BRITISH COUNCIL

Research

# Evaluating text suitability for English for Academic Purposes reading assessment: Teachers versus Lexile, Coh-Metrix and ChatGPT

Date: 10 November 2025

Author name and ORCID number: Aylin Ünaldi (0000-0003-4119-6700) & Bekir Ateş

Grant reference: RIRG-2018


This report is brought to you by English Language Research, British Council
To cite this report, please use: **Unaldi, A., & Ateş, B. (2025). Evaluating text suitability for EAP reading assessment: Teachers versus Lexile, Coh-Metrix and ChatGPT. British Council.** https://doi.org/10.57884/zass-nw57 or DOI: 10.57884/zass-nw57

# Contents

# Executive summary

The selection of appropriate texts is central to the validity of English for Academic Purposes (EAP) reading assessments. Current practices of text selection rely on both human expertise and automated text analysis tools, each with distinct strengths and limitations. This study investigates how expert EAP teachers and automated tools evaluate textual features, with a focus on identifying areas of alignment and divergence. Drawing on Khalifa and Weir's (2009) framework of contextual features, ten purposefully selected texts were analysed using Lexile, Coh-Metrix, and ChatGPT-5, alongside evaluations by EAP teachers from English core and English-medium instruction pre-sessional programmes. Findings indicate that teachers' judgements of syntactic complexity are strongly shaped by perceptions of lexical complexity, with content relevance and familiarity mitigating the effects of lexical infrequency and abstractness. Automated indices of lexical frequency and syntactic complexity, by contrast, failed to fully account for these interactions. Congruence between teachers' and tools' evaluations was strongest for benchmark texts without marked features, while divergence was evident in cases involving abstractness, subject specificity and stylistic features. The study underscores the need for contextualised, user-centred approaches to automated analysis and highlights how deeper insights into expert judgement can inform improvements in tool design and application.

# Introduction

Selecting appropriate texts for specific test-takers and assessment purposes is central to validity of reading tests. Texts must reflect linguistic, conceptual and disciplinary challenges learners encounter in academic contexts without making the assessment unnecessarily easy or difficult. Ensuring that texts contain desired target features is therefore crucial for test validity. When selecting texts for a target group of readers, several textual features need to be considered (Khalifa & Weir, 2009). Assessment frameworks such as the Common European Framework of Reference for Languages (CEFR) offer broad proficiency descriptors, but they do not provide detailed enough text complexity descriptions for different proficiency levels. Certain readability calculations based on word and sentence length can be done using facilities available online. Nevertheless, in most cases the actual task of selecting texts to match the proficiency levels of test-takers usually rests on teachers' professional judgement. On the other hand, recent advances in natural language processing (NLP) have enabled quick and reliable analysis of texts across multiple linguistic and discourse features supporting estimates of text complexity (Crossley et al., 2017; Green et al., 2013).

Numerous studies have examined the predictive ability of machine-generated complexity estimates, often statistically aligning results and expert raters' holistic judgement or comprehension-based performance scores (e.g. Benjamin, 2012; Crossley et al., 2017). While these studies are valuable for validating the efficiency of automated methods, they tend to treat human judgement as a benchmark without probing into its underlying components. Human raters bring nuanced insight into learner needs and contextual demands through their professional experience and understanding of assessment standards. We believe it is important to explore in greater depth how experts evaluate different textual features when judging the appropriateness of texts for assessment, and also to assess the extent to which automated tools reflect the sophistication of this reasoning. This will not only enhance the informed use of automated tools but also facilitate their future improvement. To this end, this study compares EAP teachers' evaluation of text suitability and the analyses of automated tools such as Lexile, Coh-Metrix and Chat GPT.

## 1.1 Textual features as part of reading construct

One of the earlier works that propose a taxonomic view of textual features in language assessment is Bachman's (1990) Communicative Language Ability framework. Bachman et al. (1995) applied this framework to analyse Cambridge-TOEFL input texts in terms of length, propositional content (vocabulary, distribution of new information, topic, genre), organisational characteristics (grammar, cohesion, rhetorical organisation) and pragmatic characteristics (illocutionary force, sociolinguistic characteristics). Weir (2005: 44) incorporates textual features into his concept of context validity, emphasising that reading processes cannot be operationally defined without clearly defining textual features. His

linguistic task demands include discourse mode, channel, text length, writer–reader relationship, nature of information, content knowledge, lexical, structural and functional. Khalifa and Weir (2009) applied this framework to the validation of Cambridge exams showing an increase in textual complexity across proficiency levels.

The CEFR (Council of Europe, 2001) offers broad level descriptors implying grammatical and lexical complexity as well as concreteness/abstractness, but lacks detail on linguistic, structural and discourse-level characteristics of texts, particularly those found in academic genres. The US Common Core State Standards provide more specific, skill-aligned guidance, covering quantitative (word/sentence length, frequency), qualitative (purpose, structure) and reader/task-based dimensions.

This review highlights the centrality of textual features to reading comprehension and provides a theoretical basis for analysing EAP text complexity whether it is done by human raters or automated tools. While widely used methods readability formulas such as Flesch-Kincaid Reading Ease (FKRE) or Grade Level (FKGL) offer simple, accessible measures, they are limited proxies, lacking comprehensive coverage of construct and criterion variables of text complexity (Sheehan et al., 2010). Section 1.2 reviews cognitively grounded approaches to text complexity that incorporate key textual features and discourse-level dimensions.

### 1.2 Automated text analysis tools and prediction of text complexity

Automated text analysis offers significant potential to support teaching and assessment practices given the abundance of available texts and lack of clear criteria for appropriateness. Substantial efforts have been directed to exploring automated text analysis methods to objectively quantify text complexity and encompass the cognitive and linguistic processes of text comprehension at higher levels, such as situation model formation (Dowell et al., 2016). For example, Coh-Metrix (McNamara et al., 2014) can compute several lexical, syntactic and discoursal features, such as propositional density, argument overlap (cohesion and coherence) and syntactic complexity. TAALES (Kyle & Crossley, 2017) and TAASSC (Kyle, 2016) parse the texts for lexical and syntactic characteristics that can identify usage-based linguistic features (such as verb–argument constructions), while Lexile (MetaMetrics, 2018) measures semantic difficulty (vocabulary frequency) and syntactic complexity (sentence length) to determine text difficulty when scaling texts for educational levels. The efficiency of the automated text analysis based on NLP has been verified in several studies through sophisticated statistical modelling against human scaling of texts. This means that human judgement is the foundation of text complexity evaluation, and understanding how experts perceive and rate textual features is essential for validating and improving automated tools.

In attempts to develop strong models of text difficulty prediction, research has explored sophisticated statistical approaches, yet paid limited attention to the details of human

judgement. For example, Sheehan et al. (2010) applied innovative modelling to texts classified by Lexile and experienced educators and identified four key predictors of complexity – academic orientation, syntactic complexity, semantic difficulty and negation – but did not investigate whether educators apply constructs comparable to those in automated analyses. Lexile (MetaMetrics, 2018) is also widely used to quantify text complexity and to grade texts according to reader ability; examples include Fitzgerald et al.'s (2016) gradation of ESL reader series and Williamson et al.'s (2016) comparison of Lexile levels of university books across the UK and USA, though Williamson et al. caution us that results are restricted to the selection of the texts in the analysed corpus. Nelson et al. (2012) found broad alignment between experts' holistic estimates, student scores and automated metrics despite certain variations by text type and grade level. However, they did not isolate the individual contributions of experts or metrics. Crossley et al. (2017) expand the approach, integrating reader comprehension, text processing and text familiarity variables with NLP-derived features in their model, outperforming traditional readability formulas, yet did not explain how readers form perceptions of complexity. These studies demonstrate the potential of automated text analysis, but show limited incorporation of human insight.

### 1.3 Automated analysis in L2 reading assessment

In second-language (L2) contexts, estimating text difficulty is complicated by varying acquisition rates, reader profiles and purposes, as well as the lack of large, well-annotated L2 corpora for robust statistical analysis (Xia et al., 2019). Nevertheless, in L2 reading assessment, similar text complexity analyses are used to evaluate the complexity of input texts. Green et al. (2013) analysed Cambridge exam texts, identifying cohesion and lexical features as key factors at different proficiency levels, but noted uncertainty in human text classifications and the construct coverage of the text analysis tools. Hamada (2015) used Coh-Metrix with Eiken passages, finding lexical and syntactic features predicted test scores more strongly than cohesion. Choi and Moon (2019) showed vocabulary and syntax strongly predicted both observed scores and experts' predictions. In both cases, expert judgements were treated holistically without detailed analysis. As in L1 reading research, L2 text analysis has yet to yield conclusive findings.

### 1.4 Large Language Models and Generative Pre-trained Transformers

A further development in text difficulty estimation is the integration of NLP in machine learning. Xia et al. (2019) ranked Cambridge exams texts into CEFR levels, treating several textual features (traditional, lexico-semantic, parse-tree syntactic, language modelling, discourse-based) as a machine learning problem in explaining readability. While more effective than earlier models, they did not identify which features may best explain CEFR levels. Balyan et al. (2020) used hierarchical algorithms to scale practice texts in iSTART, achieving increased accuracy with NLP indices than traditional readability formulas.

Large Language Models (LLMs) have transformed NLP with their ability to generate human-like text and perform diverse tasks (Benedetto et al., 2025). In language assessment, they are increasingly used for reading comprehension item generation and text production (for example Duolingo in Attali et al., 2022). Generative Pre-trained Transformers (GPTs), such as ChatGPT, can translate, generate and adapt texts without retraining, reducing reliance on human expertise and lowering costs. Hence, the emergence of research on GPTs' efficiency has been rapid. For example, Imperial and Madabushi (2022) compare the characteristics of LLM-produced stories with those of prompts, finding inconsistencies in terms of linguistic features. The researchers suggest building user-centric complexity assessment models tailored to specific groups of learners and their language abilities. Bezirhan and von Davier (2023) emphasised careful prompt design and human editing for the use of GPT-3 generated texts. Imperial et al. (2024) underline that domain experts in several fields follow strict standards for producing content. To optimise the performance of language models in generating content for language assessment, they propose an in-context learning framework that guides LLMs in aligning text generation with the CEFR and Common Core Standards. Benedetto et al. (2025) questioned whether LLMs can understand and leverage specific pedagogical requirements; specifically, whether they have knowledge of the CEFR and can apply it in readability classification (as well as in other tasks). The authors caution that LLM-based predictors cannot be directly used in educational applications, as their readability classification errors were high.

