

The Relationship between Language Complexity and Test-taker Achievement on a High-Stakes Test of Writing

Leda Lampropoulou

Irene Stoukou

David Coniam

September 2025

Authors

Leda Lampropoulou is Research Manager at LanguageCert, where she coordinates the organisation's research programme. She has over 15 years of professional experience in the development, validation, and evaluation of high-stakes language tests, with a particular focus on ensuring the reliability, validity, and fitness-for-purpose of LanguageCert's examination portfolio. In her role, she also contributes to the maintenance of quality assurance processes and alignment with international standards in language testing. She holds a BA in English Language and Philosophy from the University of London and an MA in Language Testing from Lancaster University.

Irene Stoukou is Research Associate at LANGUAGECERT. She is responsible for the analysis and monitoring of examiner performance, ensuring marking consistency and accuracy. She holds a PhD in English Literature and Culture (Aristotle University of Thessaloniki), an MA in Modern and Contemporary Literature, Culture, and Thought (University of Sussex), and a BA in English Language and Literature (Aristotle University of Thessaloniki). She has extensive EFL teaching experience in diverse educational settings. She is a member of IATEFL, UKALTA and is also affiliated with the European Association for the Study of English (ESSE), where she serves as a national correspondent for the Gender Studies Network. She is a Postdoctoral Researcher and Adjunct Lecturer in the School of English Language and Literature at Aristotle University of Thessaloniki.

David Coniam is Head of Research at LANGUAGECERT. He has been working and researching in English language teaching, education and assessment for over 50 years. His main publication and research interests are in language assessment, language teaching methodology and academic writing and publishing.

Suggested citation

Lampropoulou, L., Stoukou, I., & Coniam, D. (2025). *The relationship between language complexity and test-taker achievement on a high-stakes test of writing*. LANGUAGECERT.

Abstract

The current study explored the relationship between language complexity and test-taker proficiency at Common European Framework of Reference for Languages (CEFR) levels B1, B2, and C1 from scripts extracted from a high-stakes examination – the LANGUAGECERT Academic Writing Test. The purpose of the study involved exploring the extent to which test takers obtaining higher grades on a writing test actually produce more advanced-level writing skills, providing possible validity evidence for a high-stakes test of writing. 2,746 scripts, produced for Task 1 and Task 2 of the LANGUAGECERT Academic Writing test, which had been graded as CEFR levels B1, B2, or C1, were passed through the computational linguistic analytic tool Coh-Metrix 3.0. The scripts were analysed against nine categories of lexical, syntactic and discourse level features in 62 individual subcategories. ANOVAs were conducted on each of the 62 subcategories against CEFR level.

In total, 68% (42 of the 62) of the subcategories followed the expected progression. Texts produced by B2 test takers showed a greater number of higher-level linguistic features than B1 texts, and C1 texts in turn showed a greater number of higher-level linguistic features than those of B2 test takers. The B1 to B2 to C1 progression on the 42 subcategories was also seen to be significant. The conclusion drawn is that with over two thirds of linguistic features affirming that higher ability test takers do produce more complex texts than lower-ability test takers, the LANGUAGECERT Academic Writing Test performs as it is intended to, grading test takers appropriately.

Introduction

The Common European Framework of Reference for Languages (CEFR) provides detailed descriptions of language proficiency across the six different levels (A1 to C2). As learners progress through these levels, their language ability is expected to show increased lexical, syntactic and discourse complexity.

Against this backdrop, the relationship between linguistic complexity and L2 proficiency has been a longstanding focus of research. The underlying assumption is that the use of more complex lexical, syntactic and discourse structures within a text indicate more advanced-level writing skills (Crossley, 2020).

Crossley (2020) presents a cogent overview of the interactions between lexical, syntactic and discourse structures and L2 proficiency. While he highlights how more proficient L2 writers generally illustrate greater ability in their control of lexical, syntactic and discourse structures, he nonetheless urges that the relationship



between the features in terms of complexity and L2 proficiency is dependent upon various contextual factors and individual differences.

The focus of Crossley's (2020) study is essentially argumentative writing, where "L2 writing is often used as a proxy of language ability in both standardised tests and language acquisition studies." In the current paper, this assumption is related directly to LANGUAGECERT's Academic test of English (LCA).

The LCA is designed as a multi-level test, aligned with the CEFR, spanning levels B1 to C2. As outlined in the LANGUAGECERT Academic Qualification Handbook (Version 7.0) provided at <https://www.languagecert.org/en/language-exams/english/languagecert-academic/-/media/d4f4c8da7ef24dd6be4666c7729308d5.ashx>, the design of the LANGUAGECERT Academic test allows test takers' performance to be mapped against a continuous proficiency scale, rather than being confined to a single level. The test is underpinned by a complex set of constructs that lay out what is intended to be assessed at each CEFR level in terms of lexis, grammar, syntax and discourse features

The analytic rating scale descriptors that the markers use reflect a clear hierarchy, or cline, of linguistic and discourse competence, ensuring that achieving a score tied to a higher CEFR level requires greater mastery of complex structures, a wider range of vocabulary, and more sophisticated organisation of ideas as test takers move from B1 to C2. Thus, while all test takers engage with the same task types, the expected level of linguistic range and control, syntactic complexity, and discourse management increases progressively at each successive CEFR level.

