

LANGUAGES

Rethinking Band-Level Differences in IELTS Writing Task 2: A Multidimensional Learner-Corpus Analysis

Quoc Khanh Nguyen¹, Nhat Quang Vo^{*,2} and Thi Thuy Loan To³

¹The University of Danang, Vietnam

²Ho Chi Minh University of Banking, Vietnam

³University of Social Sciences and Humanities, Vietnam National University Ho Chi Minh City, Vietnam

*Corresponding author:

Nhat Quang Vo – Ho Chi Minh University of Banking, Vietnam.

Email: 020401250009@st.buh.edu.vn

Received: 27/07/2026. Revised: 14/08/2026. Accepted: 22/08/2026. DOI: [10.53901/tjs.2026.v01.issue03.art12](https://doi.org/10.53901/tjs.2026.v01.issue03.art12)

Abstract

This study re-examined IELTS Writing Task 2 band-level differences from a multidimensional learner-corpus perspective. Rather than assuming that higher bands can be explained by longer essays, difficult vocabulary, cohesive devices, or sentence complexity, it tested the extent to which selected textual features differentiated essays across Bands 5–8. The final analytic corpus contained 1,912 unique publicly available IELTS Writing Task 2 essays across four human-sourced approximate band groups after exact duplicate screening. Basic text-production features were calculated in Python, while lexical diversity, lexical sophistication, syntactic complexity, and cohesion indices were extracted using TAALED, TAALES, TAASSC, and TAACO. Descriptive statistics, Kruskal–Wallis tests, false discovery rate correction, Holm-adjusted pairwise comparisons, sensitivity analysis, and effect sizes were used to profile band-level variation. The results showed statistically detectable but substantively modest differences. Higher-band essays tended to show greater connector use, higher lexical sophistication, more academic vocabulary, and lower adjacent overlap, while essay length showed only a weak and sensitivity-dependent contrast. Syntactic-complexity indices were weak separators, and no single feature provided a strong band marker within the reliability limits of the approximate band labels. The study therefore characterizes band-level variation in IELTS Writing Task 2 as an uneven multidimensional textual profile rather than a single-feature pattern, offering evidence against formulaic IELTS pedagogy based on word count, rare vocabulary, mechanical connector insertion, or sentence-length expansion.

Keywords: IELTS Writing Task 2; learner corpus; lexical sophistication; syntactic complexity; writing pedagogy.

Introduction

IELTS Writing Task 2 is a high-stakes academic writing task in which test takers respond to a point of view, argument, or problem in at least 250 words within about 40 minutes [1]. In the public IELTS band descriptors, Task 2 writing is evaluated through four criteria: Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy [2]. These criteria suggest that writing quality is multidimensional rather than reducible to a single feature such as essay length, vocabulary difficulty, or sentence complexity.

Previous IELTS Writing Task 2 research has examined linguistic and discourse characteristics of learner scripts across band scores and candidate groups. Riazi and Knox [3], for example, investigated relations among first language, band score, and discourse features in Academic Writing Task 2. Barkaoui [4] examined how linguistic characteristics changed in

IELTS repeaters' Task 2 scripts over repeated test administrations. Other IELTS-related studies have focused more specifically on lexical resource, including lexical density, lexical sophistication, lexical variation, and lexical accuracy [5]. These studies confirm that textual features are relevant to IELTS writing research, but they also make one issue clear: simply showing that higher-band writing is textually different is no longer a strong novelty claim.

The present study therefore reframes the issue. Its purpose is not merely to confirm that Band 8 essays differ from Band 5 essays. Rather, it asks whether the differences are strong enough, consistent enough, and pedagogically interpretable enough to support common IELTS advice such as 'write more', 'use more advanced vocabulary', 'add more linking words', or 'write longer sentences'. This shift is important because IELTS preparation often turns broad descriptors into surface-level rules, even though band descriptors reward communicative control, flexible lexical choice,

coherence, and grammatical range rather than isolated feature frequency.

The contribution of this study is therefore not the identification of another set of statistically significant band differences, but an effect-size-aware test of whether commonly cited textual features are substantial and robust enough to support practical interpretations of IELTS writing. By combining multi-dimensional corpus indices with multiple-comparison correction, lower-higher contrasts, and sensitivity analysis, the study distinguishes statistical detectability from pedagogical importance.

The study is guided by four research questions: (1) How do IELTS Writing Task 2 learner essays differ across Band 5, Band 6, Band 7, and Band 8 in terms of length and organization? (2) How do lexical diversity and lexical sophistication vary across learner essays in Bands 5–8? (3) How do syntactic complexity and cohesion-related features vary across learner essays in Bands 5–8? (4) Which textual features show the clearest differences between lower and higher band groups?