The increased use of NLP and LLM-based predictors in language assessment has created pressure to understand whether they can perform educational tasks not only with human-level accuracy but also with the ontology required to meet pedagogical standards. High variability in their performance and the opacity of the errors they make require us to critically evaluate their dependability, interpretability and alignment with educational standards before fully integrating them into high-stakes assessment contexts.

### 1.5 Present study

This study adopts a different perspective on textual complexity analysis by focusing on the multifaceted nature of human evaluation rather than treating it as a single holistic judgement. Automated tools typically rely on texts already classified by human judges, but these judgements are usually reduced to one variable, without considering the interplay of linguistic, cognitive and contextual factors that shape expert perceptions of difficulty. The aim here is not to test the predictive accuracy of automated tools but to explore the complexity of human judgement in order to guide improvements in both tool development and practitioner use. So, we aim to show where tools can differ from human evaluation – where they can err – in the calculations of textual complexity. The study compares EAP teachers' evaluations of textual features such as topic, lexis and syntax with parameters generated by automated tools. While test developers normally consider whether a text lends itself to the assessment of certain reading skills and whether enough items could be

produced for the text, teachers in this study were asked to focus solely on textual features. The guiding research question is: How do expert judges' evaluations of selected textual features compare with automated text analysis?

# Method

## 2.1 Preparation of the qualitative text analysis tool: Text Analysis Questionnaire

In this study, the reading proficiency level is determined as pre-university EAP exit level, so the target reading context is first-year university reading as it is the case with many pre-university EAP tests such as IELTS. Texts were drawn from first-year academic books and standardised tests, with attention to the contextual features in Khalifa and Weir (2009). A subset was analysed using automated tools, and ten extracts were selected or adapted to represent these features in both favourable and unfavourable forms (e.g. simple vs complex vocabulary and syntax). In brief, Text 1 was a narrative text, representing a non-EAP genre. Texts 2, 5 and 10 represented benchmark texts with automated index measures approximating the average values in Green et. al. (2010) and McNamara et al. (2014) (see below). Text 3 included low-frequency vocabulary but simple syntax. Text 4 had average syntactic and lexical difficulty values but was quite abstract in content. Text 6 had average lexical but high syntactic difficulty. Text 7 represented field-specific subject matter. Text 8 was edited to reduce coherence by breaking the reference relations and shifting the topic in mid-paragraph. Text 9 represented a culturally specific article written in a journalistic style. Table 1 presents the numerical analyses of the texts' features as determined by automated text analysis tools. The last column in Table 1 shows Coh-Metrix norm values from McNamara et al. (2014) for Social Studies texts at the K11–College level, which we use as reference values for interpreting text complexity. For further details on how complexity changes across grade levels, the reader is referred to that source. This sample of ten texts was not intended to represent all possible features of target texts but rather to implement specific textual features to assess the sensitivity of both judges and tools to these features.

The questionnaire task had 13 questions tapping into certain textual features that were derived from the contextual parameters in Khalifa and Weir (2009). The first question (q1) was based on text purpose; the second question (q2) asked for identification of the source of the text. Questions 3–11 were five-point Likert scale questions, 1 representing 'least difficult' and 5 'most difficult'. The other questions were: **q3 audience:** whether the text is intended for general or expert readers; **q4 grammar:** syntactic complexity; **q5 vocabulary:** lexical complexity; **q6 concreteness of the information:** whether the text includes mostly concrete, factual information or abstract discussion; **q7 information density:** whether the text presents many main ideas in a relatively short span or not; **q8 topic specificity:** whether comprehension of the text requires specific field knowledge; **q9 cultural specificity:** whether the text includes several culture-specific references; **q10 sentence cohesion:** whether the sentences are connected to each other explicitly; **q11 coherence:** whether the flow of the ideas in the text can be followed easily.

**Table 1:** Automated analysis of questionnaire texts with Coh-Metrix and Lexile

| | Text1 | Text2 | Text3 | Text4 | Text5 | Text6 | Text7 | Text8 | Text9 | Text10 | Ref values |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Sound of Shell** narrative | **Globaliza-tion** ideal | **Pastoralism** dif. vocab | **Morality** abstract | **Internet** ideal | **Coal** gram. diff. | **Reading** topic spec. | **Food Prod.** incoherent | **Football** culture spec. | **Meteorite Impact** ideal | **11–college** |
| Text easability | | | | | | | | | | | |
| Narrativity% | 61.41 | 6.55 | 11.7 | 39.74 | 28.77 | 24.2 | 65.54 | 6.81 | 37.83 | 12.92 | 25.89 |
| Syntactic simplicity% | 64.06 | 52.39 | 74.86 | 46.02 | 61.41 | 0.84 | 19.77 | 57.53 | 9.18 | 50 | 47.31 |
| Word concreteness% | 81.86 | 79.67 | 95.05 | 6.55 | 11.31 | 99.16 | 16.11 | 86.86 | 43.64 | 1.83 | 51.25 |
| Referential cohesion% | 11.12 | 57.93 | 14.69 | 73.89 | 4.18 | 90.66 | 53.98 | 4.27 | 21.48 | 44.83 | 39.6 |
| Deep cohesion% | 14.01 | 36.69 | 31.92 | 88.69 | 89.8 | 94.06 | 91.62 | 77.04 | 50.8 | 50.8 | 60.03 |
| Lexical diversity MTLD, all words | 77.04 | 72.62 | 84.29 | 49.75 | 107.64 | 103.5 | 80.77 | 177.49 | 113.06 | 86.3 | 84.31 |
| Celex log freq. for all words, mean | 2.94 | 2.78 | 2.85 | 3.01 | 3.06 | 2.87 | 2.95 | 2.74 | 3.04 | 2.78 | 2.99 |
| Flesch reading ease | 86.29 | 29.02 | 46.85 | 42.2 | 54.48 | 38.15 | 64.2 | 33.75 | 50.25 | 37.64 | 49.06 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Flesch-Kincaid grade level | 4.1 | 13.91 | 10.31 | 11.87 | 10.4 | 16.51 | 9.52 | 13.27 | 12.05 | 12.73 | 11.43 |
| Coh-Metrix L2 readability | 5.47 | 13.16 | 4.86 | 20.66 | 13.77 | 9.02 | 15.76 | 8.24 | 13.12 | 11.98 | 14.04 |
| Lexile | 800–900 | 1200–1300 | 1100–1200 | 1100–1200 | 1100–1200 | 1400–1500 | 1100–1200 | 1200–1300 | 1300–1400 | 1200–1300 | |

Question 12 (q12) asked for an overall difficulty judgement of the text in question, and Question 13 (q13) was a Yes/No question for an overall judgement of whether the text was suitable for an EAP test or not (see Appendix A for the texts and the full forms of the questions). The teachers frequently responded in the comments section, especially after q13, explaining their evaluations. The majority trends are briefly reported and used as the basis for interpretations of the results.

## 2.2 Automated text analysis tools

We used two freely available automated text analysis tools: Coh-Metrix (McNamara et al., 2014) for genre, lexical, syntactic and cohesion features, and Lexile (Stenner et al., 2007) for lexical and syntactic features. Table 2 links the linguistic task demands in Khalifa and Weir's (2009) contextual validity framework to comparable automated indices. We prioritised easily interpretable, research-supported measures (for example narrativity) and limited the selection due to the study's scope.

**Table 2:** Contextual features and automated text analysis indices

| Khalifa & Weir's (2009) Linguistic task demands | Tools/indices |
|---|---|
| Discourse mode/overall text purpose | - MAT tagger<br>- Cohmetrix narrativity |
| Grammatical resources and readability indices | - Lexile score<br>- Coh-metrix L2 readability<br>- Flesch Kincaid Grade Level<br>- Cohmetrix syntactic simplicity<br>- Coh-metrix left embeddedness<br>- Coh-metrix number of modifiers per noun phrase<br>- Coh-metrix noun phrase density |
| Lexical resources | - Coh-metrix CELEX Log frequency for all words<br>- Coh-metrix Concreteness for content words |
| Nature of information | - MAT tagger<br>- Cohmetrix narrativity<br>- Coh-metrix Concreteness for content words<br>- Coh-metrix deep cohesion |
| Content knowledge | - Not available |
| Reader writer relationship | - Coh-metrix narrativity<br>- Coh-metrix deep cohesion |

## 2.3 LLM: ChatGPT-5

OpenAI's ChatGPT is one of the most sophisticated and widely used LLMs. It can integrate human feedback and align its output with user goals, so it can be fine-tuned for specialised

tasks. When it is used in text generation and the evaluation of text complexity, alignment with pedagogical standards is essential. To align the difficulty of an LLM's language output with the proficiency of a target learner group, we applied carefully engineered prompts and a difficulty model (Kogan et al., 2025). Following Imperial and Madabushi (2022), we adopted the CEFR as an expert-defined standard to guide text evaluation and implemented sample prompts with explicit instructions as recommended by Bezirhan and von Davier (2023). In general, the procedure followed in this study can be described as standard-based, criterion-driven multistep protocol that aims at reflecting experts' reasoning process in evaluating the appropriacy of texts for EAP assessment. The details are as follows.

Step 1: Establishing mental representation of standards: knowledge grounding via standard extraction; forms baseline for comparisons

> Extract the reading-related descriptors for B2–C1 levels from CEFR Companion Volume.
>
> Create operational definitions for each level.
>
> Map readability scores to CEFR levels.
>
> Create a structured list of textual features that can be used when selecting or adapting texts for an EAP reading comprehension test at B2–C1 levels.

Step 2: Benchmark calibration: few-shot learning through representative exemplars; builds internal reference for similarity judgements

> Analyse 30 IELTS texts using the structured list of textual features.
>
> Analyse 30 extracts from first-year university coursebooks using the structured list of textual features.
>
> Identify common features across IELTS and university coursebook texts.

Step 3: Criterion-driven (13 questionnaire questions) scoring of the questionnaire texts: in-context application of learned criteria for scoring and classification

> The texts used in EAP tests typically share textual characteristics with IELTS passages and first-year university coursebook extracts, such as those you have analysed previously. They also align with the text descriptions in the CEFR B2–C1 level descriptors. Analyse each text using your structured list of textual features and the common features identified across IELTS and university coursebooks when evaluating its characteristics. Then, rate each text according to the given 13 questions.