Building on the background presented above, it is reasonable to assume that texts produced by test takers are similarly graduated in terms of complexity and difficulty. Nonetheless, it is this construct that the current study examines and attempts to validate through a detailed exploration of test-taker lexical, syntactic and discourse level features. The premise underpinning the current study and driving the research question is, to reiterate the thesis just stated, that texts produced by test takers at higher CEFR levels (namely C1) will illustrate greater complexity than texts produced by test takers at lower CEFR levels (namely B1).

As a lead-in to the current study, an overview of research will now be provided as to how language proficiency supports the progression of linguistic complexity across CEFR levels. As an elaboration, some key research findings will first be provided under each heading.



1. Lexical Complexity

Lexical complexity is commonly analysed to assess linguistic ability. It generally encompasses several aspects of the vocabulary used in a text, including lexical diversity (the number of unique words), lexical density (the ratio of content words to function words), and lexical sophistication (the proportion of advanced words) (Crossley, 2020). Lexical richness and sophistication develop as learners move up the CEFR scale, and research into lexical diversity and lexical development has consistently reported that lexical complexity increases as learners progress from lower to higher proficiency levels – see González, 2017; Crossley et al., 2011; Zareva, 2007.

Read (2000) illustrates how higher CEFR learners (B2 and above) use more low-frequency words and academic or technical terms. This reflects Nation's model of vocabulary depth (2001), which shows that advanced learners develop not just more vocabulary but also a deeper understanding of word nuances, collocations, and connotations.

2. Syntactic Complexity

Syntactic complexity refers to the sophistication and variety of syntactic forms produced in language, and is considered a key aspect of L2 writing development and an indicator of more advanced writing skills. Research shows that syntactic complexity increases significantly as learners progress through the CEFR levels. There is considerable empirical evidence in the second language acquisition (SLA) literature of a strong link between the (syntactic) complexity of learners' L2 productions and their overall level of L2 development and proficiency (Ortega, 2003; Vyatkina, 2013; Wolfe-Quintero, et al., 1998).

In Martínez's (2018) analysis of syntactic complexity of writing at different proficiency levels, complexifications at sentence, clause and phrase levels of syntactic organisation revealed high correlations between scores and virtually all complexity metrics.

3. Discourse Complexity

Discourse complexity refers to the ability to construct and sustain coherent and cohesive extended texts. It encompasses the effective use of organisational patterns, logical sequencing of ideas, and appropriate cohesive devices, all of which have been shown to improve with increasing language proficiency.

Crossley and McNamara (2011) and Schiftner-Tengg (2022), for example, examined the use of discourse markers and connective devices as a sign of discourse



complexity. They illustrate how learners at lower CEFR levels rely on simple cohesive devices while more proficient learners (B2-C2) use a broader range of discourse markers to create more nuanced and sophisticated text structures.

In the context of text structure and coherence, Weigle (2002) reported how C1-C2 level writers create more structurally complex texts, with a clear opening, development, and conclusion, using cohesive ties that guide the reader through the progression of an argument. Lower-level writers (A1-B1), in contrast, struggle with maintaining coherence over extended discourse and rely on simple, sequential structures.

At higher CEFR levels, learners show the ability to handle complex argumentation and extended discourse. In research into argumentation and elaboration, for example, Byrnes and Manchón (2014) demonstrated how B2-C1 learners maintained topic continuity and developed arguments over longer stretches of writing, employing higher-level discourse strategies such as counter-arguments and concessions.

Lu (2011) reported how higher proficiency writers use a wider range of clause types, more sophisticated discourse markers, and varied lexical choices. This contrasts with lower-level learners, whose writing tends to be more formulaic, and syntactically simple.

In findings related directly to assessment, Green (2012) and Hawkins & Filipovic (2012) report how learners in higher CEFR bands consistently outperform lower-level learners in tasks involving the use of complex syntax, vocabulary, and extended discourse.

Research across syntactic, lexical, and discourse domains aligns with the CEFR's descriptions of language proficiency, confirming that as learners move through levels from A1 to C2, their ability to handle more complex grammar, richer vocabulary, and extended discourse improves. This progression is empirically supported by findings in SLA and assessment studies.

The Current Study

Building on the research reported above, the current study investigates test-taker performance at CEFR levels B1, B2, and C1 on the LANGUAGECERT Academic Writing Test (LCAWT).

The LANGUAGECERT Academic Writing Test

The LCAWT is a multilevel test designed to assess the ability to understand, use, and communicate effectively in English within an academic setting. The LCAWT lasts about 50 minutes, in which test takers are expected to produce two pieces of expository



writing: one of 150-200 words and a second of around 250 words. While the LCAWT generates scores ranging from A1 to C2, its purpose is essentially to indicate readiness for tertiary-level study; consequently, the majority of test-taker grades emerge at B1, B2, and C1 levels. Three sample test-taker scripts from the second, longer, writing task are provided in Appendix C: a B1 test taker scoring 13/32; a B2 test taker scoring 19/32; and a C1 test taker scoring 27/32 in the task.

The Writing tasks are marked against rating scales aligned to the descriptors of the CEFR. These rating scales are: *Task Fulfilment*; *Accuracy and Range of Grammar*; *Accuracy and Range of Vocabulary*; and *Organisation and Coherence*. Test-taker performance on each rating scale is rated on a nine-level scale (0–8), yielding a maximum possible score of 32 per task. To illustrate the *Accuracy and Range of Vocabulary* criterion, the Task 2 markscheme outlines specific qualitative descriptors at different score levels. For a mid-level mark, candidates “use a range of vocabulary, using simple items accurately and attempting more complex forms,” with “some errors in usage/spelling/word formation which normally do not impede meaning, but can cause some re-reading.” In the top band, candidates “use a wide range of vocabulary, including less-common items, with fluency and sophistication, and to give style,” making “very few errors in usage or spelling which only occur as slips,” and “can convey precise meaning.” Such descriptors exemplify the increasing lexical sophistication and accuracy expected across proficiency levels.