IELTS Writing Task 2 as a Multidimensional Writing Construct

IELTS Writing Task 2 requires test takers to construct an essay response to an argument, point of view, or problem under timed conditions [1]. The band descriptors indicate that performance is evaluated through multiple criteria: Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy [2]. This construct is important for the present study because it implies that band-level writing performance cannot be explained by a single linguistic measure. Recent learner-writing research similarly suggests that proficiency-sensitive textual features should be interpreted in relation to writing constructs rather than as isolated predictors [6]. Accordingly, the analytical framework links lexical diversity and lexical sophistication to Lexical Resource, syntactic complexity to Grammatical Range and Accuracy, and cohesion indices to linguistic resources relevant to Coherence and Cohesion. Length and organization are treated as contextual surface indicators rather than IELTS scoring criteria in themselves, and none of the five analytical domains is interpreted as a direct measure of an IELTS criterion.

Learner Corpus Analysis and Band-Level Textual Profiling

Learner corpus research provides a useful framework for studying systematic patterns in second-language writing. Computer learner corpora have been used to connect second language acquisition, language assessment, and foreign language teaching through empirical analysis of learner-produced texts [7]. In this tradition, learner texts are not treated merely as individual writing samples but as data that can reveal

recurring linguistic patterns across proficiency levels, task types, or learning contexts.

Lexical Diversity and Lexical Sophistication in Learner Writing

Lexical resource is one of the four IELTS Writing Task 2 criteria, but it is itself multidimensional. The public band descriptors refer to lexical range, precision, flexibility, collocation, spelling, and word formation [2]. Corpus-based research therefore needs to distinguish among lexical diversity, lexical sophistication, academic vocabulary, and lexical accuracy rather than treating vocabulary as a single construct.

Lexical diversity refers to the variety of words used in a text. However, traditional type-token ratio is sensitive to text length, which is problematic when essays vary in word count. Kyle et al. [8] note that lexical diversity indices have long been used to estimate vocabulary knowledge or lexical proficiency, but many common indices vary as a function of text length. This supports the use of more robust measures such as MTLD and MATTR, which were calculated using TAALED in the present study.

Lexical sophistication concerns the extent to which a text uses less frequent, more advanced, or more academic vocabulary. TAALES provides a large set of indices for lexical sophistication and includes measures related to frequency, range, academic vocabulary, and psycholinguistic properties [9]. This is relevant to IELTS Writing Task 2 because higher-band writing is expected to show flexible and precise lexical control, but not merely the use of rare words. Nguyen et al. [5] also show that IELTS Task 2 lexical resource can be examined through several dimensions, including lexical density, sophistication, variation, and accuracy.

Syntactic Complexity, Cohesion, and the Risk of Surface-Feature Interpretation

Grammatical Range and Accuracy is another core IELTS Writing Task 2 criterion. In corpus-based analysis, this dimension is often approached through syntactic complexity indices. However, syntactic complexity is not a single construct. Ortega [10] showed that syntactic complexity measures relate to L2 proficiency in complex ways, and different measures may capture different aspects of development.

More recent work has also cautioned against treating longer sentences or more subordination as the only signs of advanced writing. Biber et al. [11] argued that advanced academic writing often relies on phrasal elaboration rather than only clausal subordination. This distinction is important for IELTS learner essays because higher-band writing may involve more controlled nominal and phrasal packaging rather than simply longer sentences.

In this study, cohesion refers to observable linguistic links within and across sentences, including lexical overlap and connective use, whereas coherence refers

to the logical organization and interpretable progression of ideas in the text. The two are related but not equivalent. IELTS assesses them jointly under the Coherence and Cohesion criterion [2], while the automated measures used in the present study primarily capture cohesion-related textual properties. TAACO provides automated indices of local, global, and text cohesion, including overlap and connective-related measures [12]. A text may contain many cohesive devices but still develop ideas poorly, while a text with lower repetition may still be coherent if ideas are logically connected. Connector frequency and overlap indices are therefore useful descriptive signals, but they should not be interpreted as direct measures of IELTS coherence.

Research Gap and Analytical Framework

Previous IELTS Writing Task 2 research has related band scores to discourse and linguistic features [3], examined linguistic change in repeat test takers [4], and analyzed lexical resource through dimensions such as lexical density, variation, sophistication, and accuracy [5]. More recent corpus work has also examined the frequency, structure, and function of lexical bundles across IELTS writing score categories [13]. Together, these studies demonstrate the relevance of textual evidence to IELTS writing performance, but less attention has been given to asking, within one analysis, whether differences across several textual domains are large, robust to multiple-comparison control and sensitivity checks, and pedagogically substantial rather than merely statistically detectable. The present study addresses this gap by evaluating the magnitude, consistency, and practical interpretability of band-group differences across five textual domains.

Materials and Methods

Research Design

This study adopted a quantitative corpus-based design to compare textual features of IELTS Writing Task 2 learner essays across four approximate band groups: Band 5, Band 6, Band 7, and Band 8. The design was descriptive and comparative rather than predictive. It did not aim to predict IELTS scores or validate the dataset's band labels. Instead, it examined how selected linguistic and textual features varied across band groups. Because the design was cross-sectional, observed differences between band groups were not interpreted as longitudinal changes within individual writers.