Figure 1 summarises the protocol, which in effect can be taken as a cognitive model where each step builds internal representation that is needed to make reliable and justifiable

judgements on the complexity and the appropriacy of each text for EAP reading assessment.

When we began the analysis, the latest available version of ChatGPT was 4. We applied the protocol three times using ChatGPT-4, but observed some inconsistencies in the results. Shortly thereafter, ChatGPT-5 was released. We ran the protocol twice with ChatGPT-5 (Plus), obtaining relatively consistent results. In this study, we report the findings from the second ChatGPT-5 run. The table in Appendix B presents the results from the third ChatGPT-4 run and both ChatGPT-5 runs.

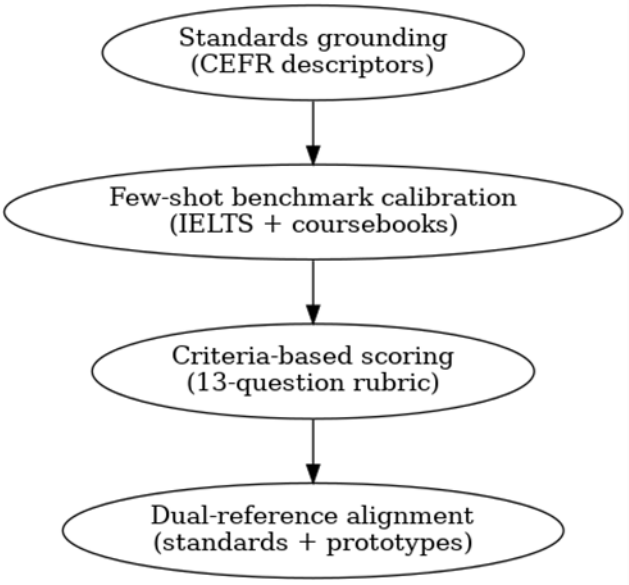


**Figure 1:** Standard-based multistep text evaluation protocol for EAP assessment

## 2.4 Participants

EAP teachers in university pre-sessional courses are ideally suited to evaluate the appropriateness and representativeness of a text for assessment at pre-university stage as they are responsible for preparing students for both entry exams and the academic demands of the first year at university. In this study, 32 EAP teachers from the School of Foreign Languages at a reputable Turkish EMI (English-medium instruction) university and 31 participants from five UK-based British universities participated. Except for one teacher with two years' experience, all had between seven and 40 years (M = 19.93, SD = 9.13). Except for 16, all the teachers had produced test materials. Teachers were given a sample task at recruitment, and those who agreed completed the questionnaire in their own time.

# Findings

## 3.1 Teachers' text evaluation

Table 3 presents the descriptive statistics of the questionnaire results, with means out of 5 and standard deviations listed below.

**Table 3:** Descriptive statistics of the questionnaire texts (1 easy – 5 difficult)

| Mean and sd | Text1 Sound of Shell | Text2 Global-ization | Text3 Pastor-alism | Text4 Morality | Text5 Internet | Text6 Coal | Text7 Reading | Text8 Food Prod. | Text9 Foot-ball | Text10 Meteorite Impact |
|---|---|---|---|---|---|---|---|---|---|---|
| q3audience | 1.33 | 2.3 | 2.44 | 3.92 | 2 | 1.78 | 3.49 | 2.51 | 2.27 | 2.44 |
| sd | 0.78 | 1.04 | 1.06 | 0.77 | 1.11 | 0.89 | 1.2 | 1.05 | 1.22 | 1.23 |
| q4grammar | 2.94 | 2.89 | 2.35 | 3.59 | 2.14 | 2.63 | 3.1 | 2.67 | 2.83 | 2.73 |
| sd | 1.24 | 1.51 | 0.83 | 1.04 | 1 | 0.99 | 1.1 | 1 | 1.11 | 1.11 |
| q5vocabulary | 4.1 | 2.84 | 3.46 | 4 | 1.94 | 2.4 | 2.7 | 2.65 | 2.9 | 2.94 |
| sd | 1.03 | 1.12 | 1.03 | 0.95 | 0.86 | 0.83 | 1.04 | 0.92 | 1.17 | 1.09 |
| q6concrete-ness | 2.68 | 2.17 | 2.21 | 4.51 | 1.79 | 1.73 | 3.4 | 2.51 | 2.81 | 2.52 |
| sd | 1.26 | 1.14 | 1.02 | 0.74 | 0.92 | 0.87 | 1.14 | 1.09 | 1.19 | 1.19 |
| q7density | 3.05 | 3.46 | 3.1 | 4.02 | 2.32 | 2.59 | 3.1 | 3.4 | 2.84 | 3.08 |
| sd | 1.29 | 1.04 | 1.06 | 0.94 | 1 | 1.08 | 1.1 | 1.07 | 1.29 | 1.08 |
| q8topicspecif | 2.37 | 2.41 | 2.62 | 3.75 | 1.76 | 1.98 | 2.95 | 2.51 | 2.94 | 2.46 |
| sd | 1.21 | 0.96 | 1.07 | 1.02 | 0.86 | 0.91 | 1.2 | 1.08 | 1.29 | 1.09 |
| q9culture-specif | 3.62 | 1.54 | 2.14 | 2.83 | 1.56 | 2.13 | 1.7 | 1.78 | 3 | 1.83 |
| sd | 1.08 | 0.76 | 1.09 | 1.49 | 0.69 | 1.11 | 1.02 | 0.83 | 1.28 | 1.04 |
| q10sent-cohesion | 2.14 | 2.02 | 1.65 | 2.06 | 1.52 | 1.73 | 2.02 | 3.3 | 2.65 | 2.1 |
| sd | 1.08 | 1.21 | 0.81 | 1.03 | 0.69 | 0.81 | 0.89 | 1.51 | 1.15 | 1 |
| q11coherence | 2.13 | 1.86 | 1.57 | 2.13 | 1.44 | 1.57 | 1.98 | 3.35 | 2.59 | 1.98 |
| sd | 1.05 | 1.12 | 0.76 | 0.99 | 0.67 | 0.71 | 0.99 | 1.56 | 1.16 | 1.01 |
| q12difficulty | 3.62 | 2.67 | 2.9 | 4.1 | 1.71 | 2.3 | 2.92 | 3.48 | 3.19 | 2.83 |
| sd | 1.07 | 1.16 | 1.22 | 1.03 | 0.81 | 0.89 | 1.14 | 1.2 | 1.2 | 1.16 |

The correlations in Table 4 show the links between complexity and suitability judgements and textual features. The strongest correlation with overall text difficulty was with vocabulary (r= .743). Grammar, information density and topic specificity are also strongly correlated with the perception of text difficulty: .658, .646 and .641 respectively. The level of abstractness (.591) and degree of coherence and cohesion within the text are moderately related with perceived text difficulty, with correlations .517 and .513 respectively. It is also

important to note that perceptions of syntactic complexity (grammar) appear to be closely related with vocabulary (.666), information density (.589), topic specificity (.537) and level of abstractness (.515) of the topic in the text. Similarly, vocabulary is correlated with information density (.596), topic specificity (.591), and level of abstractness (.540). The level of abstractness is also correlated with topic specificity (.664) and information density (.548). It is not surprising that information density is perceived as closely related to topic specificity (.608), and that perceptions of cohesion and coherence largely overlap (.890). As the data set is small, we do not report any further statistics here.

**Table 4:** Correlations between questionnaire items

| Questions N (630) | q3 | q4 | q5 | q6 | q7 | q8 | q9 | q10 | q11 |
|---|---|---|---|---|---|---|---|---|---|
| q3audience | 1 | | | | | | | | |
| q4grammar | .453** | 1 | | | | | | | |
| q5vocabulary | .365** | **.666**** | 1 | | | | | | |
| q6concreteness | **.530**** | **.515**** | **.540**** | 1 | | | | | |
| q7density | .469** | **.589**** | **.596**** | **.548**** | 1 | | | | |
| q8topicspec | **.570**** | **.537**** | **.591**** | **.664**** | **.608**** | 1 | | | |
| q9culturespec | .101** | .292** | .474** | .390** | .260** | .417** | 1 | | |
| q10sentence cohesion | .171** | .327** | .238** | .266** | .350** | .302** | .231** | 1 | |
| q11flow (coherence) | .162** | .287** | .218** | .267** | .339** | .307** | .240** | **.890**** | 1 |
| q12overalldifficulty | .409** | **.658**** | **.743**** | **.591**** | **.646**** | **.641**** | .451** | **.513**** | **.517**** |

## 3.2 Genre comparison

Table 5 presents genre/rhetorical purpose classification done by the teachers and ChatGPT-5 through the seven categories in the questionnaire (see q2 in Appendix A) based on Green et al. (2010). More than one category could be chosen to designate the rhetorical purpose of a text. As can be seen, there is a consensus between the teachers and ChatGPT on the rhetorical purpose of the texts. However, ChatGPT tended to see analysis (analyse a process...) in the texts more than the teachers. In Texts 2, 3, 5 and 10, the teachers did not choose that rhetorical purpose. Likewise, ChatGPT saw narrative characteristics in Text 9, when the teachers did not.

