	[✓ Written]	Other non-verbal (e.g. photo)							
Time for task	50 minutes for Writing test. No time limit is set for individual tasks. (10 minutes recommended for Task 1)									
Delivery	Pen and paper	[✓ Computer]								
Response format	Word completion	Gap-filling	Form filling	[✓ Short answer]	Continuous writing					
Intended genre	Interaction in a social-media context. The context for interaction may be within the public, occupational, or educational domains, reflecting real-life situations in which interactive, information-exchange forums might be used, but which do not require specialist knowledge or experience (e.g. students in an online course discussing course options, favourite subjects and educational features of the candidate's own educational context).									
Writer/intended reader relationship	The reader will be specified. The reader is not personally known to the candidate but is a participant in the same public/occupational/educational domain. Given the nature of the social media task, the message will be accessible to others.									
Discourse mode	[✓ Descriptive]	[✓ Narrative]	[✓ Expository]	Argumentative	Instructive					
Domain	[✓ Public]	[✓ Occupational]	[✓ Educational]	Personal						
Nature of task	[✓ Knowledge telling]	Knowledge transformation								
Functions targeted	Describing (people, places, job), describing likes/dislike/ interests, describing habits and routines, describing past experiences, describing feelings, emotions, attitudes, describing hopes and plans, expressing opinions, expressing agreement/disagreement									
Features of the Input / Prompt										
Description	Series of 3 prompts phrased as posts requesting information from the candidate by a member of the interactive forum.									
Length of posts	Each post requesting information should be in the form of 1–3 short sentences. Maximum length of a post is 25–30 words, with no one sentence more than 13–15 words.									
Lexical level	[✓ K1]	[✓ K2]	K3	K4	K5	K6	K7	K8	K9	K10
Grammatical level	A2 Grammatical exponents (See Guidelines on Adhering to Grammatical Level)									
Content knowledge	[✓ General]				Specific					
Cultural specificity	[✓ Neutral]				Specific					
Nature of information	Only concrete	[✓ Mostly concrete]	Fairly abstract	Mainly abstract						
Relevant domain	[✓ Public]	[✓ Occupational]	[✓ Educational]	Personal						
Information targeted	The information targeted should be familiar to the candidate and may include talking about the candidate's personal experiences, plans, etc. One question should ask the candidate to describe some aspect of the candidate's own context from a wider a perspective than the candidate's personal experience (describing features of the educational or working context in the candidate's country, subjects typically studied, etc.).									
Features of the Expected Response										
Description	A series of 3 short constructed responses. Each response needs to be structured as sentences, and the candidate must respond adequately to at least 2 questions to receive a rating of 3 or more (out of 5).									
Length of response	30–40 words per response									
Lexis/grammar	K1–K3 level lexis sufficient to complete task. Response needs to demonstrate control of B1-level grammar, writing at the short paragraph level.									
Rating scale for task	A task-specific holistic rating scale is used for the task. The rating scale is a 6-point scale from 0–5. A B1-level performance is required to achieve score bands 3–4. A score of 5 is awarded for performances beyond B1 level.									



5 **B2 (or above)**	Likely to be above the B1 level.
4 **B1.2**	Responses to all three questions are on-topic and show the following features: • Control of simple grammatical structures. Errors occur when attempting complex structures. • Punctuation and spelling mostly accurate. Errors do not impede understanding. • Vocabulary is sufficient to respond to the questions. • Uses simple cohesive devices to organise responses as a linear sequence of sentences.
3 **B1.1**	Responses to two questions are on-topic and show the following features: • Control of simple grammatical structures. Errors occur when attempting complex structures. • Punctuation and spelling mostly accurate. Errors do not impede understanding. • Vocabulary is sufficient to respond to the questions. • Uses simple cohesive devices to organise responses as a linear sequence of sentences.
2 **A2.2**	Responses to at least two questions are on-topic and show the following features: • Uses simple grammatical structures to produce writing at the sentence level. Errors with simple structures common and sometimes impede understanding. • Punctuation and spelling mistakes are noticeable. • Vocabulary is not sufficient to respond to the question(s). Inappropriate lexical choices are noticeable and sometimes impede understanding. • Responses are lists of sentences and not organised as cohesive texts.
1 **A2.1**	Response to one question is on-topic and shows the following features: • Uses simple grammatical structures to produce writing at the sentence level. Errors with simple structures common and sometimes impede understanding. • Punctuation and spelling mistakes are noticeable. • Vocabulary is not sufficient to respond to the question(s). Inappropriate lexical choices are noticeable and sometimes impede understanding. • Responses are lists of sentences and not organised as cohesive texts.
0	Performance below A2, or no meaningful language or the responses are completely off-topic (e.g. memorised script, guessing).

APPENDIX 3:
Writing task 4 specifications and scale descriptors



Test	Aptis General	Component	Writing	Task	Task 4					
Features of the Task										
Skill focus	**Integrated writing task requiring longer paragraph level writing in response to two emails. Use of both formal/informal registers required.**									
Task level (CEFR)	A1	A2	B1	[✓ B2]	C1	C2				
Task description	The candidate writes two emails in response to the task prompt which contains a short letter/notice. The first email response is an informal email to a friend regarding the information in the task prompt. The second is a more formal email to an unknown reader connected to the information (management, customer services, etc.)									
Instructions to candidates	The instructions will clearly identify the purpose by presenting a transactional email from the organisation which provides the background setting for all tasks (school offering online course, management of company, management of club/business etc.). The email will present a problem/issue/offer/opportunity which the candidate is expected to discuss in two different registers. The following is an example only: *You are a member of a travel club. You receive this email from the club: (text of short transactional email message). Write an email to your friend about your feelings and what you plan to do. Write about 50 words. Write an email to the secretary of the club. Write about your feelings and what you would like to do. Write 120–150 words.*									
Presentation of rubric	Aural	[✓ Written]	Other non-verbal (e.g. photo)							
Time for task	50 minutes for Writing test. No time limit is set for individual tasks. (10 minutes recommended for first email, and 20 minutes for the second email).									
Delivery	Pen and paper	[✓ Computer]								
Response format	Word completion	Gap-filling	Form filling	Short answer	[✓ Continuous writing]					
Intended genre	Emails, one informal, the other formal									
Writer/intended reader relationship	The readers are specified. The first reader will be known to the candidate as a participant in the same background activity as Tasks 1, 2, 3 (colleague, student studying on same online course, member of same club, etc.). Although the reader of the first email is known and the register is informal, the reader/writer relationship is defined by their roles as participants in the same activity in the public/occupational/educational domain. The intended reader of the second email will be specified but may or may not be personally known to the writer.									
Discourse mode	Descriptive	Narrative	[✓ Expository]	[✓ Argumentative]	Instructive					
Domain	[✓ Public]	[✓ Occupational]	[✓ Educational]	Personal						
Nature of task	Knowledge telling	[✓ Knowledge transformation]								
Functions targeted	Expressing opinions, giving reasons and justifications, describing hopes and plans, giving precise information, expressing abstract ideas, expressing certainty/probability/doubt, generalising and qualifying, synthesising, evaluating, speculating and hypothesising, expressing opinions tentatively, expressing shades of opinion, expressing agreement/ disagreement, expressing reaction, e.g. indifference, developing an argument systematically, conceding a point, emphasising a point/feeling/issue, defending a point of view persuasively, complaining, suggesting (based on British Council Equals Core Inventory)									
Features of the Input / Prompt										
Description	A transactional email message is presented as the starting point for both email responses to be produced. A separate instruction of 1–2 sentences is given for each email response. The instructions will specify the intended reader and the purpose/function of the email (complaining, suggesting alternatives, giving advice, etc.).									
Length of input email	50–80 words									
Lexical level	[✓ K1]	[✓ K2]	[✓ K3]	[✓ K4]	K5	K6	K7	K8	K9	K10
Content knowledge	[✓ General]				Specific					
Cultural specificity	[✓ Neutral]				Specific					
Nature of information	Only concrete	[✓ Mostly concrete]	[✓ Fairly abstract]	Mainly abstract						
Relevant domain	[✓ Public]	[✓ Occupational]	[✓ Educational]	Personal						
Information targeted	The information will be relevant to eliciting more complex and abstract functions described above.									
Features of the Expected Response										
Description	Two separate emails, one in an informal register, one in a formal register.									
Length of response	Approximately 50 words for the first email, 120–150 words for the second email.									
Lexis/grammar	K4–K5 lexis will be sufficient to complete both emails adequately. Responses must show control of B2-level grammar and cohesion and coherence across longer continuous writing texts.									
Rating scale for task	A task-specific holistic rating scale is used for the task. The rating scale is a 7-point scale from 0–6. A B2-level performance is required to achieve score bands 3–4. A score of 5 or 6 is awarded for performances beyond B2 level, with a 5 describing performance equivalent to a C1 level, and 6 for performances at a C2 level.									