Writing scripts are independently scored by two trained human markers via LANGUAGECERT’s marking environment. Markers assess scripts electronically, entering scores for each criterion. The final score for each writing task is calculated as the average of the two markers’ ratings across all four criteria. If a significant discrepancy arises between the markers’ scores, the script is referred to a Chief Examiner, whose decision is final.

To ensure scoring consistency and validity within the multi-level framework, inter-rater and intra-rater reliability analyses identify markers whose scoring may be inconsistent, leading to retraining or closer monitoring. Chief Examiners also second-mark a random sample of 10% of scripts to ensure marking quality, alongside targeted re-marking of tests.

Sample

The sample for the current study comprises 2,746 scripts—1,373 scripts from Task 1 and 1,373 scripts from Task 2—produced by approximately 1,373 test takers who sat the LCAWT between mid-2023 and mid-2024. Given that most grades fall within the CEFR levels B1, B2, and C1, the scope of the analysis and discussion in the current

study is limited to these three CEFR levels. The distribution of the productions between levels is presented in Table 1 below.

Table 1: *CEFR sample spread*

CEFR level	N of scripts
B1	1,131
B2	1,241
C1	374

Methodology and Analysis

The analysis in the current study involves the use of language feature analysis via the computational linguistic analytic tool *Coh-Metrix 3.0* (Graesser et al., 2004).

Coh-Metrix has come to be an accepted and well-established computational tool in the analysis of written text. It has been used in a variety of contexts; some of its uses in previous research are provided below.

- to assess college students' writing quality (Li & Liu, 2017; Riazi, 2016)
- to predict quality in argumentative writing (MacArthur et al., 2019)
- to investigate quality in the written output of both first- and second-language writers (Crossley & McNamara, 2011)
- to classify written assessments and automate text evaluation (McNamara et al., 2014; McNamara et al., 2015)
- to explore how textual features influence judgments of writing quality (Crossley et al., 2011)
- to explore the development of writing in novice writers (Shpit, 2022)
- to explore developments in text cohesion in writing (Wen, 2023)

Coh-Metrix indices appear as eleven sections, although the sections are not systematically arranged into groups which directly denote lexical, syntactic or discourse level features. Table 2 presents the current section order as it is laid out in Coh-Metrix 3.0 (Graesser et al., 2004).

Table 2: *Coh-Metrix sections*

Coh-Metrix section	Description
Text Descriptives	<i>Basic counts and rates (e.g., number of words, sentences, syllables, type-token ratio)</i>
Text Easability	<i>Factors derived from factor analysis, such as narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion</i>
Readability	<i>Traditional readability indices like Flesch Reading Ease and Flesch-Kincaid Grade Level</i>
Referential Cohesion	<i>Measures of how often words and concepts are repeated or overlap across sentences (e.g., stem overlap, noun overlap)</i>
Lexical Diversity	<i>Type-token ratio, word frequency, and other metrics of lexical variation</i>
Connectives	<i>Frequency and type of logical connectives (e.g., causal, temporal, adversative)</i>
L2 Readability	<i>Adapted readability indices aimed at second-language learners</i>
Syntactic Complexity	<i>Measures like mean number of words before the main verb, number of modifiers, and syntactic pattern densities</i>
Syntactic Pattern Density	<i>Specific syntactic structure frequencies (e.g., infinitive forms, passive voice)</i>
Word Information	<i>Word concreteness, imagability, familiarity, meaningfulness (from psycholinguistic norms)</i>
Latent Semantics Analysis (LSA)	<i>Measures of semantic similarity between sentences and across the text, based on LSA vector space</i>

Across the 11 sections, 108 measures, or subcategories, are reported.

As Latifi and Gierl (2021) note, incorporating all sections and subcategories would possibly ‘overshadow’ actual informativeness, while the relevance and informativeness of certain sections or subcategories of features are context-dependent; they consequently reduce the categories they use to those which provide the most information. A similar move has been taken in the current study, where the dataset has been reduced to nine sections comprising the 62 most informative subcategories. These are listed in Appendix A. Table 3 provides a summary. It should be noted that the sections have been rearranged such that lexical, syntactic and discourse level features have been grouped together.

Table 3: *Categories used in the current study*

Level	Categories	No.
Lexical	Word Information	11
N=18	Lexical Diversity	4
	Latent Semantics	3
Syntactic	Syntactic Complexity	7
N=15	Syntactic Pattern Density	8
Discourse	Text Easability	8
N=29	Referential Cohesion	10
	Situation Model	8
	Readability	3
	Total	62

With certain Coh-Metrix categories, a higher score is indicative of an easier text, while, with other categories, a higher score is indicative of a more complex text. Such detail is provided in the rightmost column of Appendix A.

Research Questions

The research question being pursued in the study may be framed from two perspectives.

To what extent do Coh-Metrix scores on the three CEFR levels of B1, B2, C1:

- (1) increase across CEFR levels (B1 to C1) in categories associated with greater linguistic complexity?
- (2) decrease across CEFR levels (B1 to C1) in categories associated with greater linguistic simplicity?