**Table 5:** Classification of rhetorical function

| Classification | Text1 Sound of Shell | Text2 Globalisation | Text3 Pastoralism | Text4 Morality | Text5 Internet | Text6 Coal | Text7 Reading | Text8 Food Prod. | Text9 Football | Text10 Meteorite Impact |
|---|---|---|---|---|---|---|---|---|---|---|
| No: Teachers (N63) | 63 Narrate an event | 57 Inform the reader | 45 Inform the reader | 45 Discuss a point.. | 54 Inform the reader | 52 Inform the reader | 40 Analyse a process | 51 Inform the reader | 46 Inform the reader | 50 Inform the read.. |
| | | 14 Discuss causes.. | 36 Describe an obj.. | 41Compare &contrast.. | 18 Discuss a point.. | 17 Analyse a process | 34 Inform the reader | 23 Discuss a point.. | 21 Describe an obj.. | 17 Describe a place |
| | | | | 39 Inform the reader | 14 Compare &contrast.. | 16 Describe an obj.. | 25 Discuss a point.. | 10 Compare &contrast.. | 11 Analyse a process | 10 Discuss a point.. |
| | | | | | | | 12 Compare &contrast.. | | | |
| Chat GPT | Narrate an event | Inform the reader | Inform the reader | Inform the reader | Inform the reader | Inform the reader | Inform the reader | Inform the reader | Narrate an event | Inform the reader |
| | | Analyse a process | Analyse a process | Compare&contrast.. | Analyse a process | Analyse a process | Discuss a point | Discuss a point | Inform the reader | Analyse a process |
| | | Discuss a point | | Discuss a point | | | | | Discuss a point | |

## 3.3 Text-by-text analysis

In this section, we present a text-by-text qualitative review of the congruence between teachers' evaluations and automated analysis. This allows closer inspection of the intricacies of teachers' judgements. For clarity, numerical values are shown in a table for each text (Tables 6–15). To interpret automated index values, we used Social Studies criterion values at the K11–College level (McNamara et al., 2014: 274–278; see Table 1). Discrepancies are defined as differences greater than 1 between teacher and ChatGPT-5 scores. Teachers' and ChatGPT-5's responses to q13 are also included. 'Sample texts' refers to the 30 IELTS and 30 coursebook extracts used in ChatGPT-5 training.

**Table 6:** Text 1 – The Sound of Shell

| Questionnaire | | ChatGPT | | Tool indices | |
|---|---|---|---|---|---|
| q3aud | 1.33 | q3aud | 1 | Narrativity | 61.41 |
| q4gram | 2.94 | q4gram | 2 | SynSim | 64.06 |
| q5vocab | 4.1 | q5vocab | 2 | WordCon | 81.86 |
| q6concrete | 2.68 | q6concrete | 1 | RefCoh | 11.12 |
| q7density | 3.05 | q7density | 1 | Deepcoh | 14.01 |
| q8topicspecif | 2.37 | q8topicspecif | 1 | LexDiv | 77.04 |
| q9cultspecif | 3.62 | q9cultspecif | 2 | Wordfreq | 2.94 |
| q10sentencecoh | 2.14 | q10sentencecoh | 2 | FKRE | 86.29 |
| q11coherence | 2.13 | q11coherence | 2 | FKGL | 4.1 |

| | | | | | |
|---|---|---|---|---|---|
| q12difficulty | 3.62 | q12difficulty | 2 | CohMet | 5.47 |
| q13suitability | 11% | q13suitability | No | Lexile | 850 |
| | | CEFR | B2 | | |

This text was selected as an example of a literary narrative and was among those that produced a discrepancy. The teachers scored almost all the values higher than the tools, indicating complexity, while the tools analysed this as a structurally simple text (except for Coh-metrix readability; 5.47). The teachers found the text to be one of the most difficult overall (3.62), and therefore unsuitable for assessment (only 11 per cent found the text suitable). They commented that the sentence structure, mostly consisting of simple sentences, was not the main problem, but the difficulty would arise from descriptive vocabulary and literary usage with metaphorical references. This is why they scored the text as lexically difficult and abstract. Some commented that slaughtering a pig might not be culturally appropriate, and the text was not suitable for academic purposes. ChatGPT-5 also agreed that despite simple syntax and low lexical sophistication, it is not a suitable text for EAP assessment as it does not resemble the sample academic texts. It did not score the text as culturally specific (2), unlike the teachers (3.62).

**Table 7:** Text 2 – Globalisation

| Questionnaire | | ChatGPT | | Tool indices | |
|---|---|---|---|---|---|
| q3aud | 2.3 | q3aud | 3 | Narrativity | 6.55 |
| q4gram | 2.89 | q4gram | 4 | SynSim | 52.39 |
| q5vocab | 2.84 | q5vocab | 4 | WordCon | 79.67 |
| q6concrete | 2.17 | q6concrete | 4 | RefCoh | 57.93 |
| q7density | 3.46 | q7density | 5 | Deepcoh | 36.69 |
| q8topicspecif | 2.41 | q8topicspecif | 4 | LexDiv | 72.62 |
| q9cultspecif | 1.54 | q9cultspecif | 1 | Wordfreq | 2.78 |
| q10sentencecoh | 2.02 | q10sentencecoh | 2 | FKRE | 29.02 |
| q11coherence | 1.86 | q11coherence | 2 | FKGL | 13.91 |
| q12difficulty | 2.67 | q12difficulty | 4 | CohMet | 13.16 |
| q13suitability | 84% | q13suitability | Yes | Lexile | 1250 |
| | | CEFR | B2 | | |

Text 2 (an IELTS text) was chosen as a benchmark text. The teachers found this text informationally dense (3.46), but otherwise having average qualities. Eighty-four per cent of the teachers deemed this text suitable for assessment. Lexile and Coh-Metrix also seemed to agree in marking the text's readability around FKGL 13. ChatGPT-5 scored this text as more complex in structural aspects (4–5), although it marked the text as B2. The teachers marked the text as of an average difficulty (2.67). Their comments acknowledged the existence of complex sentences and phrases, but these were seen as hallmarks of academic English. The text was seen as loaded with information and the topic suitable for academic readers. Similarly, ChatGPT-5 designated this text as challenging for non-specialists but matching the sample texts in quality.

Text 3 was chosen to represent a text with difficult vocabulary but average grammar difficulty. The teachers' and the tools' evaluation reflected this. However, the readability statistics look quite incongruent: low FKGL (10.31), average Lexile (1100–1200) and a very low Coh-Metrix readability (4.86). Additionally, word frequency was only slightly above the criterion level (2.85). This text was judged to be of average difficulty (2.91) and suitable by 60 per cent of the teachers. The teachers commented that the grammar difficulty was at a suitable level, but the vocabulary was unlikely to be encountered by the designated readers. Despite certain definitions provided in the text, subject-specific vocabulary might pose challenges. They marked concrete and relatively frequent vocabulary as difficult because of the irrelevance of words such as 'herder', 'grazing' and 'stapler' for the intended learner population. ChatGPT-5 evaluated the vocabulary as 'moderate, mostly concrete academic vocabulary'. The teachers who marked the text as suitable commented that it covered a lot of information, quite dense in places, but the information built up in an organised way; therefore, EAP readers should be able to process it.

**Table 8:** Text 3 – Pastoralism

| Questionnaire | | ChatGPT | | Tool indices | |
|---|---|---|---|---|---|
| q3aud | 2.44 | q3aud | 3 | Narrativity | 11.7 |
| q4gram | 2.35 | q4gram | 3 | SynSim | 74.86 |
| q5vocab | 3.46 | q5vocab | 3 | WordCon | 95.05 |
| q6concrete | 2.21 | q6concrete | 3 | RefCoh | 14.69 |
| q7density | 3.1 | q7density | 4 | Deepcoh | 31.92 |
| q8topicspecif | 2.62 | q8topicspecif | 3 | LexDiv | 84.29 |
| q9cultspecif | 2.14 | q9cultspecif | 1 | Wordfreq | 2.85 |
| q10sentencecoh | 1.65 | q10sentencecoh | 2 | FKRE | 46.85 |
| q11coherence | 1.57 | q11coherence | 2 | FKGL | 10.31 |

| | | | | | |
|---|---|---|---|---|---|
| q12difficulty | 2.91 | q12difficulty | 3 | CohMet | 4.86 |
| q13suitability | 60% | q13suitability | Yes | Lexile | 1150 |
| | | CEFR | B2 | | |

**Table 9:** Text 4 – Morality

| Questionnaire | | ChatGPT | | Tool indices | |
|---|---|---|---|---|---|
| q3aud | 3.92 | q3aud | 4 | Narrativity | 39.74 |
| q4gram | 3.59 | q4gram | 4 | SynSim | 46.02 |
| q5vocab | 4 | q5vocab | 4 | WordCon | 6.55 |
| q6concrete | 4.51 | q6concrete | 5 | RefCoh | 73.89 |
| q7density | 4.02 | q7density | 5 | Deepcoh | 88.69 |
| q8topicspecif | 3.75 | q8topicspecif | 4 | LexDiv | 49.75 |
| q9cultspecif | 2.83 | q9cultspecif | 1 | Wordfreq | 3.01 |
| q10sentencecoh | 2.06 | q10sentencecoh | 2 | FKRE | 42.2 |
| q11coherence | 2.13 | q11coherence | 2 | FKGL | 11.87 |
| q12difficulty | 4.1 | q12difficulty | 5 | CohMet | 20.66 |
| q13suitability | 37% | q13suitability | Yes | Lexile | 1150 |
| | | CEFR | B2 | | |

Text 4 was included to represent a syntactically not challenging but conceptually difficult, abstract text with relatively less frequent vocabulary. The indices were able to identify the text's abstract content (word concreteness: 6.55), but the frequency of the words was not classified as low (3.01). Three readability statistics classified this text as average difficulty for pre-university level in agreement: FKRE: 42.20, FKGL: 11.87, Lexile: 1100–1200. Coh-Metrix readability (20.66) was high, denoting a simple text. ChatGPT-5 rated the text as very difficult (5), but nevertheless classified it at the B2 level. Teachers perceived it as complex: highly abstract (4.51), informationally dense (4.02), with difficult vocabulary (4.00), a specific topic (3.75), intended for expert readers (3.92) and even syntactically complex (3.59). ChatGPT-5's evaluations aligned with the teachers' perceptions on all these aspects. The teachers rated the text most informationally dense (4.02). Only 37 per cent considered it suitable for assessment. Teachers noted its highly specific vocabulary – 'normally used by philosophers' – and said that special usages and collocations, rather than complex

structures, made it obscure and ambiguous. Some found the grammar complex, while others felt the vocabulary itself made the grammar difficult, as 'phrasing does not help in understanding the vocabulary.' Those who deemed it suitable for EAP assessment suggested using it for higher-level tasks or students in related fields. ChatGPT-5 described it as very challenging due to abstract reasoning and philosophical terminology. However, it also marked the text at B2 level and as suitable, as it closely resembles higher-level IELTS passages.