6 **C2**	Likely to be above C1 level.
5 **C1**	Response shows the following features: • Response on-topic and task fulfilled in terms of appropriateness of register. Two clearly different registers. • Range of complex grammar constructions used accurately. Some minor errors occur but do not impede understanding. • Range of vocabulary used to discuss the topics required by the task. Some awkward usage or slightly inappropriate lexical choices. • A range of cohesive devices is used to clearly indicate the links between ideas.
4 **B2.2**	Response on-topic and task fulfilled in terms of appropriateness of register: appropriate register used consistently in both responses. Response shows the following features: • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding. • Minor errors in punctuation and spelling occur but do not impede understanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. • A limited number of cohesive devices are used to indicate the links between ideas.
3 **B2.1**	Response partially on-topic and task partially fulfilled in terms of appropriateness of register: appropriate register used consistently in one response. Response shows the following features: • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding. • Minor errors in punctuation and spelling occur but do not impede understanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. • A limited number of cohesive devices are used to indicate the links between ideas.
2 **B1.2**	Response partially on-topic and task not fulfilled in terms of appropriateness of register: appropriate register not used consistently in either response. Response shows the following features: • Control of simple grammatical structures. Errors occur when attempting complex structures. • Punctuation and spelling is mostly accurate. Errors do not impede understanding. • Limitations in vocabulary make it difficult to deal fully with the task. Errors impede understanding in parts of the text. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
1 **B1.1**	Response not on-topic and task not fulfilled in terms of appropriateness of register. No evidence of awareness of register. Response shows the following features: • Control of simple grammatical structures. Errors occur when attempting complex structures. • Punctuation and spelling is mostly accurate. Errors do not impede understanding. • Limitations in vocabulary make it difficult to deal fully with the task. Errors impede understanding in most of the text. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
0 **A1/A2**	Performance below B1, or no meaningful language or the responses are completely off-topic (e.g. memorised script, guessing).

APPENDIX 4: Sample characteristics

	Country	A0	A1	A2	B1	B2	C	Grand Total
1	Afghanistan	0	0	0	1	1	0	2
2	Albania	1	6	9	57	59	10	142
3	Algeria	2	4	2	8	1	0	17
4	Armenia	1	2	5	9	5	2	24
5	Austria	1	6	26	82	49	3	167
6	Bahrain	0	2	3	8	2	0	15
7	Bangladesh	0	2	2	5	11	3	23
8	Belgium	0	0	0	1	1	0	2
9	Bosnia	0	2	0	3	3	0	8
10	Brazil	1	0	0	11	2	0	14
11	Cambodia	0	10	15	59	27	1	112
12	Chile	0	5	7	41	35	4	92
13	China	12	26	11	33	29	3	114
14	Colombia	12	86	81	187	125	4	495
15	Croatia	0	0	1	0	8	8	17
16	Cyprus	0	1	1	1	2	1	6
17	Czech Republic	1	0	0	2	6	2	11
18	Egypt	40	109	41	207	111	13	521
19	Ethiopia	1	0	0	0	1	1	3
20	France	0	0	0	0	3	0	3
21	Georgia	4	8	6	15	5	1	39
22	Germany	0	2	2	13	10	0	27
23	Ghana	0	0	15	31	33	7	86
24	Greece	0	0	0	0	3	0	3
25	Hong Kong	5	8	6	33	28	6	86
26	India	6	20	42	140	161	20	389
27	Indonesia	0	0	0	3	3	0	6
28	Iraq	0	0	0	0	1	0	1
29	Japan	0	0	1	10	9	0	20
30	Jordan	4	16	16	70	61	7	174
31	Kenya	0	0	10	24	21	18	73
32	Kuwait	0	0	0	1	3	0	4
33	Lebanon	0	0	0	9	10	3	22
34	Libya	0	1	0	0	1	1	3
35	Macedonia	0	1	1	7	5	4	18
36	Malaysia	0	0	1	3	2	0	6

	Country	A0	A1	A2	B1	B2	C	Grand Total
37	Mexico	11	68	91	242	178	12	602
38	Myanmar	2	12	43	129	89	11	286
39	Nigeria	0	0	0	4	29	21	54
40	Oman	0	8	19	27	11	1	66
41	Pakistan	1	5	4	48	41	9	108
42	Palestine	0	1	0	2	3	0	6
43	Philippines	0	0	2	1	4	1	8
44	Poland	1	3	8	17	17	2	48
45	Portugal	0	0	0	2	11	5	18
46	Qatar	1	8	11	13	6	0	39
47	Romania	2	1	0	5	11	2	21
48	Russia	0	0	0	1	2	0	3
49	Saudi Arabia	59	124	68	141	47	6	445
50	Senegal	10	35	17	53	6	0	121
51	Serbia	1	1	0	4	6	2	14
52	Singapore	0	1	6	15	13	3	38
53	Slovenia	0	0	0	0	1	1	2
54	South Africa	0	1	3	5	1	0	10
55	Spain	1	13	31	353	391	40	829
56	Sri Lanka	0	0	2	9	25	5	41
57	Sudan	45	74	15	54	10	0	198
58	Taiwan	4	8	17	151	41	2	223
59	Tanzania	0	0	0	0	2	0	2
60	Thailand	6	23	8	27	16	1	81
61	UAE	0	0	0	2	0	0	2
62	Uganda	0	0	0	21	13	5	39
63	Ukraine	0	0	0	2	3	0	5
64	Uruguay	0	0	0	1	1	3	5
65	Vietnam	1	9	15	89	108	5	227
66	Missing	1	11	24	54	25	6	121
	Grand Total	**237**	**723**	**688**	**2546**	**1948**	**265**	**6407**