Statistical Analysis

Each of the 62 Coh-Metrix subcategories was subjected to a one-way ANOVA using the software JASP (JASP Team, 2024), comparing Coh-Metrix scores across the three CEFR levels (recoded as B1=1; B2=2; C1=3). Assumptions of normality and homogeneity of variances were checked prior to analysis. For ease of readability in each category in Appendix B, a visual colour-coding scheme is used. A consistent decrease in score from B1 to B2 to C1 indicates texts coded as linguistically 'easier' across levels. Conversely, a consistent increase in score from B1 to B2 to C1 indicates texts which are coded as linguistically 'more demanding' across levels.

In the context of the research questions above, this means that **yellow** scores across Coh-Metrix subcategories indicate more complex texts being produced from B1 to C2, while **green** scores across subcategories indicate easier texts being produced from B1 to C2. This coding and detailed results can be found in Appendix B.

Results

Table 4 presents a picture of the number of Coh-Metrix subcategories where scores followed the expected progression across CEFR levels (B1 to B2 to C1). The fourth column ("Directional Progression [B1 to B2 to C1]") indicates the number of subcategories showing this trend. The fifth ("Significance [p<.001]") identifies how many of these trends were statistically significant based on the one-way ANOVAs. The final column combines the results of the fourth and fifth columns, showing the number and percentage of subcategories in which both progression and statistical significance were observed.

Table 4: Progression across the CEFR levels vis-à-vis Coh-Metrix categories

Level	Category	Subcats	Directional Progression (B1 to B2 to C1)	Significance (p<.001)	Progression & Significance
Lexical N=18	Word Information	11	9	8	73%
	Lexical Diversity	4	4	4	100%
	Latent Semantics	3	2	2	67%
	<i>Subtotal</i>	18	15	14	77.8%
Syntactic N=15	Syntactic Complexity	7	4	3	43%
	Syntactic Pattern Density	8	7	4	50%
	<i>Subtotal</i>	15	11	7	46.7%
Discourse N=29	Text Easability	8	7	6	75%
	Referential Cohesion	10	10	10	100%
	Situation Model	8	4	3	38%
	Readability	3	2	2	67%
	<i>Subtotal</i>	29	23	21	72.4%
<i>Grand total</i>		62	49	42	67.7%



As can be seen, in 42 of the 62 subcategories (i.e., 67.7%), the direction followed the B1 to B2 to C1 progression and was statistically significant. This pattern was particularly pronounced in the *Lexical* domain, where 14 of 18 indices (77.8%) showed both progression and significance. The *Syntactic* domain showed a lower rate (46.7%), while the *Discourse* domain showed a strong alignment (72.4%), particularly in Referential Cohesion, where all 10 subcategories exhibited both progression and significance.

Discussion

A discussion is provided below on each level of delicacy, summarising the details of the different categories and the subcategories beneath them.

Lexical Level

The *Lexical* level explored features related to vocabulary sophistication and variation, across 18 Coh-Metrix subcategories.

In the *Word Information* category, eight of the eleven subcategories, including content word familiarity, concreteness, imaginability, meaningfulness, and hypernymy for nouns and verbs, showed a significant increase from B1 to C1.

All four *Lexical Diversity* subcategories (e.g., type-token ratios and measures of textual and vocabulary density) also exhibited consistent and statistically significant upward trends.

In the *Latent Semantic* category, two of the three measures (e.g., sentence-to-sentence overlap), showed similar significant progression.

In total, 14 out of 18 Lexical level indices (77.8%) aligned with both the expected directional increase and statistical significance, pointing to a strong association between CEFR level and lexical development in the produced scripts.

Syntactic Level

The *Syntactic* level explored sentence structure and grammatical complexity of texts, across 15 syntactic subcategories.

Regarding *Syntactic Complexity*, three of the seven subcategories, including noun phrase modifiers, minimal edit distance (i.e., the average distance between syntactically related words), and sentence syntax similarity, showed a significant upward trend across CEFR levels.



As for *Syntactic Pattern Density*, five subcategories – adverbial, prepositional, gerund, infinitive and agentless passive voice density – followed the B1 to B2 to C1 progression and were significant.

Overall, 7 of the 15 *Syntactic* level indices (46.7%) demonstrated both directional progression and statistical significance. This lower rate compared to lexical and discourse measures may suggest that syntactic development in writing scripts is more gradual or variable, and may not be fully captured by surface-level syntactic features alone.

Discourse Level

The *Discourse* level explored how information is organised at the level of sentences and larger units, such as paragraphs and whole texts, as well as the effect of cohesion and coherence. There were 29 subcategories investigated at this level.

Concerning *Text Easability*, six of the eight measures, followed the B1 to B2 to C1 progression and were significant. These were: narrativity, word concreteness, referential cohesion, deep cohesion, verb cohesion and connectivity.

As regards *Referential Cohesion*, all ten measures exploring overlap across sentences in terms of nouns, arguments, stems, content words and anaphor followed the B1 to B2 to C1 progression and were significant.

With *Situation Model* category, three of eight indices, reflecting the extent to which texts support mental model construction, showed significant increases.

With *Readability*, two of three subcategories, including the Flesch-Kincaid Grade Level, also progressed significantly with CEFR level.

In total, 21 of the 29 *Discourse* level indices (72.4%) demonstrated both directional progression and statistical significance. These findings highlight that discourse-level features, particularly those related to cohesion and textual integration, are strongly associated with increases in writing proficiency across CEFR bands.