The analysis was organized around five domains of textual evidence: length and organization, lexical diversity, lexical sophistication, syntactic complexity, and cohesion. Python scripts in Google Colab were used for the basic production measures, while the specialized corpus tools were used for domain-specific indices: TAALED for lexical diversity, TAALES for lexical sophistication, TAASSC for syntactic complexity, and

TAACO for cohesion. These tools were chosen because they operationalize different dimensions of written-text analysis and therefore support a multidimensional profile rather than a single-feature account of band-level writing [8], [12], [14].

Dataset and Analytic Sample

The study drew on the chillies/IELTS-writing-task-2-evaluation dataset, which is hosted publicly on Hugging Face [15]. The dataset includes prompt texts, learner essays, evaluation comments, and band labels. For the present analysis, the evaluation comments were excluded because they were not part of the research questions. Only the essay texts, prompt information, and band-level groupings were retained.

The provenance of the band labels required careful interpretation. On the dataset discussion page, in response to a question about synthetic data, the dataset owner clarified that the band scores were collected from IELTS-related websites and forums where essays had been scored by IELTS tutors or experienced learners. The same clarification states that AI tools were used to generate evaluation comments, not the band scores. Therefore, the present study treats the labels as human-sourced approximate band scores, while avoiding any claim that they are official examiner ratings [16].

The dataset metadata are limited in ways that directly affect interpretation. They do not make it possible to verify the original writers, testing conditions, number of raters, rating procedures, inter-rater reliability, or essay-level provenance. Two consequences follow. First, if band labels contain rating noise, especially across adjacent bands, the observed differences may be attenuated because measurement error can reduce the apparent strength of group differences. Second, if some higher-band texts include model essays, heavily edited essays, or AI-assisted texts from public IELTS websites and forums, some observed differences may be inflated rather than attenuated. The corpus is therefore described as a publicly available IELTS Writing Task 2 essay dataset with human-sourced approximate band labels, not as a verified corpus of official IELTS scripts or fully authenticated learner production. Accordingly, the dataset is used as a source of approximate proficiency groupings for examining the robustness of aggregate textual patterns, not as evidence for validating official IELTS band distinctions.

Legal and ethical transparency was also considered. Although the Hugging Face dataset card did not display an explicit license in its metadata, the dataset creator stated publicly that the dataset may be used for academic research under the CC BY 4.0 license [17]. The present study therefore uses the dataset for academic research with attribution to the creator. Individual essay texts are not reproduced in this article; only aggregate statistical results are reported,

and no attempt was made to identify individual writers or reconstruct essay provenance.

The working dataset was prepared in Google Colab. Essays with band labels from 5.0 to 8.5 were grouped into four approximate band categories: Band 5 (5.0–5.5), Band 6 (6.0–6.5), Band 7 (7.0–7.5), and Band 8 (8.0–8.5). From each category, 500 essays were selected using reproducible random sampling with a fixed seed of 42, producing an initial balanced sample of 2,000 essays. Exact duplicate essay texts were then screened after whitespace normalization. One record was retained when identical texts carried the same band-group label, while all instances of an identical text with conflicting band-group labels were excluded. The final analytic sample contained 1,912 unique essays: 478 in Band 5, 479 in Band 6, 486 in Band 7, and 469 in Band 8. Writer identifiers were unavailable in the source dataset, so repeated authorship could not be independently verified. The core columns used in the analysis were `essay_id`, `prompt`, `essay`, `band_clean`, `band_group`, `essay_word_count`, and `paragraph_count`. Prompt distribution was examined descriptively using whitespace-normalized prompt text. The final sample contained 995 distinct prompt strings, with 351 represented in Band 5, 358 in Band 6, 345 in Band 7, and 341 in Band 8. Only 34 prompt strings occurred in all four band groups, and the most frequent prompt accounted for 77 essays (4.0% of the sample; Band 5 = 18, Band 6 = 16, Band 7 = 24, Band 8 = 19). Because prompt assignment was not controlled and most prompt strings were sparsely represented, prompt was treated as contextual metadata rather than modeled as an independent variable.

Data Screening and Feature Extraction

Text length was handled cautiously because several automated indices may be influenced by the amount of text available for analysis. For lexical diversity, the study prioritized indices such as MTLT and MATTR, which are designed to be less dependent on text length than simple type-token ratio [8]. This does not mean, however, that all selected indices are fully length-independent. Cohesion-related overlap indices and some syntactic-complexity indices may still be affected by word count, sentence count, or paragraph structure.

Feature extraction proceeded in two linked stages. First, Python was used to generate basic production features: word count, paragraph count, sentence count, average sentence length, simple type-token ratio, long-word ratio, academic connector counts, and first-person pronoun counts. Second, the essays were submitted to TAALED, TAALES, TAASSC, and TAACO to obtain finer-grained indices of lexical diversity, lexical sophistication, syntactic structure, and cohesion. The resulting outputs were merged with the main file by `essay_id`.