APPENDIX 5:
Metrics analysed with Text Inspector

Class of metrics to be investigated	Metrics
Standard measures	Sentence count
	Token count
	Type count
	Syllable count
	Type/token ratio
	Words with more than 2 syllables
	Words with more than 2 syllables (%)
	Average syllables per sentence
	Average syllables per word
	Syllables per 100 words
	Average sentence length
Readability measures	Flesch Reading Ease
	Flesch-Kincaid Grade
	Gunning Fog Index
Lexical diversity	Lexical diversity (VOCD)
	Lexical variance (MTLD)
	Total verb elements
	Total noun elements
	Total verbal elements per sentence
	Total noun elements per sentence
English Vocabulary Profile (EVP)	EVP: Token % for all levels (A1-C2)
	EVP: Type % for all levels (A1-C2)
British National Corpus (BNC) and Corpus of Contemporary American English (COCA) metrics	Type percent (0-1K)
	Type percent (1-2K)
	Type percent (2-3K)
	Type percent (3-4K)
	Type percent (4-5K)
	Type percent (5-6K)
	Type percent (6-7K)
	Type percent (7-8K)
	Type percent (8-9K)
	Type percent (9-10K)

Class of metrics to be investigated	Metrics
	Token percent (0-1K)
	Token percent (1-2K)
	Token percent (2-3K)
	Token percent (3-4K)
	Token percent (4-5K)
	Token percent (5-6K)
	Token percent (6-7K)
	Token percent (7-8K)
	Token percent (8-9K)
	Token percent (9-10K)
Academic Word List (AWL)	AWL all types %
	AWL all tokens %
Metadiscourse markers	Total type and token %
	Announce Goals type and token %
	Attitude marker type and token %
	Code gloss type and token %
	Emphatic type and token %
	Endophoric type and token %
	Evidential type and token %
	Hedge type and token %
	Label stage type and token %
	Logical connective type and token %
	Person marker type and token %
	Relational marker type and token %
	Sequencing type and token %
	Topic shift type and token %

APPENDIX 6: Metadiscourse markers analysed using Text Inspector

Announce Goals (Frame marker)			
here I will	my purpose	the aim	I intend
I seek	I wish	I argue	I propose
I suggest	I discuss	I would like to	I will focus on
we will focus on	I will emphasise	we will emphasise	my goal is
in this section	in this chapter	here I do this	here I will
Code glosses			
put another way	for example	for instance	e.g.
i.e.	that is	that is to say	namely
in other words	this means	which means	in fact
Viz.	specifically	such as	
known as	defined as	called	
Endophorics			
see	noted	discussed below	discussed above
discussed earlier	discussed later	discussed before	section
chapter	fig	figure	table
example	page		
Hedges			
apparently	appear to be	approximately	assume
believed	certain extent	certain level	certain amount
could	couldn’t	doubt	essentially
estimate	frequently	generally	in general
indicate	largely	likely	mainly
may	maybe	might	mostly
often	perhaps	plausible	possible
possibly	presumably	probable	probably
relatively	seems	sometimes	somewhat
suggest	suspect	unlikely	uncertain
unclear	usually	would	wouldn’t
little	not understood	almost	
Logical connectives			
but	therefore	thereby	so
so as to	in addition	similarly	equally
likewise	moreover	furthermore	in contrast
by contrast	as a result	the result is	result in
since	because	consequently	as a consequence
accordingly	on the other hand	on the contrary	however
besides	also	whereas	while
although	even though	though	yet
nevertheless	nonetheless	hence	thus
leads to	or	and	

Relational markers			
incidentally	determine	consider	imagine
by the way	let us	let’s	lets
let	notice	our	recall
note	us	we	you
our	one’s	assume	think about
your			
Attitude markers			
admittedly	I agree	amazingly	unusually
appropriately	correctly	curiously	disappointing
disagree	even	fortunately	have to
hopefully	important	importantly	interest
interestingly	prefer	pleased	must
ought	understandably	remarkable	surprisingly
unfortunate	unfortunately		
Emphatics (Boosters)			
actually	always	apparent	
I believe	certain that	certainly	certainty
clearly	it is clear	conclusively	decidedly
definitely	demonstrate	determine	doubtless
essential	establish	in fact	the fact that
indeed	know	it is known that	must
never	no doubt	beyond doubt	obvious
obviously	of course	prove	show
sure	true	undoubtedly	well known
won’t	even if	should	by far
Evidentials			
literature	according to	cite	cites
quote	established	said	says
points out	points to	point to	point out
argues	argue	claim	claims
believe	believes	suggests	suggest
show	shows	proves	prove
demonstrates	demonstrate	found that	studies
research			
Label stages (Frame marker)			
in conclusion	to sum up	in sum	summarise
summarise	overall	on the whole	all in all
so far	thus far	to repeat	
Person markers			
I	we	me	my
our	mine		
Sequencing (Frame marker)			
first	firstly	second	secondly
third	thirdly	fourth	fourthly
fifthly	next	to begin	to start with
last	lastly	finally	subsequently
two	three	four	five
to conclude			
Topic shifts (Frame marker)			
well	to move on	to look more closely	
to come back to	in regard to	with regard to	to digress

APPENDIX 7:

Descriptive statistics for significant findings

	CEFR	Mean	SE	95% CI Lower Bound	95% CI Upper Bound	SD	Min	Max	Range	IQR	Med.
Sentence count	A0	2.25	0.04	2.17	2.32	2.98	1	23	22	1	1
	A1	5.27	0.07	5.13	5.40	5.43	1	36	35	3	6
	A2	12.16	0.11	11.95	12.38	8.67	1	48	47	10	13
	B1	15.92	0.11	15.70	16.14	8.96	1	52	51	16	14
	B2	21.53	0.10	21.34	21.72	7.66	1	51	50	22	9
	C	23.81	0.08	23.66	23.96	6.17	7	40	33	24	7.5
Token count	A0	54.95	0.74	53.50	56.40	58.99	1	392	391	35	59
	A1	134.18	1.04	132.14	136.21	82.79	1	394	393	123	123
	A2	231.05	1.07	228.96	233.14	85.07	31	642	611	236	132
	B1	291.55	0.85	289.89	293.22	67.82	81	636	555	301	84
	B2	334.42	0.70	333.06	335.78	55.53	105	749	644	332	51
	C	345.40	0.55	344.31	346.48	44.25	243	543	300	339	41.25
Type count	A0	34.67	0.38	33.92	35.41	30.51	1	185	184	26	39
	A1	70.27	0.44	69.41	71.12	34.84	1	186	185	68	50.75
	A2	112.19	0.41	111.38	112.99	32.90	24	269	245	317	176
	B1	138.50	0.34	137.83	139.16	27.02	57	247	190	140	35
	B2	160.47	0.29	159.90	161.04	23.24	61	280	219	160	29
	C	174.60	0.24	174.14	175.07	18.91	127	253	126	173	25
Syllable count	A0	77.87	1.08	75.76	79.98	85.86	2	634	632	49	88
	A1	182.49	1.40	179.74	185.25	112.07	1	545	544	165.5	168.75
	A2	312.84	1.44	310.02	315.67	115.06	40	855	815	317	176
	B1	396.94	1.17	394.65	399.24	93.43	108	846	738	407	115
	B2	464.63	1.01	462.65	466.61	80.75	141	1038	897	460	83
	C	493.39	0.81	491.79	494.98	64.82	315	816	501	489	63
Words with more than 2 syllables	A0	4.13	7.03	3.96	4.30	9.45	0	68	68	3	2
	A1	8.49	7.38	8.31	8.67	4.12	0	43	43	10	6
	A2	15.13	8.29	14.90	15.30	2.72	0	49	49	12	14
	B1	19.97	8.39	19.80	20.20	2.44	2	66	64	11	19
	B2	26.06	9.50	25.80	26.30	2.49	3	80	77	12.5	25
	C	31.93	9.07	31.70	32.20	2.44	9	70	61	12	31
Lexical Diversity (VOCD)	A0	26.55	0.54	25.50	27.61	43.05	0	200	200	0	49.71
	A1	46.11	0.39	45.34	46.89	31.44	0	200	200	46.19	30.51
	A2	63.63	0.23	63.17	64.08	18.38	0	172.37	172.37	63.40	23.87
	B1	71.08	0.20	70.68	71.47	16.05	28.74	148.63	119.89	69.50	20.44
	B2	78.21	0.19	77.84	78.58	14.97	42.85	149.36	106.51	76.78	19.55
	C	86.04	0.17	85.70	86.38	13.95	57.09	130.25	73.16	84.25	17.56
EVP A1 type %	A0	54.83	0.75	51.88	57.77	14.15	0.78	84.00	83.22	18.47	57.33
	A1	61.68	0.24	60.87	62.48	9.97	16.49	95.24	78.75	11.92	62.14
	A2	63.57	0.19	63.00	64.14	7.56	42.67	86.59	43.92	10.12	63.83
	B1	61.03	0.09	60.77	61.28	6.51	36.54	84.85	48.31	8.94	61.26
	B2	58.36	0.10	58.08	58.64	6.26	36.51	78.74	42.23	8.48	58.33
	C	54.27	0.26	53.64	54.89	5.18	41.77	68.57	26.80	7.62	54.21
EVP B2 type %	A0	1.21	0.32	0.83	1.59	1.84	0.00	9.26	9.26	2.33	0.00
	A1	1.12	0.12	1.00	1.23	1.47	0.00	11.21	11.21	1.79	0.62
	A2	1.77	0.12	1.64	1.90	1.72	0.00	9.09	9.09	2.57	1.43
	B1	2.49	0.06	2.41	2.56	1.90	0.00	12.77	12.77	2.45	2.16
	B2	3.63	0.07	3.53	3.72	2.19	0.00	14.21	14.21	2.90	3.27
	C	5.32	0.20	5.04	5.61	2.35	0.55	12.17	11.62	3.43	5.22

EVP C1 type %	**A0**	0.53	0.28	0.34	0.73	0.92	0.00	3.57	3.57	1.04	0.00
	A1	0.58	0.11	0.51	0.64	0.83	0.00	3.95	3.95	1.09	0.00
	A2	0.68	0.11	0.60	0.75	0.96	0.00	8.73	8.73	1.09	0.00
	B1	0.84	0.05	0.80	0.87	0.87	0.00	5.88	5.88	1.32	0.70
	B2	1.12	0.06	1.08	1.17	0.94	0.00	6.55	6.55	1.10	0.90
	C	1.65	0.17	1.52	1.78	1.05	0.00	5.85	5.85	1.51	1.62
EVP A1 token %	**A0**	64.77	0.74	61.86	67.68	13.96	1.54	87.50	85.96	13.11	66.67
	A1	72.45	0.22	71.74	73.15	8.77	17.36	97.62	80.26	9.78	72.95
	A2	75.49	0.18	75.01	75.97	6.39	39.11	93.00	53.89	8.25	75.81
	B1	74.28	0.08	74.07	74.49	5.32	53.25	90.78	37.53	7.32	74.50
	B2	72.48	0.09	72.26	72.70	4.96	55.56	87.23	31.67	6.94	72.59
	C	69.52	0.23	69.03	70.01	4.04	59.44	81.71	22.27	6.13	69.22
EVP B2 token %	**A0**	0.92	0.30	0.60	1.23	1.52	0.00	7.59	7.59	1.59	0.00
	A1	0.77	0.12	0.68	0.86	1.08	0.00	7.47	7.47	1.19	0.32
	A2	1.09	0.11	1.00	1.17	1.14	0.00	6.83	6.83	1.61	0.81
	B1	1.49	0.06	1.45	1.54	1.26	0.00	8.63	8.63	1.52	1.22
	B2	2.16	0.07	2.10	2.23	1.44	0.00	8.90	8.90	1.85	1.89
	C	3.25	0.18	3.07	3.44	1.54	0.30	7.54	7.24	2.27	3.06
EVP C1 token %	**A0**	0.40	0.28	0.25	0.55	0.73	0.00	3.57	3.57	0.66	0.00
	A1	0.40	0.11	0.35	0.45	0.62	0.00	3.39	3.39	0.69	0.00
	A2	0.42	0.10	0.37	0.47	0.63	0.00	5.05	5.05	0.64	0.00
	B1	0.52	0.05	0.50	0.54	0.58	0.00	3.95	3.95	0.81	0.37
	B2	0.69	0.06	0.67	0.72	0.61	0.00	4.18	4.18	0.73	0.60
	C	1.02	0.16	0.93	1.10	0.70	0.00	4.95	4.95	0.88	0.92
BNC type % 2-3K	**A0**	3.15	0.33	2.67	3.64	2.32	0.00	13.51	13.51	2.79	2.82
	A1	3.62	0.13	3.44	3.80	2.20	0.00	12.77	12.77	2.73	3.45
	A2	4.26	0.12	4.10	4.41	2.06	0.00	12.38	12.38	2.57	4.12
	B1	4.78	0.06	4.72	4.85	1.78	0.00	11.05	11.05	2.46	4.72
	B2	5.41	0.07	5.33	5.49	1.74	0.59	11.90	11.31	2.39	5.29
	C	5.95	0.18	5.75	6.15	1.68	2.30	10.56	8.26	2.52	5.81
BNC type % 4-5K	**A0**	0.70	0.30	0.39	1.02	1.50	0.00	10.00	10.00	1.20	0.00
	A1	0.80	0.12	0.71	0.89	1.15	0.00	7.69	7.69	1.43	0.00
	A2	1.17	0.11	1.09	1.26	1.09	0.00	7.37	7.37	1.81	1.01
	B1	1.34	0.06	1.30	1.38	1.02	0.00	6.21	6.21	1.28	1.25
	B2	1.64	0.06	1.59	1.68	1.03	0.00	7.98	7.98	1.55	1.55
	C	1.87	0.17	1.75	1.98	0.96	0.00	4.84	4.84	1.34	1.79
BNC type % 6-7K	**A0**	0.34	0.28	0.19	0.48	0.69	0.00	2.50	2.50	0.00	0.00
	A1	0.45	0.11	0.39	0.52	0.81	0.00	6.00	6.00	0.87	0.00
	A2	0.74	0.10	0.68	0.81	0.83	0.00	4.23	4.23	1.27	0.66
	B1	0.88	0.05	0.85	0.91	0.78	0.00	4.35	4.35	1.35	0.74
	B2	1.05	0.06	1.01	1.08	0.77	0.00	4.23	4.23	0.95	1.01
	C	1.30	0.17	1.20	1.39	0.81	0.00	4.00	4.00	1.15	1.18
BNC token % 4-5K	**A0**	0.56	0.30	0.27	0.84	1.36	0.00	10.34	10.34	0.79	0.00
	A1	0.55	0.11	0.48	0.62	0.87	0.00	5.88	5.88	0.87	0.00
	A2	0.70	0.10	0.65	0.76	0.73	0.00	5.07	5.07	1.08	0.59
	B1	0.78	0.05	0.76	0.81	0.65	0.00	5.17	5.17	0.82	0.66
	B2	0.95	0.06	0.92	0.98	0.64	0.00	4.81	4.81	0.86	0.87
	C	1.15	0.16	1.07	1.23	0.66	0.00	3.34	3.34	0.92	1.09
BNC token % 6-7K	**A0**	0.23	0.27	0.13	0.33	0.49	0.00	2.22	2.22	0.00	0.00
	A1	0.33	0.11	0.27	0.38	0.69	0.00	6.45	6.45	0.48	0.00
	A2	0.48	0.10	0.43	0.52	0.58	0.00	2.78	2.78	0.79	0.32
	B1	0.53	0.05	0.51	0.55	0.50	0.00	2.91	2.91	0.84	0.40
	B2	0.61	0.06	0.59	0.63	0.49	0.00	2.68	2.68	0.61	0.57
	C	0.77	0.16	0.71	0.84	0.51	0.00	2.64	2.64	0.74	0.70
COCA type % 5-6K	**A0**	0.58	0.28	0.38	0.78	0.96	0.00	4.35	4.35	1.15	0.00
	A1	0.81	0.12	0.72	0.90	1.08	0.00	6.58	6.58	1.45	0.00
	A2	1.06	0.11	0.99	1.14	1.01	0.00	6.12	6.12	1.68	0.94
	B1	1.20	0.06	1.16	1.24	0.97	0.00	5.47	5.47	1.24	1.14
	B2	1.41	0.06	1.37	1.45	0.95	0.00	7.21	7.21	1.34	1.30
	C	1.73	0.17	1.61	1.85	0.98	0.00	5.32	5.32	1.18	1.68