Conclusion

The current study explored the relationship between lexical, syntactic and discourse features and CEFR proficiency levels in the high-stakes LANGUAGECERT Academic Writing Test, focusing on texts rated at B1, B2, and C1. The purpose of the study was to provide external validation, or triangulation, of the assumption that higher CEFR-level grades reflect greater linguistic complexity, and thus, higher language ability in writing.



LANGUAGECERT's English language writing assessments are grounded in CEFR-aligned constructs that define expectations for lexis, grammar, syntax, and discourse features at each level. The associated rating scales similarly describe an increasing level of complexity across CEFR bands. It is therefore expected that performance should reflect this progression in measurable ways. The current study sought to validate this expectation by analysing linguistic features in test-taker scripts using the computational tool Coh-Metrix 3.0, a widely recognised instrument for text analysis.

The research hypothesis driving the study was that texts produced by test takers achieving higher CEFR levels (such as C1) would illustrate greater complexity than texts produced by test takers at lower CEFR levels (such as B1).

A total of 2,746 test-taker scripts, produced for both Task 1 and Task 2 across 62 Coh-Metrix subcategories spanning lexical, syntactic, and discourse domains, were analysed. One-way ANOVAs were conducted to examine differences across proficiency levels. Results showed that 49 out of 62 subcategories (79.0%) displayed a directional increase from B1 to B2 to C1, and 42 subcategories (67.7%) demonstrated statistically significant differences.

At the lexical level, 14 of the 18 measures (77.8%) emerged as following the B1 to B2 to C1 progression and as being significant.

At the syntactic level, 7 of the 15 measures (46.7%) emerged as following the B1 to B2 to C1 progression and as being significant.

At the discourse level, 21 of the 29 measures (72.4%) emerged as following the B1 to B2 to C1 progression and as being significant.

These findings provide considerable empirical support for the construct validity of the LANGUAGECERT Academic Writing Test: more proficient candidates produce texts that are measurably more complex, particularly in terms of vocabulary use and discourse cohesion.

The conclusion that may be drawn is that with over two thirds of lexical, syntactic and discourse level features affirming that higher-ability test takers do produce more complex texts than lower-ability test takers, the LANGUAGECERT Academic Writing test functions as intended, effectively distinguishing among levels of proficiency. While the notion might appear somewhat simplistic that a multilevel test should assign more able test takers to higher levels and lower-ability ones to lower levels, the current research confirms that this is the case. The thesis possibly also needs to be considered from the opposite perspective: if the categories did not in general follow the expected progression, this might well give cause for concern.



While the study provides strong validation evidence, it is limited to three CEFR levels and the written modality. Further research would aim to cover the whole gamut of the six-level CEFR scale and extend the analysis to include the speaking test. Additionally, while this study employed a quantitative approach, further research could complement it with qualitative analysis of linguistic features in candidate texts (e.g., through detailed discourse analysis or expert ratings), to explore how complexity manifests in more nuanced ways.

Overall, the findings from this large-scale analysis offer strong support for the construct validity of the LANGUAGECERT Academic Writing Test and reinforce its alignment with CEFR proficiency levels. By applying computational linguistic tools to authentic test-taker output, this study contributes to both the validation of high-stakes assessment and the broader understanding of language development. As demand grows for transparent, evidence-based language testing, this research demonstrates the value of integrating automated linguistic analysis into the validation process, helping ensure that test scores are both interpretable and meaningful for stakeholders.

Acknowledgement


We are grateful to the developers of Coh-Metrix for permitting us access to the Coh-Metrix 3.0 software.

References

- Byrnes, H., & Manchón, R. M. (Eds.). (2014). *Task-Based Language Learning: Insights from and for L2 Writing*. John Benjamins Publishing Company. <https://doi.org/10.1002/tesq.388>
- Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3), 415-443.
- Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life-Long Learning*, 21(2/3), 170-191. <http://doi.org/10.1504/IJCEELL.2011.040197>
- Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2), 115-135. <https://doi.org/10.1111/j.1467-9817.2010.01449.x>
- Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29, 243-263. <https://doi.org/10.1177/0265532211419331>
- González, M. C. (2017). The contribution of lexical diversity to college-level writing. *TESOL Journal*, 8(4), 899-919. <https://doi.org/10.1002/tesj.342>
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193-202. <https://doi.org/10.3758/BF03195564>
- Green, A. (2012). *Language functions revisited: Theoretical and empirical bases for language construct definition across the ability range*. Cambridge University Press.
- Hawkins, J. A., & Filipović, L. (2012). *Criterial FEATURES in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge University Press.
- JASP Team (2024). JASP (Version 0.19.0)[Computer software]. <https://jasp-stats.org/>
- Latifi, S., & Gierl, M. (2021). Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing. *Language Testing*, 38(1), 62-85. <https://doi.org/10.1177/0265532220929918>
- Lu, X. F. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1), 36-62. <https://doi.org/10.5054/tq.2011.240859>