Feature Selection and Statistical Analysis

To avoid treating hundreds of automated outputs as equally informative, 36 interpretable features were retained across five analytical domains: length and organization (5 features), lexical diversity (4), lexical sophistication (9), syntactic complexity (8), and cohesion (10). Feature selection emphasized construct relevance, established indices from the corpus-analysis tools, interpretability, and coverage across domains rather than exhaustive inclusion of all available outputs [8], [9], [12], [14]. The automated indices were treated as textual proxies rather than direct operationalizations of IELTS scoring criteria. Definitions, source tools, and interpretation notes for all 36 features are provided in Additional file 1.

For reproducibility, the external-tool versions were recorded as TAALED 1.4.1 for lexical diversity, TAALES 2.2 for lexical sophistication, TAASSC 1.3.8 for syntactic complexity, and TAACO 2.0.4 for cohesion-related indices. Data screening and statistical preparation were conducted in Google Colab using Python.

Descriptive statistics were calculated for each feature within each band group. Selected results are reported in the main text, while complete descriptive and omnibus inferential results for all 36 features are provided in Additional file 1. Median, standard deviation, minimum, and maximum values were also inspected during data screening to determine whether extreme values or skewed distributions affected interpretation. Inferential analysis was then conducted to test whether the selected features differed across Band 5, Band 6, Band 7, and Band 8. Because several textual features were not normally distributed and variance homogeneity could not be assumed for all variables, Kruskal-Wallis tests were used as the main group-comparison procedure [18].

The analysis involved 36 omnibus Kruskal-Wallis tests across the selected feature set. To reduce the risk of Type I error, Benjamini-Hochberg false discovery rate correction was applied to the 36 omnibus *p*-values before substantive interpretation [19]. When an omnibus test remained statistically significant after correction, pairwise Mann-Whitney *U* tests were conducted, with Holm correction applied within each feature [20]. Effect size was reported using epsilon squared for Kruskal-Wallis tests and rank-biserial correlation for pairwise Mann-Whitney *U* tests. Epsilon-squared values were interpreted using conventional eta-squared benchmarks adapted for non-parametric group comparisons: below 0.01 = negligible, 0.01 to below 0.06 = small, 0.06 to below 0.14 = moderate, and 0.14 or above = large [21]. These thresholds were used as interpretive guides, with priority given to effect magnitude, consistency, and pedagogical interpretability.

Results

This section reports the results in relation to the four research questions. The analysis included 1,912 unique IELTS Writing Task 2 essays: 478 in Band 5, 479 in Band 6, 486 in Band 7, and 469 in Band 8. Kruskal-Wallis tests were followed by Holm-adjusted pairwise Mann-Whitney U tests where relevant [18], [20]. Effect sizes are reported as epsilon squared and interpreted together with statistical significance. After Benjamini-Hochberg FDR correction, 18 of the 36 features remained statistically significant in the deduplicated sample. The clearest statistically detectable features were concentrated in lexical sophistication, adjacent argument overlap, academic connector use, academic vocabulary, and essay length, while most syntactic-complexity and cohesion-related indices remained substantively weak because their epsilon-squared values were negligible. Features with marginal adjusted p values and negligible effect sizes are therefore not treated as substantive band-level markers. Complete band-level descriptive statistics and omnibus results for all 36 features are provided in Additional file 1.

Length and Organization

RQ1 asked how the essays differed across Band 5, Band 6, Band 7, and Band 8 in terms of length and organization. In this domain, academic connectors per 100 words, essay word count, and paragraph count remained statistically significant after FDR correction, whereas average sentence length and sentence count did not. Academic connector use per 100 words showed the clearest length-and-organization difference, $H(3) = 32.47$, raw $p < 0.001$, FDR-adjusted $p < 0.001$, $\epsilon^2 = 0.0154$. Mean connector use increased from Band 5 ($M = 1.92$) and Band 6 ($M = 1.95$) to Band 7 ($M = 2.08$) and Band 8 ($M = 2.17$). Essay word count also differed across groups, $H(3) = 29.51$, raw $p < 0.001$, FDR-adjusted $p < 0.001$, $\epsilon^2 = 0.0139$, with mean word count increasing from Band 5 ($M = 296.95$) and Band 6 ($M = 297.78$) to Band 7 ($M = 303.79$) and Band 8 ($M = 308.23$). However, the sensitivity analysis reported below shows that this word-count effect weakened after essays below 250 words were excluded. Paragraph count remained statistically significant after FDR correction, $H(3) = 15.67$, raw $p = 0.001$, FDR-adjusted $p = 0.004$, but its effect size was negligible, $\epsilon^2 = 0.0066$. Average sentence length did not survive FDR correction, $H(3) = 5.99$, raw $p = 0.112$, FDR-adjusted $p = 0.161$, $\epsilon^2 = 0.0016$, and sentence count showed no meaningful difference, $H(3) = 2.57$, raw $p = 0.463$, FDR-adjusted $p = 0.537$, $\epsilon^2 \approx 0.0000$. Overall, connector use was the most stable feature in this domain, while word count was weak and sensitivity-dependent, and paragraph count, average sentence length, and sentence count were weak or negligible separators.