COCA type % 7-8K	**A0**	0.47	0.28	0.30	0.64	0.80	0.00	3.70	3.70	1.10	0.00
	A1	0.56	0.11	0.49	0.62	0.81	0.00	6.52	6.52	1.10	0.00
	A2	0.63	0.10	0.58	0.69	0.74	0.00	5.19	5.19	1.02	0.61
	B1	0.88	0.05	0.85	0.91	0.76	0.00	3.70	3.70	1.35	0.74
	B2	1.04	0.06	1.01	1.08	0.74	0.00	4.69	4.69	0.96	1.03
	C	1.30	0.17	1.20	1.40	0.82	0.00	4.12	4.12	1.15	1.16
COCA token % 7-8K	**A0**	0.38	0.27	0.24	0.52	0.67	0.00	2.78	2.78	0.65	0.00
	A1	0.49	0.11	0.43	0.56	0.86	0.00	6.54	6.54	0.79	0.00
	A2	0.51	0.10	0.45	0.56	0.74	0.00	5.93	5.93	0.76	0.30
	B1	0.80	0.05	0.76	0.83	0.82	0.00	5.34	5.34	1.21	0.58
	B2	0.92	0.06	0.88	0.96	0.81	0.00	4.76	4.76	1.03	0.73
	C	1.11	0.17	1.02	1.21	0.78	0.00	3.76	3.76	1.08	0.94
AWL all types %	**A0**	1.28	0.30	0.96	1.59	1.51	0.00	6.00	6.00	2.44	0.54
	A1	1.30	0.12	1.18	1.42	1.53	0.00	10.38	10.38	2.22	1.04
	A2	1.68	0.11	1.56	1.80	1.59	0.00	10.85	10.85	2.57	1.38
	B1	2.43	0.06	2.36	2.50	1.86	0.00	12.35	12.35	2.44	2.11
	B2	3.51	0.07	3.41	3.62	2.32	0.00	11.88	11.88	3.08	3.07
	C	4.96	0.20	4.66	5.26	2.49	0.00	12.93	12.93	3.32	4.74

APPENDIX 8: Kruskal-Wallis test results for all metrics across CEFR thresholds

Text Inspector Metric		CEFR boundary				
		A0–A1	A1–A2	A2–B1	B1–B2	B2–C
1	Sentence count	***	***	***	***	**
2	Token count	***	***	***	***	*
3	Type count	**	***	***	***	***
4	Syllable count	***	***	***	***	***
5	Type/token ratio	***	***	***		***
6	Words with more than 2 syllables	***	***	***	***	***
7	Words with more than 2 syllables – percentage				***	***
8	Average syllables per sentence	***	***		***	
9	Average syllables per word	*			***	***
10	Syllables per 100 words	*			***	***
11	Flesch Reading Ease		***		***	
12	Flesch-Kincaid Grade	**	***		***	
13	Gunning Fog Index	**	***		***	
14	Average sentence length	***	***		***	
15	Lexical diversity (VOCD)	**	***	***	***	***
16	Lexical diversity (MTLD)		***	***	***	***
17	Tokens per type	***	***	***		***
18	Elements	**	***	***	***	***
19	Total verb elements	**	***	***	***	***
20	Verbal elements per sentence	***	***		***	
21	Total noun elements	***	***	***	***	***
22	Noun elements per sentence		***	***	***	
23	EVP A1 type %	***	*	***	***	***
24	EVP A2 type %		***	***	***	***
25	EVP B1 type %		***	***	***	***
26	EVP B2 type %	*	***	***	***	***
27	EVP C1 type %	**	***	***	***	***
28	EVP C2 type %		**		***	***
29	EVP unlisted type %	***	***	***	***	***
30	EVP A1 token %	***	***	***	***	***
31	EVP A2 token %		***	***	***	***
32	EVP B1 token %		***	***	***	***
33	EVP B2 token %	**	***	***	***	***
34	EVP C1 token %	**	***	***	***	***
35	EVP C2 token %		**		***	***
36	EVP unlisted token %	**	***	***	***	***
37	BNC type percent (0-1K)	***	***	**		***
38	BNC type percent (1K-2K)	***	***		***	
39	BNC type percent (2K-3K)	***	***	***	***	***