- Li, X., & Liu, J. (2017). Automatic essay scoring based on Coh-Metrix feature selection for Chinese English learners. In T. Huang, R. Lau, Y. Huan, M. Spaniol, & C. Yuen. (Eds.), *International symposium on emerging technologies for education* (pp. 382–393). Springer. https://doi.org/10.1007/978-3-319-52836-6_40
- MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing*, 32(6), 1553–1574. <https://doi.org/10.1007/s11145-018-9853-6>
- Martínez, A. C. L. (2018). Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels. *Assessing Writing*, 35, 1–11. <https://doi.org/10.1016/j.asw.2017.11.002>
- McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23, 35–59. <https://doi.org/10.1016/j.asw.2014.09.002>
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511894664>
- Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge University Press.
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4), 492–518. <https://doi.org/10.1093/applin/24.4.492>
- Read, J. (2000). *Assessing Vocabulary*. Cambridge University Press.
- Riazi, A. M. (2016). Comparing writing performance in TOEFL-iBT and academic assignments: An exploration of textual features. *Assessing Writing*, 28, 15–27. <https://doi.org/10.1016/j.asw.2016.02.001>
- Schiftner-Tengg, B. (2022). Analysing discourse coherence in students' L2 writing: Rhetorical structure and the use of connectives. In Berger, A., Heaney, H., Resnik, P., Rieder-Bünemann, A., Savukova, G. (Eds.), *Developing advanced English language competence: A research-informed approach at tertiary level* (pp. 237–255). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-79241-1_23

- 
- Shpit, E. I. (2022). The use of Coh-Metrix by individual Russian novice writers for developing self-assessment and self-correction skills. *International Journal of English for Specific Purposes*, 3(1), 6-33.
- Vyatkina, N. (2013). Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. *Modern Language Journal*, 97(S1), 11–30.
<https://doi.org/10.1111/j.1540-4781.2012.01421.x>
- Weigle, S. C. (2002). *Assessing Writing*. Cambridge University Press.
- Wen, J. (2023). The changes in text cohesion of senior high school students measured by Coh-Metrix as a function of grade level. *International Journal of Social Science and Education Research*, 6(8), 109-119. [https://doi.org/10.6918/IJOSSER.202308_6\(8\).0014](https://doi.org/10.6918/IJOSSER.202308_6(8).0014)
- Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, Hi: University of Hawaii, Second Language Teaching & Curriculum Center.
- Zareva, A. (2007). Structure of the second language mental lexicon: How does it compare to native speakers' lexical organization?. *Second Language Research*, 23(2), 123–153.
<https://doi.org/10.1177/0267658307076543>

Appendix A: Coh-Metrix Categories and Subcategories Used in the Analysis

Table A1: *Lexical level*

LEXICAL LEVEL			
No.	Subcategory	Category and subcategory gloss	Higher score =
Word Information			
95	WRDFRQa	CELEX Log frequency for all words	Easier
96	WRDFRQmc	CELEX Log minimum frequency for content words	Easier
97	WRDAOAc	Age of acquisition for content words	More Complex
98	WRDFAMc	Familiarity for content words	Easier
99	WRDCNCc	Concreteness for content words	Easier
100	WRDIMGc	Imagability for content words	Easier
101	WRDMEAc	Meaningfulness, content words	Easier
102	WRDPOLc	Polysemy for content words	More Complex
103	WRDHYPn	Hypernymy for nouns	More Complex
104	WRDHYPv	Hypernymy for verbs	More Complex
105	WRDHYPnv	Hypernymy for nouns and verbs	More Complex
Lexical Diversity			
48	LDTTRc	Lexical diversity, type-token ratio, content word lemmas	More Complex
49	LDTTRa	Lexical diversity, type-token ratio, all words	More Complex
50	LDMTLDa	Lexical diversity, MTLD, all words	More Complex
51	LDVOCDa	Lexical diversity, VOCD, all words	More Complex
Lexical Semantic Analysis (LSA)			
40	LSASS1	LSA overlap, adjacent sentences	Easier
42	LSASSp	LSA overlap, all sentences in paragraph	Easier
46	LSAGN	LSA given/new, sentences	Easier

Table A2: Syntactic level

SYNTACTIC LEVEL			
No.	Subcategory	Category and subcategory gloss	Higher score =
		Syntactic Complexity	
69	SYNLE	Left embeddedness, words before main verb	More Complex
70	SYNNP	Number of modifiers per noun phrase	More Complex
71	SYNMEDpos	Minimal Edit Distance, part of speech	More Complex
72	SYNMEDwrd	Minimal Edit Distance, all words	More Complex
73	SYNMEDlem	Minimal Edit Distance, lemmas	More Complex
74	SYNSTRUTa	Sentence syntax similarity, adjacent sentences.	More Complex
75	SYNSTRUTt	Sentence syntax similarity, all combinations, across p'graphs	More Complex
		Syntactic Pattern Density	
76	DRNP	Noun phrase density, incidence	More Complex
77	DRVP	Verb phrase density, incidence	More Complex
78	DRAP	Adverbial phrase density, incidence	More Complex
79	DRPP	Preposition phrase density, incidence	More Complex
80	DRPVAL	Agentless passive voice density, incidence	More Complex
81	DRNEG	Negation density, incidence	Easier
82	DRGERUND	Gerund density, incidence	More Complex
83	DRINF	Infinitive density, incidence	More Complex