Lexical Diversity and Lexical Sophistication

RQ2 focused on lexical diversity and lexical sophistication. MTLD, MATTR-50, Kuperman Age of Acquisition, AWL normed proportion, and MRC familiarity all showed statistically significant omnibus differences. However, the effect sizes were small or negligible. Kuperman AoA and AWL normed proportion provided clearer lexical sophistication signals than the lexical diversity indices. This suggests that higher-band essays were associated with slightly later-acquired and more academic vocabulary, but lexical diversity did not function as a strong separator of adjacent bands. Table 2 presents the Kruskal-Wallis results for the selected lexical diversity and lexical sophistication features.

Syntactic Complexity and Cohesion-Related Features

RQ3 examined syntactic complexity and cohesion-related features. Selected syntactic features, including nominal dependencies and prepositions per clause, showed statistically significant differences, but the effect sizes were negligible. Adjacent argument overlap showed a small effect, decreasing from lower to higher band groups. Adjacent content-word overlap also decreased descriptively, but its effect size was negligible. These findings suggest that higher-band essays may rely slightly less on local repetition and slightly more on varied lexical or referential continuity, but the automated cohesion results should not be interpreted as direct evidence of stronger or weaker coherence. Table 3 reports the Kruskal-Wallis results for the selected syntactic complexity and cohesion-related features.

Lower-Higher Band Differences

RQ4 asked which textual features showed the clearest lower-higher band differences. The strongest pairwise contrasts generally appeared between Band 5 and Band 8. Among the pairwise contrasts summarized in Table 4, the clearest features were Kuperman AoA, adjacent argument overlap, AWL normed proportion, and academic connector use. Even so, the rank-biserial correlations remained modest. This pattern suggests that textual features can detect broad lower-higher tendencies, but they are not strong enough to justify a single-feature explanation of band-level writing differences. Table 4 summarizes the four clearest Holm-adjusted pairwise contrasts between lower and higher band groups.

Robustness and Sensitivity Analyses

As an additional robustness check, the four approximate band groups were collapsed into a lower-band group (Band 5–6) and a higher-band group (Band 7–8). This comparison was intended to reduce the influence of noisy adjacent-band labels, because misclassification is more likely between neighboring bands than between broader lower and higher proficiency groupings. The lower-higher comparison

Table 1: Kruskal–Wallis results for length and organization features.

Feature	H(3)	raw p	FDR-adjusted p	ε^2	Effect size
Academic connectors per 100 words	32.47	< 0.001	< 0.001	0.0154	Small
Essay word count	29.51	< 0.001	< 0.001	0.0139	Small
Paragraph count	15.67	0.001	0.004	0.0066	Negligible
Average sentence length	5.99	0.112	0.161	0.0016	Negligible
Sentence count	2.57	0.463	0.537	0.0000	Negligible

Note: FDR = false discovery rate; ε^2 = epsilon squared.

Table 2: Kruskal–Wallis results for selected lexical features.

Feature	H(3)	raw p	FDR-adjusted p	ε^2	Effect size
Kuperman AoA, all words	36.11	< 0.001	< 0.001	0.0174	Small
AWL normed proportion	27.70	< 0.001	< 0.001	0.0129	Small
MRC familiarity, all words	22.28	< 0.001	< 0.001	0.0101	Small
MTLD, all words	22.42	< 0.001	< 0.001	0.0102	Small
MATTR-50, all words	13.84	0.003	0.008	0.0057	Negligible

Note: FDR = false discovery rate; ε^2 = epsilon squared.

Table 3: Kruskal–Wallis results for selected syntactic and cohesion-related features.

Feature	H(3)	raw p	FDR-adjusted p	ε^2	Effect size
Adjacent argument overlap	34.86	< 0.001	< 0.001	0.0167	Small
Nominal dependencies in noun nodes	18.45	< 0.001	0.001	0.0081	Negligible
Prepositions per clause	15.54	0.001	0.004	0.0066	Negligible
Adjacent content-word overlap	20.89	< 0.001	< 0.001	0.0094	Negligible
All logical connectives	11.23	0.011	0.024	0.0043	Negligible

Note: FDR = false discovery rate; ε^2 = epsilon squared.

Table 4: Four clearest Holm-adjusted pairwise contrasts.