40	BNC type percent (3K-4K)	***	***		***	*
41	BNC type percent (4K-5K)	***	***	***	***	*
42	BNC type percent (5K-6K)	***	***			
43	BNC type percent (6K-7K)	***	***	***	***	***
44	BNC type percent (7K-8K)	***	***	**	**	
45	BNC type percent (8K-9K)	***		***	***	*
46	BNC type percent (9K-10K)		***			***
47	BNC type percent (Off-list)					
48	BNC token percent (0-1K)	***	***	***		***
49	BNC token percent (1K-2K)	***	***			
50	BNC token percent (2K-3K)	***		***	***	***
51	BNC token percent (3K-4K)	***	***		*	
52	BNC token percent (4K-5K)	***	***	**	***	***
53	BNC token percent (5K-6K)	***	***	***		
54	BNC token percent (6K-7K)	***	***	***	***	***
55	BNC token percent (7K-8K)	***	*	***	***	
56	BNC token percent (8K-9K)	***		***	***	*
57	BNC token percent (9K-10K)		***			
58	BNC token percent (Off-list)					
59	COCA type percent (0-1K)	***	***		*	***
60	COCA type percent (1K-2K)	***	***		***	
61	COCA type percent (2K-3K)	***	***		***	***
62	COCA type percent (3K-4K)	***	***	***	***	
63	COCA type percent (4K-5K)	***	***		***	*
64	COCA type percent (5K-6K)	***	***	*	***	***
65	COCA type percent (6K-7K)		**		***	***
66	COCA type percent (7K-8K)	***	***	***	***	***
67	COCA type percent (8K-9K)		***	***	***	**
68	COCA type percent (9K-10K)	***	***			
69	COCA type percent (Off-list)					
70	COCA token percent (0-1K)	***	***	***		***
71	COCA token percent (1K-2K)	***	***			
72	COCA token percent (2K-3K)	***	***		***	
73	COCA token percent (3K-4K)	***	***	**	***	
74	COCA token percent (4K-5K)	***	***		***	**
75	COCA token percent (5K-6K)	***	***		***	***
76	COCA token percent (6K-7K)		*		***	***
77	COCA token percent (7K-8K)	***	**	***	***	**
78	COCA token percent (8K-9K)		***	***	***	***
79	COCA token percent (9K-10K)	***	***			
80	COCA token percent (Off-list)					
81	AWL all types	**	***	***	***	***
82	AWL all tokens	**	***	***	***	***

Each column tests the null hypothesis that the sample distributions are the same across CEFR groups.
*** = significant at p < .001; ** = significant at p < .01; * = significant at p < .05. Shaded = not significant.
Yellow = most discriminating across all score boundaries. Green = most discriminating across upper score boundaries (A2–C). Blue = most discriminating across lower score boundaries (A0-A2)

APPENDIX 9: Percentages of metadiscourse markers used across CEFR bands

Metric (token percent)	CEFR	Mean	SE	95% CI Lower	95% CI Upper	SD	Min	Max	Range	IQR	Med.
83. Metadiscourse token %	A0	17.22	0.71	15.81	18.63	6.78	0.00	31.28	31.28	8.78	17.31
	A1	20.74	0.21	20.32	21.16	5.20	4.63	35.29	30.66	6.33	21.05
	A2	21.44	0.15	21.15	21.73	3.86	8.87	34.07	25.20	5.10	21.53
	B1	21.49	0.07	21.36	21.62	3.31	10.82	34.50	23.68	4.52	21.53
	B2	20.66	0.07	20.52	20.81	3.25	10.57	31.90	21.33	4.47	20.72
	C	19.50	0.17	19.16	19.84	2.79	11.71	28.02	16.31	3.81	19.54
84. Token Announce goals %	A0	0.02	0.02	-0.02	0.05	0.16	0.00	1.52	1.52	0.00	0.00
	A1	0.13	0.02	0.10	0.16	0.38	0.00	2.68	2.68	0.00	0.00
	A2	0.19	0.02	0.16	0.22	0.42	0.00	3.64	3.64	0.30	0.00
	B1	0.25	0.01	0.23	0.26	0.36	0.00	2.83	2.83	0.38	0.00
	B2	0.33	0.01	0.31	0.34	0.34	0.00	2.02	2.02	0.54	0.30
	C	0.34	0.02	0.30	0.37	0.31	0.00	1.65	1.65	0.55	0.31
85. Token Attitude marker %	A0	0.10	0.04	0.03	0.17	0.36	0.00	2.13	2.13	0.00	0.00
	A1	0.30	0.02	0.26	0.35	0.57	0.00	3.57	3.57	0.47	0.00
	A2	0.39	0.02	0.35	0.43	0.52	0.00	3.23	3.23	0.66	0.00
	B1	0.46	0.01	0.44	0.48	0.49	0.00	3.57	3.57	0.70	0.34
	B2	0.50	0.01	0.48	0.52	0.44	0.00	2.32	2.32	0.48	0.38
	C	0.49	0.03	0.44	0.54	0.43	0.00	2.13	2.13	0.43	0.33
86. Token Code gloss %	A0	0.05	0.02	0.00	0.10	0.23	0.00	1.42	1.42	0.00	0.00
	A1	0.07	0.01	0.05	0.09	0.24	0.00	1.92	1.92	0.00	0.00
	A2	0.13	0.01	0.10	0.15	0.31	0.00	2.53	2.53	0.00	0.00
	B1	0.11	0.00	0.11	0.12	0.24	0.00	1.92	1.92	0.00	0.00
	B2	0.12	0.00	0.11	0.13	0.21	0.00	1.98	1.98	0.28	0.00
	C	0.10	0.01	0.08	0.13	0.20	0.00	1.36	1.36	0.26	0.00
87. Token Emphatic %	A0	0.51	0.08	0.34	0.68	0.81	0.00	3.19	3.19	1.04	0.00
	A1	0.53	0.03	0.46	0.59	0.78	0.00	7.14	7.14	0.88	0.00
	A2	0.74	0.03	0.69	0.80	0.73	0.00	5.68	5.68	1.24	0.62
	B1	0.90	0.01	0.88	0.93	0.68	0.00	4.20	4.20	0.90	0.82
	B2	0.99	0.01	0.96	1.02	0.63	0.00	3.98	3.98	0.77	0.92
	C	1.12	0.04	1.04	1.19	0.61	0.00	2.87	2.87	0.89	1.06