Table A3: *Discourse level*

		DISCOURSE LEVEL	
No.	Subcategory	Category and subcategory gloss	Higher score =
		Text Easability Principal Components (PC)	
12	PCNARz	PC Narrativity	Easier
14	PCSYNz	PC Syntactic simplicity	Easier
16	PCCNCz	PC Word concreteness	Easier
18	PCREFz	PC Referential cohesion	Easier
20	PCDCz	PC Deep cohesion	Easier
22	PCVERBz	PC Verb cohesion	Easier
24	PCCONNz	PC Connectivity	Easier
26	PCTEMPz	PC Temporality	Easier
		Referential Cohesion	
28	CRFNO1	Noun overlap, adjacent sentences	Easier
29	CRFAO1	Argument overlap, adjacent sentences	Easier
30	CRFSO1	Stem overlap, adjacent sentences	Easier
31	CRFNOa	Noun overlap, all sentences	Easier
32	CRFAOa	Argument overlap, all sentences	Easier
33	CRFSOa	Stem overlap, all sentences	Easier
34	CRFCWO1	Content word overlap, adjacent sentences	Easier
36	CRFCWOa	Content word overlap, all sentences	Easier
38	CRFANP1	Anaphor overlap, adjacent sentences	Easier
39	CRFANPa	Anaphor overlap, all sentences	Easier
		Situation Model	
61	SMCAUSv	Causal verb incidence	Easier
62	SMCAUSvp	Causal verbs and causal particles incidence	Easier
63	SMINTEp	Intentional verbs incidence	Easier
64	SMCAUSr	Ratio of casual particles to causal verbs	Easier
65	SMINTER	Ratio of intentional particles to intentional verbs	Easier
66	SMCAUSlsa	LSA verb overlap	Easier
67	SMCAUSwn	WordNet verb overlap	Easier
68	SMTEMP	Temporal cohesion, tense & aspect repetition	Easier
		Readability	
106	RDFRE	Flesch Reading Ease	Easier
107	RDFKGL	Flesch-Kincaid Grade Level	More Complex
108	RDL2	Coh-Metrix L2 Readability	Easier

Appendix B: Analysis of Coh-Metrix Subcategories by CEFR Level

Colour coding key:

Green: Higher scores = easier, advanced writers should score lower

Yellow: Higher scores = more demanding, advanced writers should score higher

Table B1: *Lexical level*

	Subcategory	B1 mean	B2 mean	C1 mean	Significance	Progression & Significance
Word Information	WRDFRQa	3.10	3.06	2.99	0.001	✓
	WRDFRQmc	0.87	0.94	0.83		
	WRDAOAc	371.81	375.26	378.25	0.001	✓
	WRDFAMc	579.32	576.42	573.40	0.001	✓
	WRDCNCc	375.46	374.92	370.65	0.02	✓
	WRDIMGc	405.90	405.78	402.68	0.08	
	WRDMEAc	434.43	433.53	431.26	0.001	✓
	WRDPOLc	3.70	3.70	3.67		
	WRDHYPn	5.90	6.13	6.39	0.001	✓
	WRDHYPv	1.45	1.51	1.58	0.001	✓
	WRDHYPnv	1.62	1.69	1.77	0.001	✓
Lexical Diversity						
	LDTRc	0.72	0.74	0.76	0.001	✓
	LDTRa	0.55	0.56	0.58	0.001	✓
	LDMTLD	77.82	84.46	98.38	0.001	✓
Latent Semantics	LDVOCd	75.86	82.50	94.00	0.001	✓
	LSASS1	0.22	0.21	0.20	0.001	✓
	LSASSp	0.20	0.19	0.18	0.01	✓
	LSAGN	0.27	0.28	0.28		

Table B2: *Syntactic level*

	Subcategory	B1 mean	B2 mean	C1 mean	Significance	Progression & Significance
Syntactic Complexity	SYNLE	7.21	6.41	6.47		
	SYNNP	0.70	0.75	0.79	0.001	✓
	SYNMEDpos	0.63	0.63	0.63		
	SYNMEDwrd	0.85	0.87	0.88	0.001	✓
	SYNMEDlem	0.83	0.85	0.86	0.001	✓
	SYNSTRUTa	0.08	0.09	0.10	0.07	
	SYNSTRUTt	0.08	0.08	0.08		
Syntactic Pattern Density						
	DRNP	411.45	402.70	392.42	0.001	
	DRVP	199.74	200.64	202.19	0.20	
	DRAP	31.92	32.40	34.02	0.001	✓
	DRPP	115.76	121.32	119.74		
	DRPVAL	4.90	6.66	7.01	0.001	✓
	DRNEG	7.10	6.48	6.44	0.07	
	DRGERUND	18.34	21.05	24.03	0.001	✓
	DRINF	17.25	18.87	20.86	0.001	✓

Table A3: *Discourse level*

	Subcategory	B1 mean	B2 mean	C1 mean	Significance	Progression & Significance
Text Easability	PCNARz	0.15	-0.07	-0.27	0.001	✓
	PCSYNz	-1.31	-0.98	-0.95	0.001	✓
	PCCNCz	0.23	0.19	0.14	0.001	✓
	PCREFz	0.60	0.22	-0.20	0.001	✓
	PCDCz	0.75	0.62	0.59	0.06	
	PCVERBz	0.60	0.37	0.02	0.001	✓
	PCCONNz	-2.10	-2.19	-2.25	0.04	✓
	PCTEMPz	-0.95	-0.97	-0.96		
Referential Cohesion	CRFNO1	0.58	0.53	0.49	0.001	✓
	CRFAO1	0.68	0.64	0.61	0.001	✓
	CRFSO1	0.64	0.61	0.58	0.001	✓
	CRFNOa	0.54	0.50	0.46	0.001	✓
	CRFAOa	0.64	0.60	0.57	0.001	✓
	CRFSOa	0.61	0.58	0.56	0.001	✓
	CRFCWO1	0.15	0.12	0.10	0.001	✓
	CRFCWOa	0.13	0.11	0.09	0.001	✓
	CRFANP1	0.37	0.32	0.31	0.001	✓
	CRFANPa	0.20	0.14	0.13	0.001	✓
Situation Model	SMCAUSv	20.39	22.62	22.57		
	SMCAUSvp	33.71	35.57	33.73		
	SMINTEp	10.78	10.74	10.56	0.97	
	SMCAUSr	0.71	0.59	0.47	0.001	✓
	SMINTER	2.10	1.77	1.76	0.001	✓
	SMCAUSlsa	0.11	0.10	0.09	0.001	✓
	SMCAUSwn	0.50	0.51	0.49		
	SMTMP	0.76	0.76	0.76		
Readability	RDFRE	52.27	52.33	46.68	0.001	✓
	RDFKGL	13.13	12.12	12.89		
	RDL2	20.62	17.49	14.71	0.001	✓

Appendix C: Sample Test-Taker Scripts

Task 2 Prompt

Read the following statement and write about the topic.