**Table 10:** Text 5 – Internet

| Questionnaire | | ChatGPT | | Tool indices | |
|---|---|---|---|---|---|
| q3aud | 2 | q3aud | 3 | Narrativity | 28.77 |
| q4gram | 2.14 | q4gram | 3 | SynSim | 61.41 |
| q5vocab | 1.94 | q5vocab | 3 | WordCon | 11.31 |
| q6concrete | 1.79 | q6concrete | 3 | RefCoh | 4.18 |
| q7density | 2.32 | q7density | 4 | Deepcoh | 89.8 |
| q8topicspecif | 1.76 | q8topicspecif | 3 | LexDiv | 107.64 |
| q9cultspecif | 1.56 | q9cultspecif | 1 | Wordfreq | 3.06 |
| q10sentencecoh | 1.52 | q10sentencecoh | 2 | FKRE | 54.48 |
| q11coherence | 1.44 | q11coherence | 2 | FKGL | 10.4 |
| q12difficulty | 1.71 | q12difficulty | 3 | CohMet | 13.77 |
| q13suitability | 83% | q13suitability | Yes | Lexile | 1150 |
| | | CEFR | B2 | | |

Text 5 was included as a benchmark text, a textbook example, but slightly less demanding than the other two ideal texts (Text 2 and Text 10). The tools did not classify the text as complex (Lexile 1100–1200, Coh-Metrix 13.77), except for a low concreteness index (11.31) and high lexical density (107.64). The teachers agreed with this simplicity, scoring all complexity indices below average, including vocabulary and information density (1.94 and 2.32 respectively). The overall difficulty was the lowest (1.71) among all the texts, and the suitability score was 83 per cent. ChatGPT-5 scored the text as slightly more complex. The teachers generally commented on the straightforwardness of the sentences, easy-access vocabulary and jargon-free nature of the text. For the teachers who saw the text as suitable, it was a well-organised academic text with relevant subject matter at a suitable degree of difficulty. ChatGPT-5's comments were in line with these.

Table 11: Text 6 – Coal

| Questionnaire | | ChatGPT | | Tool indices | |
|---|---|---|---|---|---|
| q3aud | 1.78 | q3aud | 3 | Narrativity | 24.2 |
| q4gram | 2.63 | q4gram | 3 | SynSim | 0.84 |
| q5vocab | 2.4 | q5vocab | 3 | WordCon | 99.16 |
| q6concrete | 1.73 | q6concrete | 3 | RefCoh | 90.66 |
| q7density | 2.59 | q7density | 4 | Deepcoh | 94.06 |
| q8topicspecif | 1.98 | q8topicspecif | 3 | LexDiv | 103.5 |
| q9cultspecif | 2.13 | q9cultspecif | 2 | Wordfreq | 2.87 |
| q10sentencecoh | 1.73 | q10sentencecoh | 2 | FKRE | 38.15 |
| q11coherence | 1.57 | q11coherence | 2 | FKGL | 16.51 |
| q12difficulty | 2.3 | q12difficulty | 3 | CohMet | 9.02 |
| q13suitability | 87% | q13suitability | Yes | Lexile | 1450 |
| | | CEFR | C1 | | |

Text 6 was chosen to represent a text with complex grammar but average vocabulary difficulty. The automated tool indices showed very low syntactic simplicity (0.84) and low readability (Coh-Metrix: 9, FKRE: 38.15, FKGL: 16.5, Lexile: 1400–1500), and high lexical density (103.5), reflecting high complexity congruently. Despite being marked as difficult by the tools, teachers did not rate this text as such (overall difficulty: 2.3; grammar: 2.63), and ChatGPT-5's scores aligned with theirs. Most (87%) deemed it suitable for assessment. They noted complex sentence structures, but felt that EAP students at the target level could manage them, as comprehension was not hindered. Vocabulary was accessible and not overly subject-specific, and the text was conceptually straightforward, with information clearly organised. Teachers acknowledged its academic, informative style and sophisticated grammar, but said accessible vocabulary supported suitability. ChatGPT-5 likened it to IELTS and university coursebook passages in style and density.

Table 12: Text 7 – Reading

| Questionnaire | | ChatGPT | | Tool indices | |
|---|---|---|---|---|---|
| q3aud | 3.49 | q3aud | 4 | Narrativity | 65.54 |
| q4gram | 3.1 | q4gram | 4 | SynSim | 19.77 |
| q5vocab | 2.7 | q5vocab | 4 | WordCon | 16.11 |
| q6concrete | 3.4 | q6concrete | 4 | RefCoh | 53.98 |
| q7density | 3.1 | q7density | 4 | Deepcoh | 91.62 |
| q8topicspecif | 3.95 | q8topicspecif | 4 | LexDiv | 80.77 |
| q9cultspecif | 1.7 | q9cultspecif | 1 | Wordfreq | 2.95 |
| q10sentencecoh | 2.02 | q10sentencecoh | 2 | FKRE | 64.2 |
| q11coherence | 1.98 | q11coherence | 2 | FKGL | 9.52 |
| q12difficulty | 2.92 | q12difficulty | 4 | CohMet | 15.76 |
| q13suitability | 56% | q13suitability | Yes | Lexile | 1150 |
| | | CEFR | B2+ | | |

Text 7 represented a highly topic-specific but structurally simple academic text. Tools rated syntactic simplicity (19.77) and word concreteness (16.11) as low; Lexile (1100–1200) and Coh-Metrix (15.76) placed it at grade 11–12, while FKRE (64.2) and FKGL (9.5) labelled it relatively simple. Teachers identified it as specialist-oriented (3.49) and highly topic-specific (3.95), with ChatGPT-5 in agreement. Overall difficulty was moderate (2.92), yet only 56 per cent deemed it suitable for assessment. Most teachers saw it as highly academic but accessible, suitable for intensive reading; they saw syntactic complexity, word abstractness and content as appropriate to the target group. The abstractness of vocabulary did not matter, as words such as 'process', 'translation' and 'argument' should be familiar to EAP learners. Those opposed cited subject specificity requiring background knowledge. ChatGPT-5 noted its similarity to sample texts.

Text 8 was edited to disrupt coherence to test its impact on automated tools and teacher perceptions. Readability indices placed it on the difficult side (FKRE: 33.75; FKGL: 13; Coh-Metrix: 8.24; Lexile: 1200–1300). Vocabulary frequency was low (2.74) and lexical diversity high (177.49); Coh-Metrix also showed high deep cohesion (77.04). Teachers rated information density (3.4), lack of sentence cohesion (3.6) and lack of coherence (3.65) high, judging overall difficulty as high (3.48). Only 23 per cent of the teachers found it suitable for assessment, citing 'jumpiness' and sentences not following each other logically. ChatGPT-5, however, did not detect incoherence (2), found clear logical sequencing and likened it to sample texts.

Table 13: Text 8 – Food

| Questionnaire | | ChatGPT | | Tool indices | |
|---|---|---|---|---|---|
| q3aud | 2.51 | q3aud | 3 | Narrativity | 6.81 |
| q4gram | 2.67 | q4gram | 3 | SynSim | 57.53 |
| q5vocab | 2.65 | q5vocab | 3 | WordCon | 86.86 |
| q6concrete | 2.51 | q6concrete | 3 | RefCoh | 4.27 |
| q7density | 3.4 | q7density | 4 | Deepcoh | 77.04 |
| q8topicspecif | 2.51 | q8topicspecif | 3 | LexDiv | 177.49 |
| q9cultspecif | 1.78 | q9cultspecif | 2 | Wordfreq | 2.74 |
| q10sentencecoh | 3.6 | q10sentencecoh | 2 | FKRE | 33.75 |
| q11coherence | 3.65 | q11coherence | 2 | FKGL | 13.27 |
| q12difficulty | 3.48 | q12difficulty | 3 | CohMet | 8.24 |
| q13suitability | 23% | q13suitability | Yes | Lexile | 1250 |
| | | CEFR | B2+ | | |

Table 14: Text 9 – Football

| Questionnaire | | ChatGPT | | Tool indices | |
|---|---|---|---|---|---|
| q3aud | 2.27 | q3aud | 3 | Narrativity | 37.83 |
| q4gram | 2.83 | q4gram | 3 | SynSim | 9.18 |
| q5vocab | 2.9 | q5vocab | 3 | WordCon | 43.64 |
| q6concrete | 2.81 | q6concrete | 3 | RefCoh | 21.48 |
| q7density | 2.84 | q7density | 3 | Deepcoh | 50.8 |
| q8topicspecif | 2.94 | q8topicspecif | 3 | LexDiv | 113.06 |
| q9cultspecif | 3 | q9cultspecif | 3 | Wordfreq | 3.04 |
| q10sentencecoh | 2.65 | q10sentencecoh | 2 | FKRE | 50.25 |
| q11coherence | 2.59 | q11coherence | 2 | FKGL | 12.05 |
| q12difficulty | 3.19 | q12difficulty | 3 | CohMet | 13.12 |
| q13suitability | 17% | q13suitability | Yes | Lexile | 1350 |
| | | CEFR | B2+/C1 | | |

Text 9 was included in the study to represent a culturally specific text. It had low syntactic simplicity (9.18) and high Lexile (1300–1400) scores, though other measures did not indicate notable pre-university complexity (Coh-Metrix: 13.12; FKRE: 50; FKGL: 12). Teachers scored most features as average, with the highest rating for overall difficulty (3.19). Only 17 per cent considered it suitable for assessment. They found the grammar complex but accessible, yet noted heavy jargon and colloquial usages. Those rejecting it cited the need for football-specific knowledge, terminology and journalese. The frequent cultural and context-specific references in this text were seen as impediments to comprehension. ChatGPT-5 judged it similar to IELTS-style articles on famous individuals, integrating analytical discussion like sport-science texts.

**Table 15:** Text 10 – Meteorite Impact

| Questionnaire | | ChatGPT | | Tool indices | |
|---|---|---|---|---|---|
| q3aud | 2.44 | q3aud | 4 | Narrativity | 12.92 |
| q4gram | 2.73 | q4gram | 4 | SynSim | 50 |
| q5vocab | 2.94 | q5vocab | 4 | WordCon | 1.83 |
| q6concrete | 2.52 | q6concrete | 4 | RefCoh | 44.83 |
| q7density | 3.08 | q7density | 4 | Deepcoh | 50.8 |
| q8topicspecif | 2.46 | q8topicspecif | 4 | LexDiv | 86.3 |
| q9cultspecif | 1.83 | q9cultspecif | 1 | Wordfreq | 2.78 |
| q10sentencecoh | 2.1 | q10sentencecoh | 2 | FKRE | 37.64 |
| q11coherence | 1.98 | q11coherence | 1 | FKGL | 12.73 |
| q12difficulty | 2.83 | q12difficulty | 4 | CohMet | 11.98 |
| q13suitability | 76% | q13suitability | Yes | Lexile | 1250 |
| | | CEFR | B2+/C1 | | |

Text 10 was chosen as a benchmark. Tools flagged only very low word concreteness (1.83); other features were average, as expected. Readability indices aligned on complexity (Lexile: 1200–1300; Coh-Metrix: 13.12; FKGL: 12), though ChatGPT-5's scores were mostly higher (around 4). Teachers rated most features as average (vocabulary: 2.94; grammar: 2.73), except for slightly above-average information density (3.08). Seventy-six per cent considered it suitable for assessment, noting frequent higher-level vocabulary and subject-specific concepts but no grammatical difficulty. They felt the topic and academic style suited EAP assessment. ChatGPT-5 said it demands strong reading skills and closely matches sample text features.