88. Token	Endophoric %	A0	0.16	0.06	0.04	0.28	0.58	0.00	4.17	4.17	0.00	0.00
		A1	0.18	0.02	0.14	0.21	0.45	0.00	4.41	4.41	0.00	0.00
		A2	0.14	0.01	0.12	0.16	0.29	0.00	2.28	2.28	0.00	0.00
		B1	0.13	0.00	0.12	0.13	0.24	0.00	2.97	2.97	0.29	0.00
		B2	0.11	0.00	0.10	0.12	0.20	0.00	1.50	1.50	0.27	0.00
		C	0.07	0.01	0.06	0.09	0.15	0.00	1.01	1.01	0.00	0.00
89. Token	Evidential %	A0	0.07	0.04	0.00	0.15	0.35	0.00	2.86	2.86	0.00	0.00
		A1	0.05	0.01	0.03	0.07	0.22	0.00	2.38	2.38	0.00	0.00
		A2	0.08	0.01	0.06	0.09	0.21	0.00	1.68	1.68	0.00	0.00
		B1	0.12	0.00	0.11	0.13	0.23	0.00	1.63	1.63	0.28	0.00
		B2	0.17	0.01	0.16	0.18	0.26	0.00	1.89	1.89	0.31	0.00
		C	0.26	0.02	0.22	0.30	0.32	0.00	1.59	1.59	0.34	0.24
90. Token	Hedge %	A0	0.30	0.06	0.18	0.43	0.60	0.00	3.30	3.30	0.40	0.00
		A1	0.37	0.03	0.32	0.43	0.65	0.00	3.75	3.75	0.62	0.00
		A2	0.56	0.03	0.51	0.61	0.66	0.00	4.55	4.55	0.87	0.39
		B1	0.72	0.01	0.69	0.75	0.70	0.00	4.22	4.22	1.10	0.59
		B2	0.95	0.02	0.92	0.99	0.74	0.00	4.43	4.43	1.00	0.84
		C	1.18	0.05	1.09	1.27	0.76	0.00	3.74	3.74	1.00	1.06
91. Token	Label stage %	A0	0.06	0.04	-0.02	0.15	0.41	0.00	3.57	3.57	0.00	0.00
		A1	0.02	0.01	0.01	0.04	0.14	0.00	1.43	1.43	0.00	0.00
		A2	0.02	0.01	0.01	0.03	0.13	0.00	1.69	1.69	0.00	0.00
		B1	0.03	0.00	0.02	0.03	0.10	0.00	1.11	1.11	0.00	0.00
		B2	0.05	0.00	0.04	0.06	0.13	0.00	0.96	0.96	0.00	0.00
		C	0.06	0.01	0.04	0.07	0.13	0.00	0.64	0.64	0.00	0.00
92. Token	Logical connective %	A0	4.39	0.36	3.69	5.10	3.39	0.00	17.86	17.86	4.84	4.35
		A1	5.46	0.11	5.25	5.67	2.57	0.00	15.29	15.29	3.52	5.27
		A2	5.46	0.08	5.31	5.61	1.97	0.00	12.71	12.71	2.54	5.22
		B1	5.35	0.03	5.28	5.41	1.71	0.73	13.42	12.69	2.28	5.26
		B2	5.04	0.03	4.97	5.11	1.53	0.44	12.07	11.63	2.04	5.00
		C	4.55	0.08	4.40	4.70	1.24	1.54	9.14	7.60	1.62	4.49
93. Token	Person marker %	A0	7.06	0.51	6.04	8.08	4.90	0.00	20.00	20.00	8.04	6.41
		A1	10.35	0.16	10.05	10.66	3.77	1.09	23.91	22.82	5.38	10.49
		A2	10.19	0.11	9.98	10.40	2.80	1.14	20.00	18.86	3.69	10.32
		B1	9.60	0.05	9.51	9.69	2.34	1.58	18.18	16.60	3.19	9.50
		B2	8.61	0.05	8.52	8.71	2.07	1.93	15.03	13.10	2.88	8.57
		C	7.76	0.11	7.55	7.98	1.77	3.72	12.50	8.78	2.34	7.72
94. Token	Relational marker %	A0	3.45	0.39	2.68	4.23	3.72	0.00	22.86	22.86	3.89	2.90
		A1	2.48	0.09	2.30	2.66	2.23	0.00	13.95	13.95	3.07	2.07
		A2	2.50	0.06	2.38	2.62	1.62	0.00	9.20	9.20	2.14	2.30
		B1	2.73	0.03	2.67	2.78	1.37	0.00	9.85	9.85	1.87	2.62
		B2	2.67	0.03	2.62	2.72	1.17	0.00	7.11	7.11	1.56	2.56
		C	2.49	0.06	2.37	2.61	1.00	0.00	5.60	5.60	1.49	2.49

95. Token	Sequencing %	A0	0.95	0.12	0.71	1.18	1.14	0.00	4.55	4.55	1.67	0.66
		A1	0.70	0.04	0.62	0.78	0.97	0.00	6.12	6.12	1.11	0.39
		A2	0.89	0.03	0.83	0.95	0.81	0.00	3.91	3.91	1.16	0.75
		B1	0.92	0.01	0.89	0.94	0.68	0.00	4.71	4.71	0.93	0.84
		B2	0.89	0.01	0.86	0.91	0.62	0.00	4.60	4.60	0.88	0.82
		C	0.83	0.03	0.76	0.89	0.54	0.00	2.56	2.56	0.75	0.68
96. Token	Topic shift %	A0	0.09	0.04	0.01	0.17	0.40	0.00	3.00	3.00	0.00	0.00
		A1	0.10	0.01	0.07	0.12	0.33	0.00	3.28	3.28	0.00	0.00
		A2	0.15	0.01	0.12	0.17	0.34	0.00	2.64	2.64	0.00	0.00
		B1	0.18	0.01	0.17	0.19	0.31	0.00	2.80	2.80	0.31	0.00
		B2	0.23	0.01	0.22	0.24	0.30	0.00	1.85	1.85	0.34	0.00
		C	0.26	0.02	0.22	0.29	0.29	0.00	1.54	1.54	0.34	0.28

Metric (type percent)		CEFR	Mean	SE	95% CI Lower	95% CI Upper	SD	Min	Max	Range	IQR	Med.
97. Metadiscourse type %		A0	11.06	0.40	10.19	11.94	4.19	0.00	26.67	26.67	5.92	11.63
		A1	12.15	0.15	11.87	12.44	3.52	2.33	27.78	25.45	4.67	11.95
		A2	12.98	0.13	12.78	13.19	2.67	5.48	23.16	17.68	3.45	12.95
		B1	13.41	0.06	13.31	13.50	2.38	5.38	23.61	18.23	3.18	13.33
		B2	13.58	0.07	13.48	13.68	2.17	7.57	22.06	14.49	2.90	13.43
		C	13.54	0.19	13.29	13.79	2.05	8.64	20.26	11.62	2.68	13.51
98. Type	Announce goals %	A0	0.03	0.26	-0.03	0.08	0.25	0.00	2.38	2.38	0.00	0.00
		A1	0.19	0.11	0.15	0.23	0.51	0.00	3.17	3.17	0.00	0.00
		A2	0.26	0.10	0.22	0.29	0.47	0.00	2.38	2.38	0.60	0.00
		B1	0.35	0.05	0.34	0.37	0.44	0.00	2.60	2.60	0.69	0.00
		B2	0.45	0.06	0.44	0.47	0.39	0.00	1.99	1.99	0.67	0.57
		C	0.47	0.16	0.42	0.51	0.37	0.00	2.05	2.05	0.62	0.55
99. Type	Attitude marker %	A0	0.14	0.27	0.04	0.24	0.46	0.00	2.13	2.13	0.00	0.00
		A1	0.41	0.11	0.36	0.47	0.71	0.00	3.80	3.80	0.85	0.00
		A2	0.59	0.10	0.53	0.64	0.70	0.00	4.40	4.40	1.02	0.00
		B1	0.68	0.05	0.66	0.70	0.62	0.00	3.70	3.70	1.09	0.67
		B2	0.75	0.06	0.73	0.78	0.58	0.00	3.07	3.07	0.69	0.66
		C	0.74	0.16	0.67	0.81	0.58	0.00	2.70	2.70	0.65	0.61
100. Type	Code gloss %	A0	0.06	0.26	0.00	0.13	0.29	0.00	2.13	2.13	0.00	0.00
		A1	0.11	0.11	0.08	0.14	0.38	0.00	2.44	2.44	0.00	0.00
		A2	0.20	0.10	0.17	0.23	0.43	0.00	2.53	2.53	0.00	0.00
		B1	0.19	0.05	0.18	0.21	0.36	0.00	2.04	2.04	0.00	0.00
		B2	0.20	0.06	0.19	0.22	0.34	0.00	1.69	1.69	0.55	0.00
		C	0.18	0.16	0.14	0.22	0.32	0.00	1.68	1.68	0.50	0.00