Feature	Comparison	Group 1 M	Group 2 M	Holm-adjusted p	$ r_{rb} $
Kuperman AoA, all words	Band 5 vs Band 8	5.788	5.907	< 0.001	0.220
Adjacent argument overlap	Band 5 vs Band 8	0.134	0.112	< 0.001	0.196
AWL normed proportion	Band 5 vs Band 8	0.082	0.091	< 0.001	0.196
Academic connectors per 100 words	Band 5 vs Band 8	1.919	2.170	< 0.001	0.184

Note: M = mean; $|r_{rb}|$ = absolute rank-biserial correlation.

produced the same general profile as the four-band analysis. After FDR correction, 19 of the 36 features remained statistically significant. The clearest lower-higher contrasts appeared in adjacent argument overlap ($M = 0.117$ for Band 7–8 vs. 0.132 for Band 5–6, FDR-adjusted $p < 0.001$, $|r_{rb}| = 0.143$), academic connectors per 100 words ($M = 2.12$ vs. 1.93 , FDR-adjusted $p < 0.001$, $|r_{rb}| = 0.142$), Kuperman AoA ($M =$

5.885 vs. 5.818 , FDR-adjusted $p < 0.001$, $|r_{rb}| = 0.131$), and essay word count ($M = 305.97$ vs. 297.37 , FDR-adjusted $p < 0.001$, $|r_{rb}| = 0.131$). Other significant contrasts, including MRC familiarity, adjacent content-word overlap, nominal dependencies, AWL normed proportion, MTLD, and prepositions per clause, also remained modest in magnitude. This pattern supports the interpretation that the corpus contains detectable

lower–higher textual tendencies, while also confirming that these tendencies do not amount to strong single-feature band separation.

A second sensitivity analysis excluded the 76 essays below the 250-word IELTS Writing Task 2 threshold. After this exclusion, 17 of the 36 features remained statistically significant after FDR correction. The central profile remained largely stable: Kuperman AoA remained small ($\epsilon^2 = 0.0163$), adjacent argument overlap remained small ($\epsilon^2 = 0.0183$), academic connectors per 100 words remained small and became slightly stronger ($\epsilon^2 = 0.0215$), and AWL normed proportion remained small ($\epsilon^2 = 0.0118$). However, essay word count weakened from a small effect in the full sample ($\epsilon^2 = 0.0139$) to a negligible effect in the 250-word-plus sample ($\epsilon^2 = 0.0045$). HD-D, which was significant in the full deduplicated sample, no longer remained significant after FDR correction. Therefore, the inclusion of below-threshold essays was not the main driver of the lexical sophistication, connector-use, or adjacent-overlap findings, but it did partly contribute to the apparent word-count contrast. This strengthens the central interpretation that length-related differences should not be treated as strong evidence of band separation.

Discussion

The central finding is not that higher approximate band groups differed from lower groups, but that most textual differences were modest despite a large sample. Lexical sophistication, academic vocabulary, connector use, adjacent overlap, and essay length showed detectable variation. However, the effects were generally small and clearer in broader lower–higher contrasts than between adjacent bands. This pattern weakens any interpretation of single textual features as reliable indicators of band-group separation and instead supports a multidimensional, effect-size-aware account.

A central interpretive issue is the reliability of the approximate band labels. Rating noise may attenuate adjacent-band contrasts. Conversely, the possible presence of model essays, heavily revised texts, or AI-assisted essays may inflate some textual differences. The lower–higher analysis was therefore used to examine whether broader contrasts were more stable. Accordingly, the findings should be interpreted as a cautious textual profile of a public essay dataset with human-sourced approximate band labels, not as a direct estimate of official IELTS band separation.

This interpretation is consistent with the IELTS Writing Task 2 construct. The band descriptors evaluate task response, coherence and cohesion, lexical resource, and grammatical range and accuracy [2]. They do not suggest that a score can be obtained by increasing word count, inserting more connectors, or adding rare words. The present results reinforce this construct-aware view, but the robustness checks refine the interpretation. Academic connector use was

the most stable length-and-organization feature. It remained significant after FDR correction and was clear in the lower–higher comparison. Its effect was also not weakened by excluding essays below 250 words. Even so, its effect size was small, so connector frequency should not be interpreted as a strong band marker or as evidence that learners should simply insert more linking expressions. Essay word count showed a small full-sample effect and a significant lower–higher contrast. However, its effect size became negligible after essays below the 250-word threshold were removed. This means that word count is better understood as a minimum condition for task fulfillment and idea development than as a reliable indicator of higher-band writing quality. Paragraph count, average sentence length, and sentence count were weak or negligible signals. In particular, average sentence length did not survive FDR correction in the four-band analysis, which weakens any formulaic advice that higher-band IELTS writing can be achieved by writing longer sentences.

The lexical findings also require a balanced interpretation. Higher-band essays showed clearer patterns in lexical sophistication and academic vocabulary than in lexical diversity. This pattern suggests that lexical differences across band groups may relate more to appropriate and, where necessary, academic vocabulary than to lexical diversity alone. This finding aligns with Nguyen et al. [5], who treated lexical resource as multidimensional rather than as a single vocabulary measure. However, the small effect sizes also show that lexical sophistication alone cannot function as a reliable band marker, especially when the band labels are approximate rather than official examiner ratings.