## 3.4 Summary of the results

As our data is based on ten cases of texts with differing features, the findings cannot be generalised. Nonetheless, the following observations can be made.

1. Correlation data indicated that vocabulary correlated most strongly with overall text difficulty perceptions, followed by grammar, information density and topic specificity.
2. Rhetorical purpose of texts can be identified accurately by the automated text analysis tools. However, narrative features in non-narrative texts can be misleading (Text 7, Text 9). Text analysis tools may not judge the complexity of narrative texts accurately due to literary usage and cultural specificity (Text 1).
3. In the analysis of benchmark texts – those without skewed features such as high abstractness or content specificity – the results from automated tools, including ChatGPT, aligned with the teachers' evaluations. Such automated methods may work well with texts that reflect the typical features of EAP texts (Text 2, Text 5).
4. In general, ChatGPT-5's text evaluations aligned with those of the teachers; however, for judgements of subject specificity, appropriate abstractness, incoherence and cultural specificity, ChatGPT-5 may be less effective than teachers' evaluations (Text 1, Text 3, Text 4, Text 7, Text 8, Text 9).
5. Automated readability indices were roughly congruent with each other except for one case of discrepancy (Text 6).
6. Readability statistics may not reflect the difficulty brought about by subject-specificity (Text 3, Text 4, Text 7).
7. High word frequency, as measured by automated tools, does not necessarily indicate word concreteness; high-frequency words may be used in abstract senses (Text 4). Texts with simple syntax but high-frequency words used abstractly can create a high conceptual load, yet still appear to have high readability in automated analyses. ChatGPT-5 may be more accurate in detecting abstractness.
8. For EAP teachers, vocabulary complexity overrides grammatical complexity in importance. While subject-specific vocabulary is seen as a source of complexity, complex syntax may not be seen as such when the vocabulary in the text is accessible (Text 3, Text 6). Automated text analysis tools may be inefficient in reflecting this.
9. ChatGPT-5 may not designate the CEFR level of the texts accurately, often overgeneralising to B2 level (Text 2, Text 3, Text 4).

# Discussion

In this study, we investigated the textual features that shape EAP teachers' perceptions of text difficulty and compared their evaluations with automated text analysis in a qualitative case-based manner. The analyses highlighted areas requiring more attention for developing a robust system of text complexity analysis.

Vocabulary emerged as the strongest variable explaining teachers' evaluations. This aligns with previous research that has established vocabulary as the primary factor correlated with the judgements on syntactic complexity (Hamada, 2015). At the pre-university level, learners are generally able to process a wide range of syntactic features; thus, unless texts pose unusual syntactic difficulty, text complexity often resides in vocabulary, especially in EAP contexts where field-specific lexis is central. For example, Text 6 (Coal) displayed relatively high syntactic complexity but was not judged as syntactically complex because its vocabulary was concrete and accessible (e.g. soap, destination). By contrast, Text 4 (Morality) was not syntactically complex, contained frequent words but used them in abstract senses. It was judged as syntactically and lexically difficult. This was a case of frequent words being loaded with conceptual complexity depending on the context, for example 'reason' meaning 'cause', a frequent word, and 'good judgement', relatively rare usage reflecting a complex concept. In contrast, Text 5 (Internet) contained abstract, low-frequency words that were nonetheless accessible to learners. These examples are also indicative of the limitations of automatic tools to distinguish between relative difficulty of lexical items and relevance to the learners.

These findings resonate with long-standing views on the primacy of vocabulary in L2 (Barcroft, 2004). Healy and Sherrod (1994, in Barcroft, 2004) underline that 'grammar knowledge actually resides at the lexical level in connections between words and groups of words developed over time' (p. 201). Such perspectives align with lexical, usage-based explanations of language acquisition (i.e. Tomasello, 2015). As Khalifa and Weir (2009) underline, syntactic processing is affected by the difficulty or ease with which the lexical items can be accessed, and this intricate interplay has to be accounted for.

Our study also demonstrated that lexical complexity cannot be reduced to frequency or abstractness. Instead, judgements of lexical complexity are shaped by the synthesis and relative weight of frequency, abstractness, relevance to learners and subject familiarity from the perceptions of teachers. This highlights the need for a more sophisticated treatment of lexical complexity in L2 text analysis. For example, subject-specific EAP corpora based on first-year university coursebooks with analysis of lexical combinations (usage-based categories) should be available for automatic analysis. Within such corpora, lexical primacy could be identified by relevance to academic domains. NLP methods used in user interest detection (for example TF–IDF in Xia, 2024) could be adapted to calculate lexical relevance.

Teachers' judgements also reflected interaction among vocabulary, syntax, content relevance, familiarity and stylistic features. As atomistic calculations may not capture such complexity, a layered, hierarchical model of EAP text complexity weighing content relevance and familiarity over grammar would be more accurate. Under-analysis of the complexity of human judgement may also explain why previous studies identified factors influencing textual difficulty but failed to produce conclusive, congruent results (see also Imperial & Madabushi, 2022).

Automated tools often operationalise coherence as word overlap across sentences, but this may not always capture whether ideas are logically connected to form a unified message. Deeper measures of discourse connectedness can be achieved through discourse analysis, such as Rhetorical Structure Theory (Sun & Xiong, 2019). Automated measures also tend to underestimate the difficulty of narrative texts, likely due to their stylistic and cultural features. Narratives with literary style and culturally embedded content may demonstrate sophistication beyond syntactic and lexical complexity (Nelson et al., 2012; Sheehan et al., 2010).

In our analysis, ChatGPT-5 exhibited congruence with teacher judgements in general. However, it was not as efficient in evaluating subject specificity, appropriate abstractness, incoherence and cultural specificity. While ChatGPT-5 generally aligned with teacher evaluations for benchmark texts and standard academic features, it showed limitations in detecting certain qualitative aspects, especially those dependent on human awareness of context, reader background and educational appropriateness. Following Bezirhan and von Davier (2023) and Kogan et al. (2025), we argue that cognitive models trained on appropriate data are necessary for LLMs to address text difficulty more effectively. For ChatGPT-5, we created a training corpus of domain-specific texts (university coursebook texts) and expert-aligned IELTS texts (Green et al., 2010) using CEFR as expert-defined standard (Imperial et al., 2024). Despite this, we observed inconsistencies both within and across runs of ChatGPT-5, with a tendency to overgeneralise to B2 classifications, a finding also reported by Benedetto et al. (2025). This raises questions about computations and constraints underlying ChatGPT-5's assignments and highlights the current limits of model controllability.

# Conclusion

Selecting suitable texts for specific learner groups requires understanding both readers and texts and judging whether chosen analytical methods are appropriate (Benjamin, 2012). Although automated text complexity analysis has provided valuable support for L2 reading assessment, it cannot yet match the sophistication of human evaluation, which considers learners' educational background, the level of conceptual and cognitive challenge they can manage, and the subject or cultural specificity of texts in context. Automated analysis could benefit greatly from user-centric profiling of CEFR-aligned texts tagged for educational and cultural contexts, language-use domains and age-related subcategories. Hierarchical models of text complexity that incorporate these dimensions alongside lexical-grammatical interactions and discoursal features may offer more valid and context-sensitive analysis. While our suggestions are based on a limited, small-scale qualitative comparison of human judgement and automated text analysis, they are intended to highlight areas for future development. Future research can build on this work by combining larger datasets with fine-grained human evaluations to develop more robust and pedagogically sound models of text complexity. Presently, these tools should be used with awareness of their shortcomings and applied in a critical manner.

# References


Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence, 5*, 903077. https://doi.org/10.3389/frai.2022.903077

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford University Press.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. C. (1995). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study*. Cambridge University Press.

Balyan, R., McCarthy, K. S., & McNamara, D. S. (2020). Applying natural language processing and hierarchical machine learning approaches to text difficulty classification. *International Journal of Artificial Intelligence in Education, 30*(3), 337–370. https://doi.org/10.1007/s40593-020-00201-7

Barcroft, J. (2004). Second language vocabulary acquisition: A lexical input processing approach. *Foreign Language Annals, 37*, 200–208. https://doi.org/10.1111/j.1944-9720.2004.tb02193.x

Benedetto, L., Gaudeau, G., Caines, A., & Buttery, P. (2025). Assessing how accurately large language models encode and apply the Common European Framework of Reference for Languages. *Computers and Education: Artificial Intelligence, 8*, 100353.

Benjamin, R. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24*(1), 63–88. https://doi.org/10.1007/s10648-011-9181-8

Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence, 5*, 100161. https://doi.org/10.1016/j.caeai.2023.100161

Choi, I. C., & Moon, Y. (2019). Predicting the difficulty of EFL tests based on corpus linguistic features and expert judgment. *Language Assessment Quarterly, 17*(1), 18–42. https://doi.org/10.1080/15434303.2019.1674315

Common Core State Standards: The Standards' approach to text complexity https://www.isbe.net/Documents/5-determining-text-complexity.pdf

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes, 54*(5–6), 340–359. https://doi.org/10.1080/0163853X.2017.1296264

Dowell, M. M. N., Graesser, A. C., & Cai, Z. (2016). Language and discourse analysis with Coh-Metrix: Applications from material to learning environments at scale. *Journal of Learning Analytics, 3*(3), 72–95. https://doi.org/10.18608/jla.2016.33.5

Fitzgerald, J., Elmore, J., Hiebert, E. H., Koons, H. H., Bowen, K., Sanford-Moore, E. E., & Stenner, A. J. (2016). Examining text complexity in the early grades. *Phi Delta Kappan*, *97*(8), 60–65. https://doi.org/10.1177/0031721716647023

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

Green, A., Ünaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing, 27*(2), 191–211. https://doi.org/10.1177/0265532209349471

Hamada, A. (2015). Linguistic variables determining the difficulty of Eiken reading passages. *JLTA Journal, 18*, 57–77. https://doi.org/10.20622/jltajournal.18.0_57

Imperial, J. M., & Madabushi, H. T. (2022, April 11). *Uniform complexity for text generation*. arXiv. https://doi.org/10.48550/arXiv.2204.05185

Imperial, J. M., Forey, G., & Madabushi, H. T. (2024, February 19). *Standardize: Aligning language models with expert-defined standards for content generation*. arXiv. https://doi.org/10.48550/arXiv.2402.12593

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.