101. Type	**Emphatic %**	**A0**	0.62	0.28	0.42	0.81	0.94	0.00	3.28	3.28	1.35	0.00
		A1	0.71	0.11	0.64	0.79	0.94	0.00	8.33	8.33	1.30	0.00
		A2	1.00	0.10	0.93	1.07	0.89	0.00	5.26	5.26	1.56	0.88
		B1	1.20	0.05	1.17	1.23	0.79	0.00	4.76	4.76	1.01	1.19
		B2	1.32	0.06	1.29	1.35	0.73	0.00	3.97	3.97	1.08	1.27
		C	1.50	0.16	1.41	1.59	0.74	0.00	4.19	4.19	0.87	1.47
102. Type	**Endophoric %**	**A0**	0.18	0.27	0.07	0.30	0.57	0.00	2.70	2.70	0.00	0.00
		A1	0.24	0.11	0.19	0.28	0.53	0.00	2.86	2.86	0.00	0.00
		A2	0.22	0.10	0.19	0.25	0.40	0.00	2.27	2.27	0.00	0.00
		B1	0.21	0.05	0.19	0.22	0.35	0.00	1.69	1.69	0.58	0.00
		B2	0.18	0.06	0.17	0.20	0.30	0.00	1.57	1.57	0.53	0.00
		C	0.13	0.15	0.10	0.16	0.24	0.00	0.72	0.72	0.00	0.00
103. Type	**Evidential %**	**A0**	0.08	0.26	0.01	0.16	0.36	0.00	2.04	2.04	0.00	0.00
		A1	0.08	0.10	0.06	0.11	0.31	0.00	2.33	2.33	0.00	0.00
		A2	0.13	0.10	0.11	0.15	0.32	0.00	2.13	2.13	0.00	0.00
		B1	0.21	0.05	0.20	0.23	0.37	0.00	2.36	2.36	0.56	0.00
		B2	0.29	0.06	0.27	0.30	0.40	0.00	2.13	2.13	0.60	0.00
		C	0.41	0.16	0.35	0.46	0.47	0.00	2.50	2.50	0.62	0.48
104. Type	**Hedge %**	**A0**	0.40	0.28	0.25	0.56	0.74	0.00	2.63	2.63	0.54	0.00
		A1	0.52	0.11	0.45	0.59	0.82	0.00	4.55	4.55	1.06	0.00
		A2	0.83	0.10	0.77	0.90	0.87	0.00	4.21	4.21	1.39	0.77
		B1	1.03	0.05	0.99	1.06	0.88	0.00	4.58	4.58	1.56	0.83
		B2	1.25	0.06	1.22	1.29	0.82	0.00	4.64	4.64	1.15	1.21
		C	1.44	0.17	1.34	1.54	0.83	0.00	4.49	4.49	1.28	1.34
105. Type	**Label stage %**	**A0**	0.07	0.26	-0.01	0.15	0.37	0.00	2.27	2.27	0.00	0.00
		A1	0.04	0.10	0.02	0.06	0.23	0.00	1.96	1.96	0.00	0.00
		A2	0.04	0.10	0.02	0.05	0.20	0.00	1.96	1.96	0.00	0.00
		B1	0.05	0.05	0.04	0.06	0.19	0.00	1.61	1.61	0.00	0.00
		B2	0.09	0.06	0.08	0.10	0.23	0.00	1.44	1.44	0.00	0.00
		C	0.10	0.15	0.08	0.13	0.22	0.00	0.72	0.72	0.00	0.00
106. Type	**Logical connective %**	**A0**	2.54	0.30	2.23	2.84	1.47	0.00	6.38	6.38	2.04	2.47
		A1	3.45	0.12	3.32	3.59	1.67	0.00	12.50	12.50	2.08	3.23
		A2	3.41	0.11	3.33	3.50	1.15	0.00	7.69	7.69	1.42	3.33
		B1	3.34	0.06	3.31	3.38	1.01	0.58	8.20	7.62	1.33	3.29
		B2	3.23	0.06	3.19	3.27	0.92	0.67	7.94	7.27	1.19	3.18
		C	3.10	0.17	2.99	3.21	0.90	1.07	5.65	4.58	1.10	3.08
107. Type	**Person marker %**	**A0**	3.61	0.34	3.07	4.15	2.59	0.00	20.00	20.00	2.61	3.57
		A1	3.69	0.12	3.58	3.80	1.38	0.79	10.26	9.47	1.69	3.42
		A2	3.21	0.11	3.13	3.29	1.04	1.29	10.20	8.91	1.35	3.06
		B1	2.86	0.05	2.83	2.89	0.72	0.77	6.25	5.48	0.96	2.80
		B2	2.55	0.06	2.53	2.58	0.56	0.68	4.85	4.17	0.77	2.53
		C	2.37	0.16	2.31	2.43	0.49	1.12	3.60	2.48	0.65	2.38

108. Type	Relational marker %	A0	2.10	0.31	1.74	2.47	1.76	0.00	7.89	7.89	1.86	1.85
		A1	1.67	0.12	1.57	1.77	1.23	0.00	7.14	7.14	1.36	1.56
		A2	1.71	0.10	1.65	1.78	0.84	0.00	6.45	6.45	1.09	1.69
		B1	1.81	0.05	1.78	1.84	0.74	0.00	4.76	4.76	0.95	1.77
		B2	1.77	0.06	1.74	1.80	0.65	0.00	4.80	4.80	0.89	1.76
		C	1.73	0.16	1.65	1.81	0.65	0.00	4.19	4.19	0.93	1.69
109. Type	Sequencing %	A0	1.11	0.29	0.86	1.37	1.22	0.00	5.17	5.17	2.00	1.16
		A1	0.91	0.12	0.82	0.99	1.11	0.00	6.67	6.67	1.50	0.77
		A2	1.18	0.11	1.11	1.26	0.98	0.00	5.88	5.88	1.16	1.08
		B1	1.23	0.05	1.20	1.26	0.83	0.00	7.14	7.14	1.06	1.18
		B2	1.17	0.06	1.14	1.21	0.76	0.00	4.94	4.94	0.97	1.14
		C	1.04	0.16	0.96	1.12	0.66	0.00	3.35	3.35	0.74	1.04
110. Type	Topic shift %	A0	0.12	0.27	0.02	0.21	0.45	0.00	2.70	2.70	0.00	0.00
		A1	0.13	0.11	0.10	0.16	0.40	0.00	3.23	3.23	0.00	0.00
		A2	0.20	0.10	0.17	0.23	0.39	0.00	2.27	2.27	0.00	0.00
		B1	0.23	0.05	0.22	0.25	0.34	0.00	1.77	1.77	0.60	0.00
		B2	0.30	0.06	0.28	0.31	0.32	0.00	1.85	1.85	0.60	0.00
		C	0.33	0.16	0.30	0.37	0.30	0.00	1.11	1.11	0.60	0.50

British Council Assessment Research Awards and Grants

If you're involved or work in research into assessment, then the British Council Assessment Research Awards and Grants might interest you.

These awards recognise achievement and innovation within the field of language assessment and form part of the British Council's extensive support of research activities across the world.

RESEARCHING LEXICAL THRESHOLDS AND LEXICAL PROFILES ACROSS THE COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES (CEFR) LEVELS ASSESSED IN THE APTIS TEST

AR-G/2021/1

Nathaniel Owen
Prithvi Shrestha
Stephen Bax
Open University

ARAGs RESEARCH REPORTS ONLINE

ISSN 2057-5203