The syntactic and cohesion findings further challenge simple assumptions about how higher-band writing differs from lower-band writing. In this dataset, syntactic-complexity indices were statistically detectable in some cases but substantively weak, with negligible effect sizes. This means that the results do not support a pedagogical claim that higher-band IELTS essays are clearly distinguished by longer or more complex sentence structures alone. If anything, the syntactic results suggest that broad sentence-length measures and selected fine-grained syntactic indices are insufficient as stand-alone band markers. This interpretation is compatible with earlier arguments that advanced academic writing often involves controlled phrasal and nominal packaging rather than a simple increase in sentence length or clausal subordination [11]. Similarly, lower adjacent overlap in higher-band essays should not be read mechanically as weaker cohesion. It may reflect reduced local repetition and greater use of paraphrase, synonymy, referencing, or lexical variation. Automated overlap indices therefore need to be interpreted in relation to discourse function rather than treated as direct proxies for coherence.

The main pedagogical implication is that IELTS writing instruction should move beyond single-feature

targets. Teachers should not tell learners simply to write longer essays, use harder words, insert more linking expressions, or produce longer sentences. Instead, instruction should help learners develop a balanced textual profile. This includes sufficient idea development, appropriate paragraphing, flexible lexical control, academic vocabulary where relevant, controlled syntactic packaging, and cohesion that supports argument progression. Word count should be treated as a minimum condition for task fulfillment rather than evidence of quality by itself. Feedback should instead focus on whether connectors, lexical repetition, and syntactic choices support communicative effectiveness.

Conclusion

This study investigated how IELTS Writing Task 2 learner essays differed across four approximate band groups in a deduplicated corpus of 1,912 essays. By combining Python-based text features with indices from TAALED, TAALES, TAASSC, and TAACO, the study provided a multidimensional profile of textual variation across human-sourced approximate band groups.

The findings indicate that higher-band essays were not distinguished by one dominant feature. Instead, band-level differences appeared across several dimensions, including essay length, academic connector use, lexical sophistication, academic vocabulary, lexical diversity, selected syntactic-complexity features, and cohesion-related overlap indices. However, the effect sizes were generally small, so these differences should be interpreted as modest textual tendencies rather than strong markers of band separation.

The robustness checks reinforced this interpretation. Although 18 of 36 features remained significant after FDR correction, only seven reached the small effect-size range and none were moderate or large. Broader lower-higher contrasts were clearer but still modest, while excluding essays below 250 words weakened the word-count effect but left connector use, Kuperman AoA, AWL proportion, and adjacent argument overlap relatively stable. These patterns argue against surface-level prescriptions for IELTS Writing Task 2.

Several limitations define the scope of these findings. The dataset contains human-sourced approximate band scores rather than official IELTS examiner ratings, and its metadata do not permit verification of rating procedures, inter-rater reliability, writer background, testing conditions, or essay-level provenance. Noisy labels may attenuate differences, while model, heavily edited, or AI-assisted texts may inflate them. FDR correction, lower-higher comparisons, and the 250-word sensitivity analysis reduce but do not remove these concerns. Prompt assignment was uncontrolled and prompt was not modeled as an independent factor; some indices may also remain length-sensitive; the results should therefore be read as exploratory

textual tendencies rather than evidence about official IELTS score assignment.

Within these boundaries, the study shows that detectable differences across approximate band groups are modest and distributed rather than concentrated in one feature. Future research should use officially rated and authenticated scripts, report inter-rater reliability, control prompt effects, examine possible AI-assisted or model-essay contamination, and combine automated indices with qualitative analysis of argument quality, coherence, lexical appropriacy, and grammatical control.

Acknowledgements

The authors would like to thank the creator of the chillies/IELTS-writing-task-2-evaluation dataset for making the data publicly available. The authors also acknowledge the developers of TAALED, TAALES, TAASSC, and TAACO, whose tools supported the corpus-based analyses reported in this study.

Ethics Approval and Consent to Participate

This study was a secondary analysis of a publicly accessible, pre-existing textual dataset. No participants were recruited or contacted, and the authors did not collect new identifiable personal data. No institutional ethics approval was obtained for this secondary analysis. The source dataset did not provide writer identifiers or sufficient metadata to verify the original consent or ethics-review status of the essays. All analyses were conducted at aggregate level, and no attempt was made to identify individual writers or reconstruct essay provenance.

Funding

No external funding was received for this research.

Competing Interests

The authors declare no competing interests.

Additional File

Additional File 1 is available at: <https://doi.org/10.5281/zenodo.22810095>

References

- [1] IELTS, "IELTS Academic Writing sample tasks," British Council, IDP IELTS, and Cambridge University Press & Assessment, 2023. [Online]. Available: <https://ielts.org/cdn/Sample-tests/ielts-academic-writing-sample-tasks-2023.pdf>. [Accessed: Jul. 18, 2026].