Kogan, D., Schumacher, M., & Nguyen, S. (2025, June 16). *Ace-CEFR: A dataset for automated evaluation of the linguistic difficulty of conversational texts for LLM applications*. arXiv. https://doi.org/10.48550/arXiv.2506.14046

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Doctoral dissertation, University of Georgia]. https://doi.org/10.57709/8501051

Kyle, K., & Crossley, S. A. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing, 34*(4), 513–535. https://doi.org/10.1177/0265532217712554

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

MetaMetrics. (2018). *Lexile framework for reading*. https://lexile.com/educators/tools-to-support-reading-at-school/tools-to-determine-a-books-complexity/the-lexile-analyzer/

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance.* Student Achievement Partners. https://achievethecore.org/page/1196/measures-of-text-difficulty-testing-their-predictive-value-for-grade-levels-and-student-performance

OpenAI. (2025, August 7). *Introducing GPT-5.* https://openai.com/index/introducing-gpt-5/

Sheehan, K. M., Kostin, I., Futagi, Y., & Flor, M. (2010). *Generating automated text complexity classifications that are aligned with targeted text complexity standards.* Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02235.x

Stenner, A. J., Sanford-Moore, E. E., & Burdick, D. S. (2007). *The Lexile Framework for Reading: Technical report*. Metametrics, Inc. https://metametricsinc.com/research-publications/lexile-technical-report/

Sun, K., & Xiong, W. (2019). A computational model for measuring discourse complexity. *Discourse Studies, 21*(6), 690–712. https://doi.org/10.1177/1461445619866985

Tomasello, M. (2015). The usage-based theory of language acquisition. In E. L. Bavin & L. R. Naigle (Eds.), *The Cambridge handbook of child language* (2nd ed., pp. 89–106). Cambridge University Press. https://doi.org/10.1017/CBO9781316095829

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave. https://doi.org/10.1057/9780230514577

Williamson, G., Sandvik, T., Stenner, J., & Johnson, A. (2016). *Complexity of university texts in the United Kingdom*. MetaMetrics Inc. https://metametricsinc.com/research-publications/complexity-university-texts-united-kingdom/?full_article=true

Xia, K. (2024). *Personalized recommendation for network teaching courses based on combined filtering of deep learning and k-means*. In Z. Hou (Ed.), *Intelligent computing technology and automation* (pp. 1159–1167). IOS Press. https://doi.org/10.3233/ATDE231299

Xia, M., Kochmar, E., & Briscoe, T. (2019). *Text readability assessment for second language learners*. arXiv. https://doi.org/10.48550/arXiv.1906.07580

# Appendices

## Appendix A – Text Analysis Questionnaire

**TEXT 1:** THE SOUND OF THE SHELL

They found a piglet caught in a curtain of creepers, throwing itself at the elastic traces in all the madness of extreme terror. Its voice was thin, needle-sharp and insistent. The three boys rushed forward and Jack drew his knife again with a flourish. He raised his arm in the air. There came a pause, a hiatus, the pig continued to scream and the creepers to jerk, and the blade continued to flash at the end of a bony arm. The pause was only long enough for them to understand what an enormity the downward stroke would be. Then the piglet tore loose from the creepers and scurried into the undergrowth. They were left looking at each other and the place of terror. Jack's face was white under the freckles. He noticed that he still held the knife aloft and brought his arm down replacing the blade in the sheath. Then they all three laughed ashamedly and began to climb back to the track.

"I was choosing a place," said Jack. "I was just waiting for a moment to decide where to stab him."

"You should stick a pig," said Ralph fiercely. "They always talk about sticking a pig."

"You cut a pig's throat to let the blood out," said Jack, "otherwise you can't eat the meat."

1. Which of the following does this text do? (you can choose more than one.)

   (1) narrate an event, (2) describe an object, place etc., (3) inform the reader on a point, (4) compare and contrast things or phenomena, (5) analyze a process, (6) discuss a point from different perspectives (7) defend a point

2. This extract is probably taken from … (you can choose more than one.)

   (1) newspaper article, (2) magazine article, (3) research article, (4) textbook chapter, (5) book chapter, (6) novel/story, (7) other: _____.

3. This text is written for …

   (general audiences) 1 2 3 4 5 (experts)

4. For typical Freshmen students, the grammar of this text is …
   (consider passives, compound/complex sentences and phrases etc.)

   (easy) 1 2 3 4 5 (difficult)

   Comments:

   _____

5. For typical Freshmen students, the vocabulary of this text is ...

   (basic/frequent) 1 2 3 4 5 (difficult)

   Comments: _____

6. The concepts discussed in this text are ...

   (concrete) 1 2 3 4 5 (abstract)

   Comments: _____

7. If a text presents a lot of information in a short space, we call it an informationally dense text. The information in this text is ...

   (not dense) 1 2 3 4 5 (very dense)

   Comments: _____

8. The reading of this text requires _____ amount of topic specific knowledge.

   (a minimum) 1 2 3 4 5 (a very high)

   Comments: _____

9. If a text can be understood by readers from different cultural backgrounds, we call it a culture-free text. The topic of the text is ...

   (culture-free) 1 2 3 4 5 (culture-specific)

   Comments: _____

10. Sentences in the text are connected to each other ...

    (very clearly) 1 2 3 4 5 (not clearly)

    Comments: _____

11. The flow of the ideas in the text is ...

    (clear) 1 2 3 4 5 (not clear)

    Comments: _____

12. A student at the beginning of the first year will read this text ...

    (easily) 1 2 3 4 5 (with difficulty)

    Comments: _____

13. Do you think this is a suitable text to be used in an exam at the end of the prep. year? Please, circle Yes or No. (Disregard the length)

    YES NO

Explain your decision briefly:

**TEXT 2:** GLOBALIZATION

...Globalization has given rise to conditions that have facilitated the emergence and spread of diseases. Constant urbanization and population growth, especially in developing countries, have increased population density, allowing communicable diseases to spread more easily. Global trade and travel have opened new routes for the spread of diseases. However, the importance of tourism revenue has prevented countries from reporting epidemics, allowing diseases to spread even further. Development and its destruction of native habitats have introduced diseases that were previously isolated in nature. Food borne illness outbreaks have increased as a result of the rise of global trade in the late 20th century. The increasing frequency of natural disasters related to climate change can lead to a higher incidence of the disease outbreaks that often follow. Such outbreaks and potential pandemics may also result in widespread public fear and panic. The possibility of global pandemics resulting in millions of deaths and severe negative impacts on economy has led to the rise of national and international emergency response planning and the use of new technology to create early detection and warning systems. Problems also include the tendency of some nations to cover up disease outbreaks. This prevents research and implementation of measures designed to quickly halt the spread of diseases.

**TEXT 3:** PASTORALISM

...Pastoralism as a way of life involves the herding of sheep, goats, and cattle. It emerged around 5500 BCE, essentially at the same time that full-time farmers appeared. The first pastoralists were closely affiliated with the inhabitants of agricultural villages growing wheat and barley, which required large parcels of land. Pastoralists produced meat and dairy products, as well as wool for textiles. Additionally, they bartered these products with the agriculturalists for grain, pottery, and other staples. In the fertile crescent surrounding the Mesopotamian alluvium, many extended families farmed and herded at the same time. They were cultivating crops on large estates and grazing their herds in the foothills and mountains nearby. These herders moved their livestock seasonally. They usually pastured their flocks in higher lands during summer and in valleys in winter. This movement over short distances is called transhumance and did not require herders to vacate their primary locations in the mountain valleys. Nomadic pastoralism is another form of pastoralism. It is based on the herding of cattle and other livestock. It flourished in various settings, most notably in the steppe lands north of the agricultural zone of southern Eurasia. This way of life was characterized by horse-riding herders of livestock.

**TEXT 4:** MORALITY

...The rationalists and empiricists carried their debate into the area of moral knowledge. The rationalists claimed that our knowledge of moral principles is a type of metaphysical knowledge, implanted in us by God, and discoverable by reason as it deduces general principles about human nature. On the other hand, the Scottish empiricists, especially David Hume and Adam Smith, argued that morality is founded entirely on the contingencies of human nature. They claim that morality is based on desire. Morality concerns making people happy, fulfilling their reflected desires, and reason is just a practical means of helping them fulfil their desires. There is nothing of special importance in reason in its own right. It is mainly a rationalizer and servant of the passions. As Hume said, "Reason is and ought only to be a slave of the passions and can never pretend to any other office than to serve and obey them." Morality is founded on our feeling of sympathy with other people's sufferings, on fellow feeling. For such empiricists then, morality is contingent upon human nature. If we had a different nature, then we would have different feelings and desires, and so we would have different moral principles.

**TEXT 5:** THE INTERNET

...Over the past decade or so, research within the social sciences has come to use the Internet more and more. Three uses will be briefly outlined here. The first can be found in using the Internet to gain relatively straightforward access to data on all manner of worldwide issues. In many ways using a search engine is a good starting point for almost any social research, and sometimes it may prove to be all you need as data on the Web is like secondary data that is open to analysis. A second use can be to deploy research tools on the Internet. The most obvious example here is email interviewing. Having found a sample or special subject, one can ask questions by email and the respondent replies, which can lead to further and fuller questioning. A third approach is to investigate the nature of online life itself. Increasingly, we spend more of our time 'living online', so it becomes of sociological interest to see how people use the Internet. However, as researchers come to use websites more and more for their basic materials, they can come up with a huge amount of websites that are unreliable and even useless for the purposes of accurate information. Anybody can make a website after all, and what is to stop people putting misinformation on the site either deliberately or out of ignorance.