The internet allows the individual to feel part of a global online community. Some people believe that this community can bring diverse people together and help solve global problems. Others disagree, fearing that it just leads to disagreements that drive people further apart.

Discuss both of these views and give your own opinion.

Write about 250 words.

About the Samples

The following are responses from candidates at three different CEFR levels (B1, B2, C1) in response to the same Task 2 prompt from the LANGUAGECERT Academic Writing Test. Each script is preceded by a set of identifying details in the following format:

<<Level – Total Raw Marks – Gender – Age – L1 – Word Count>>

<<B1 - 14 - Male - 22 - Chinese - 268>>

Nowadays, it is fact that the internet excatly enable us to become a part of the global online community. However, whether the community can bring diverse people together and help solve global problems or can leads to disagreements is a hot issues. My view is that, internet is a useful tool that can let our lifestyle be more efficiently.

Firstly, it is obvious that we can easily to communicate with each other online. This is because there are a large number of softwares that can let us easily to chat each other. Like, wechat, telegram, facebook and so forth. Therefore, all of us from different countries can easily chat each other.

Moreover, we can easily contact the large number of news from different countries online and deal with some global problems. For example, it is convenient for us to by using searching engines like google or watching youtube or bbc news. So, we can rapidly to get the latestly news on time.

That is not to say that internet is always great. There are, of course, disagreements that drive people further apart. For instance, sometimes we ignore chatting by face to face with our friends in the real society. It mean that we will probably lose the relationship between our friends. Therefore, we should improve our relationship with in our community.

In summary, we should correctly to use internet. Meanwile, we should use internet on a regular basis. And the end of the day. Only by doing so can we let the internet be efficient to server us. And we can be together and solve some global problems.

<< B2 - 21 - Male - 21 - Tamil - 250 >>

The internet is considered to be the *utopia* of the current generation. People tend to spend more time on the internet compare to the real-life. This creates a conflict-of-interest between different communities of people with different opinions. In this essay, I will discuss on both the views on the impact of internet on people.

Internet can be both a boon and bane to society. One of the advantages is that the people can connect with more people and create a network of communities based on their interests and opinions. This seems to be a great thing for the people who want to communicate with people from diverse background and exchange their opinions. Some people like this approach as they think that this approach broadens their minds on wide thoughts and learn some new things.

Eventhough, the internet feels like a wonderland, it also has some negative impact on the people, according to some communities. Some people feels that communication only through the internet, seperates the people apart. They think that the people will tend to lack in interactive skill to communicate with other people in real-life. This shows the different opinion of people with different reasons of beneficial approach. Although, there are always a solution for a problem.

In my opinion, people need to interact with other people in real-life, with less time spent on the internet. This approach can not only benefit the interactive knowlege of a person, but also diminishing the characteristic of introvertness in the future days.

<< C1 – 28 – Male – 44 – Greek – 313 >>


The internet is considered by many as the biggest technological advance of the last 30 years that affected most of the human's activities, including every-day communications.

Communications through internet have many forms; simple emails and chat rooms in the first years to instant multimedia messaging, social media, photo and video sharing and even virtual reality rooms in the so-called 'meta-verse'.

The fact that so many people can have instant communication with each other have obvious advantages, Everyone can have access to a large community of people having similar concerns or even better have solutions to similar problems. The bias that the traditional networks (tv, newspapers etc) may have can be overcome by community communications in a freely-run internet. The truth cannot be hidden by governments or large capital organizations, so the people can easily organise and act as a union to a common goal. In politics, one can access a large audience without spending a serious amount of money or having relations to traditional media (tv, newspapers etc.), so we could consider that even the democracy becomes more accessible to anyone.

On the other hand, there are many disadvantages of the communications that the internet brought to our lives. Mostly because of the anonymity, usually the ideas spreading in the community have actually been the closest to the edges in the political spectrum, bringing people even more far away from each other. Ideas backed by fake news circulated in the unregulated internet have brought populist parties in many countries, inflating the problems that the people had.

Internet is a tool and as a tool it can be used to bring benefits or to do harm. My opinion is that the users of the internet should have a very good training from their schools and their families in order to be able to judge what is beneficiary and what is not and act accordingly.



MILLIONS OF EXAMS DELIVERED WORLDWIDE

LANGUAGECERT is an Awarding Organisation recognised by Ofqual. It spearheads innovations in language assessment and certification, providing high-quality services to the global learners' community. It is a UK-based member of PeopleCert Group, a global leader in the certification industry, that delivers millions of exams in over 200 countries.

Learn more about LANGUAGECERT exams at:
www.languagecert.org

LANGUAGECERT is a business name of PeopleCert
Qualifications Ltd, UK company number 09620926