- [2] IELTS, "Writing Task 2 band descriptors," British Council, IDP IELTS, and Cambridge University Press & Assessment, 2023. [Online]. Available: https://takeielts.britishcouncil.org/sites/default/files/ielts_writing_band_descriptors.pdf. [Accessed: Jul. 24, 2026].
- [3] A. M. Riazi and J. S. Knox, "An investigation of the relations between test-takers' first language and the discourse of written performance on the IELTS Academic Writing Test, Task 2," *IELTS Research Reports Online Series*, 2013. [Online]. Available: <https://ielts.org/researchers/our-research/research-reports/an-investigation-of-the-relations-between-test-takers-first-language-and-the-discourse-of-written-performance-on-the-ielts-academic-writing-test-task-2>. [Accessed: Aug. 2, 2026].
- [4] K. Barkaoui, "What changes and what doesn't? An examination of changes in the linguistic characteristics of IELTS repeaters' Writing Task 2 scripts," *IELTS Research Report Series*, no. 3, IELTS, 2016. [Online]. Available: <https://ielts.org/cdn/Research/what-changes-and-what-doesnt-barkaoui-2016.pdf>. [Accessed: Aug. 7, 2026].
- [5] B. T. T. Nguyen, T. T. B. Nguyen, and N. Q. Phan, "Vietnamese IELTS learners' perceptions of IELTS Writing Task 2 and their performance-based lexical resource," *IELTS Research Reports Online Series*, no. 5/22, IELTS, 2022. [Online]. Available: <https://ielts.org/cdn/Research/vietnamese-ielts-learners-perceptions-of-ielts-writing-task-2-nguyen-et-al-2022.pdf>. [Accessed: Aug. 12, 2026].
- [6] S. A. Crossley, "Linguistic features in writing quality and development: An overview," *Journal of Writing Research*, vol. 11, no. 3, pp. 415–443, 2020, doi: [10.17239/jowr-2020.11.03.01](https://doi.org/10.17239/jowr-2020.11.03.01).
- [7] S. Granger, J. Hung, and S. Petch-Tyson, Eds., *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam, The Netherlands: John Benjamins, 2002, doi: [10.1075/llt.6](https://doi.org/10.1075/llt.6).
- [8] K. Kyle, S. A. Crossley, and S. Jarvis, "Assessing the validity of lexical diversity indices using direct judgements," *Language Assessment Quarterly*, vol. 18, no. 2, pp. 154–170, 2021, doi: [10.1080/15434303.2020.1844205](https://doi.org/10.1080/15434303.2020.1844205).
- [9] K. Kyle, S. A. Crossley, and C. M. Berger, "The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0," *Behavior Research Methods*, vol. 50, no. 3, pp. 1030–1046, 2018, doi: [10.3758/s13428-017-0924-4](https://doi.org/10.3758/s13428-017-0924-4).
- [10] L. Ortega, "Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing," *Applied Linguistics*, vol. 24, no. 4, pp. 492–518, 2003, doi: [10.1093/applin/24.4.492](https://doi.org/10.1093/applin/24.4.492).
- [11] D. Biber, B. Gray, and K. Poonpon, "Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?" *TESOL Quarterly*, vol. 45, no. 1, pp. 5–35, 2011, doi: [10.5054/tq.2011.244483](https://doi.org/10.5054/tq.2011.244483).
- [12] S. A. Crossley, K. Kyle, and D. S. McNamara, "The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion," *Behavior Research Methods*, vol. 48, no. 4, pp. 1227–1237, 2016, doi: [10.3758/s13428-015-0651-7](https://doi.org/10.3758/s13428-015-0651-7).
- [13] A. Saadatara, G. R. Kiany, and H. Talebzadeh, "Bundles to beat the band in high-stakes tests: Pedagogical applications of an exploratory investigation of lexical bundles across band scores of the IELTS writing component," *Journal of English for Academic Purposes*, vol. 61, Art. no. 101208, 2023, doi: [10.1016/j.jeap.2022.101208](https://doi.org/10.1016/j.jeap.2022.101208).
- [14] S. Kim, P. Williams, and L. McCallum, "Modelling the use of the tool for the automatic analysis of syntactic sophistication and complexity (TAASSC)," *Research Methods in Applied Linguistics*, vol. 3, no. 1, Art. no. 100087, 2024, doi: [10.1016/j.rmal.2023.100087](https://doi.org/10.1016/j.rmal.2023.100087).
- [15] chillies, "IELTS-writing-task-2-evaluation" [Data set], Hugging Face, 2024. [Online]. Available: <https://huggingface.co/datasets/chillies/IELTS-writing-task-2-evaluation>. [Accessed: Aug. 18, 2026].
- [16] chillies, "Comment on 'Synthetic generated data'" [Discussion comment], Hugging Face, Aug. 30, 2025. [Online]. Available: <https://huggingface.co/datasets/chillies/IELTS-writing-task-2-evaluation/discussions/7>. [Accessed: Aug. 21, 2026].
- [17] chillies, "Dataset permissions" [Discussion comment], Hugging Face, May 22, 2025. [Online]. Available: <https://huggingface.co/datasets/chillies/IELTS-writing-task-2-evaluation/discussions/6>. [Accessed: Aug. 25, 2026].
- [18] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952, doi: [10.1080/01621459.1952.10483441](https://doi.org/10.1080/01621459.1952.10483441).
- [19] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995, doi: [10.1111/j.2517-6161.1995.tb02031.x](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x).
- [20] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. [Online]. Available: <https://www.jstor.org/stable/4615733>
- [21] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1988.