**TEXT 6:** COAL

...England's forests were never fully restored; however, fuel shortages were reduced to some extent by burning coal in the place of wood. Despite people's worries about the harmful gases given off by burning coal, it came to be widely used for domestic heating, and as a source of heat for the production of sugar, bricks, soap, glass, and iron. More than simply being a substitute for wood, by the end of the nineteenth century coal had become the

basis of industrial civilization, as the rich coal deposits of Britain significantly contributed to that country's unique position as "the Workshop of the World." Much of the industrial age was the era of coal, as coal fired steam engines powered factories, pulled railroad trains, generated electricity, and pushed ships to distant destinations. Yet, just when coal had established its primacy as the most important energy source for industrial society, hard questions were being asked about the continued practicality of technologies based on coal. By the end of the nineteenth century it was becoming clear that stocks of coal, while still large, were being reduced at ever increasing rates, and the projection of established trends seemed to offer clear proof that Britain was running out of coal.

### TEXT 7: THE NATURE OF READING

...If theorists are not yet agreed on what skills are involved in the reading process, is it at least possible to find some consensus on what happens when we read? What kinds of tasks characterize the activity involved in reading? Clearly, reading involves perceiving the written form of language visually. Here we already encounter the first problem: do readers then relate the printed form of language to the spoken form? If so, then once that translation has taken place, reading is the same sort of activity as listening. And the only specific aspect of reading that we need to concern ourselves with as testers is the process of transformation from print to speech. One argument, put forward by theorists like Smith (1971), is that readers proceed directly to meaning, and do not go via sound. They claim that readers can process print much faster than sounds, and so there would be an upper limit on the speed with which we read if we had to go from print to sound. Fluent reading is done at speeds up to three times as fast as many people speak in everyday conversation.

### TEXT 8: FOOD PRODUCTION

... Two things distinguish food production from all other productive activities. First, every single person needs food for each day and has a right to it; and second, it is hugely dependent on nature. Four unique aspects, one political, another natural make food production highly vulnerable and similar to all other businesses. At the same time, cultural values are highly fixed in food and agricultural systems worldwide. Farmers everywhere have major advantages, including weather, long-term climate change, and price instability in input and product markets. However, smallholder farmers in developing countries must in addition deal with difficult environments, both natural in terms of soil quality, rainfall, etc., and human in terms of infrastructure, financial systems, markets, knowledge and technology. Participants in the online debate argued that the biggest challenge is to address the underlying causes of the agricultural system's ability to ensure sufficient food for all. And they identified our dependency on fossil fuels and encouraging government policies as the main reasons of this problem. In order to document the risks farmers face,

most experts call for greater state intervention. They argue that governments can enhance job prospects by providing basic services like roads to get produce to the markets, or water and food storage facilities to reduce losses.

### TEXT 9: FOOTBALL

... For AS Roma, the key to shaping its future is not forgetting its past. That past now includes Francesco Totti, who concluded his remarkable 24-year career with Roma on May 28, in a home match at the Stadio Olimpico against Genoa. The 40-year-old striker provides an interesting case study of football longevity for the club's director of performance Darcy Norman. Norman is interested in using a supply chain management and systems thinking approach, borrowed from the world of big business and applied to European football. It's an approach based on the idea that knowing that every action sets off a chain of events that will impact performance. As for Totti, Norman cites a "complex system" that includes "good genetics" and balanced lifestyle that allowed the Roma attacker to make effective appearances at his age. Totti is very much in tune with his "performance mind set," says Norman. "His ability to read the game, and be at the right place at the right time can compensate for the fact that he may not be as explosive as before," he adds. Darcy also notes that there are "definitely things to learn" from Totti, along with the careers of Roma midfielder Daniele De Rossi, and Juventus' 39-year-old goalkeeper Gianluigi Buffon.

### TEXT 10: METEORITE IMPACTS

... Impacts by meteorites represent one mechanism that could cause global catastrophes and seriously influence the evolution of life all over the planet. According to some estimates, the majority of all extinctions of species may be due to such impacts. Such a perspective fundamentally changes our view of biological evolution. The standard criterion for the survival of a species is its success in competing with other species and adapting to slowly changing environments. Yet an equally important criterion is the ability of a species to survive random global ecological catastrophes due to impacts.

Earth is a target in a cosmic shooting gallery, subject to random violent events that were unsuspected a few decades ago. In 1991 the United States Congress asked NASA to investigate the hazard posed today by large impacts on Earth. The group conducting the study concluded from a detailed analysis that impacts from meteorites can indeed be hazardous. Although there is always some risk that a large impact could occur, careful study shows that this risk is quite small.

## Appendix B – ChatGPT-5 Results

| | Text1 | Text1 | Text1 | Text2 | Text2 | Text2 | Text3 | Text3 | Text3 | Text4 | Text4 | Text4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| q1genre | inform reader analyse a process | narrate an event | narrate an event | inform reader compare & contrast discuss a point | textbook chapter book chapter research article | inform reader analyse a process discuss a point | inform reader discuss a point | inform reader analyse a process describe an object/place | inform reader analyse a process | inform reader analyse a process | inform reader compare & contrast discuss a point | inform reader compare & contrast discuss a point |
| q2source | textbook chapter | novel/ story | novel/ story | magazine article | magazine article text book chapter | textbook chapter magazine/ research based article | research article | text book chapter book chapter | textbook chapter book chapter | textbook chapter | text book chapter book chapter | text book chapter book chapter |
| q3audience | 4 | 1 | 1 | 3 | 4 | 3 | 5 | 3 | 3 | 4 | 5 | 4 |
| q4grammar | 4 | 2 | 2 | 3 | 4 | 4 | 4 | 3 | 3 | 3 | 5 | 4 |
| q5vocabulary | 4 | 2 | 2 | 3 | 4 | 4 | 5 | 3 | 3 | 4 | 5 | 4 |
| q6concretenes | 4 | 1 | 1 | 3 | 4 | 4 | 5 | 3 | 3 | 4 | 5 | 5 |
| q7density | 4 | 1 | 1 | 3 | 5 | 5 | 5 | 4 | 4 | 4 | 5 | 5 |
| q8topicspecif | 4 | 1 | 1 | 3 | 4 | 4 | 5 | 3 | 3 | 4 | 5 | 4 |
| q9culturespecif | 2 | 3 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 1 |
| q10sentcohesi | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 |
| q11coherence | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 |
| q12diffculty | 2 | 1 | 2 | 2 | 4 | 4 | 4 | 3 | 3 | 3 | 5 | 5 |
| q13 suitability | Yes | No | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| CEFR | C1 | C1 | B2 | B2+ | C1 | B2 | C1 | B2+ | B2 | B2+ | C1 | B2 |
| | **Text5** | **Text5** | **Text5** | **Text6** | **Text6** | **Text6** | **Text7** | **Text7** | **Text7** | **Text8** | **Text8** | **Text8** |
| q1genre | inform reader defend a point | inform reader analyse a process discuss a point | inform reader analyse a process | inform reader analyse a process discuss a point | inform reader analyse a process | inform reader analyse a process | inform reader analyse a process | inform reader discuss a point | inform reader discuss a point | inform reader analyse a process | inform reader discuss a point | inform reader discuss a point |
| q2source | newspaper article | textbook chapter magazine article | textbook chapter book chapter | book chapter | textbook chapter book chapter | textbook chapter book chapter | textbook chapter | textbook chapter research article | textbook chapter book chapter | textbook chapter | textbook chapter magazine article | textbook chapter magazine article |
| q3audience | 3 | 3 | 3 | 4 | 3 | 3 | 3 | 4 | 4 | 4 | 3 | 3 |
| q4grammar | 3 | 3 | 3 | 4 | 4 | 3 | 3 | 4 | 4 | 4 | 3 | 3 |
| q5vocabulary | 3 | 3 | 3 | 4 | 4 | 3 | 3 | 4 | 4 | 4 | 3 | 3 |
| q6concretenes | 3 | 3 | 3 | 4 | 3 | 3 | 3 | 4 | 4 | 4 | 3 | 3 |
| q7density | 3 | 4 | 4 | 4 | 4 | 4 | 3 | 4 | 4 | 4 | 4 | 4 |
| q8topicspecif | 3 | 3 | 3 | 4 | 3 | 3 | 3 | 4 | 4 | 4 | 3 | 3 |
| q9culturespecif | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 |
| q10sentcohesi | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 |
| q11coherence | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 |
| q12diffculty | 2 | 3 | 3 | 3 | 4 | 3 | 2 | 4 | 4 | 3 | 3 | 3 |
| q13 suitability | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| CEFR | B2 | B2+ | B2 | C1 | C1 | C1 | B2+ | C1 | B2+ | C1 | B2+ | B2+ |
| | **Text9** | **Text9** | **Text9** | **Text10** | **Text10** | **Text10** | | | | | | |
| q1genre | inform reader analyse a process | narrate an event inform reader discuss a point | inform reader analyse a process | inform reader analyse a process | inform reader analyse a process | inform reader analyse a process | | | | | | |
| q2source | magazine article | magazine article newspaper article | magazine article newspaper article | textbook chapter | textbook chapter magazine article | textbook chapter magazine article | | | | | | |
| q3audience | 3 | 2 | 3 | 4 | 3 | 4 | | | | | | |
| q4grammar | 3 | 3 | 3 | 4 | 3 | 4 | | | | | | |
| q5vocabulary | 3 | 3 | 3 | 4 | 3 | 4 | | | | | | |
| q6concretenes | 3 | 2 | 3 | 4 | 3 | 4 | | | | | | |
| q7density | 3 | 3 | 3 | 4 | 4 | 4 | | | | | | |
| q8topicspecif | 3 | 2 | 3 | 4 | 3 | 4 | | | | | | |
| q9culturespecif | 1 | 4 | 3 | 1 | 1 | 1 | | | | | | |
| q10sentcohesi | 1 | 2 | 2 | 1 | 2 | 2 | | | | | | |
| q11coherence | 1 | 2 | 2 | 1 | 2 | 1 | | | | | | |
| q12diffculty | 2 | 3 | 3 | 3 | 3 | 4 | | | | | | |
| q13 suitability | Yes | Marginally | Yes | Yes | Yes | Yes | | | | | | |
| CEFR | B2+ | B2 | B2+/C1 | C1 | B2+ | B2+/C1 | | | | | | |