Text Quality and Consistency of G-TELP™ Level 2 Reading Passages

2025

ITSC Research Series

Report ITSC-2025-01

Text Quality and Consistency of G-TELP™ Level 2 Reading Passages

ITSC Research Series

Report ITSC-2025-01

Prepared by Karl Sarvestani, PhD

Abstract

This study examines the consistency of four key linguistic features—readability, lexical diversity, syntactic sophistication, and global textual cohesion—across 144 reading passages from the G-TELP Level 2 reading test. The analysis employs automated text processing tools to assess the alignment of these features with the intended proficiency level of the test (CEFR B2–C2). Results indicate that while lexical diversity and global cohesion demonstrate consistent patterns across genres, other features such as syntactic sophistication exhibit more variability. Notably, magazine article passages show a bimodal distribution in readability scores, and business letters contain shorter clauses overall, affecting both readability and syntactic complexity scores. Notwithstanding these variations, findings suggest that the passages largely align with the expected linguistic features for the intended level of the assessment. This research highlights how consistency across multiple iterations of a proficiency test can be achieved to ensure fairness in testing and will assist in refining development practices among G-TELP writers and editors. Potential avenues for future research are discussed, including incorporating test-taker performance data and exploring genre-specific effects on text quality.

Background

The design of reading comprehension sections within English proficiency exams poses a formidable challenge for assessment companies. A crucial concern in the development of such exams is ensuring that the linguistic features of reading passages are consistent and appropriately matched to the intended proficiency levels of the candidates. To this end, this report examines the consistency of several key linguistic features across a set of reading passages for the Level 2 exam of the General Test of English Language Proficiency

(G-TELP). This reading exam is composed of four parts, each of which represents a different genre of text. Therefore, the effects of text type on these linguistic features are also examined. Specifically, this investigation focuses on four aspects of the passages: readability, lexical diversity, syntactic sophistication, and global textual cohesion. These elements are essential in determining the quality of the reading materials, which in turn affects how accurately the exam can assess test takers' language abilities (Crossley, 2020; McNamara, 2010).

G-TELP Level 2 reading passages undergo a rigorous development process involving several rounds of drafting, editing, and review before they are passed along to question writers, sent for external review, and finalized and released for use in testing sessions. The level of success these efforts have in ensuring text quality and consistency can be explored by examining quantifiable linguistic features that characterize text quality, including readability, lexical and syntactic properties, and discourse properties of text cohesiveness. By examining these linguistic features, we can assess the effectiveness of the development process in producing high-quality reading passages that better ensure reliable and consistent measurement of test-taker abilities.

Readability refers to the ease with which a passage can be understood, typically influenced by factors such as sentence length, word complexity, and overall structure. It is important for a text to be suited to the appropriate reading level of its readers to provide sufficient challenge and engagement. Lexical diversity measures the range of vocabulary used in a passage and is important for evaluating both the breadth and depth of a candidate’s vocabulary knowledge. Syntactic sophistication pertains to the complexity of sentence structures, which reflects the grammatical and syntactical proficiency required for understanding the text. Finally, global textual cohesion involves the coherence and logical flow of ideas within a passage, which is critical for evaluating how well a candidate can grasp the overarching meaning and relationships between concepts in the text. All of these elements can contribute to high-quality reading passages.

A strong relationship has consistently been found between the lexical properties of a text and reading comprehension (e.g., Wright & Cevetti, 2017), with readers' vocabulary coverage of the words in a text being a primary factor. More specifically, second language readers are best able to understand text when they recognize 95–98% of the vocabulary used (Laufer & Ravenhorst-Kalovski, 2010). Prior research has measured the lexical sophistication of reading passages based on the presence of specific word types, including academic (Douglas, 2013), low frequency (McNamara et al., 2010), multisyllabic, and unfamiliar (Crossley et al., 2011) as well as on the presence of specificity, meaning, and imagery (McNamara et al., 2013). Quantifiable linguistic features that correlate with these measures of lexical sophistication include word frequency, word range, n-gram frequency, and psycholinguistic word properties as well as indices related to word recognition norms, contextual distinctiveness, word neighborhood, semantic network, n-gram range, and n-gram strength of association (Kyle et al., 2018).

Regarding syntactic sophistication, a meta-analysis conducted by Jeon & Yamashita (2014) found that readers' grammatical knowledge correlates even more strongly with reading comprehension than lexical knowledge. Thus, both lexical and syntactic features of a text must be considered when gauging its suitability for use in language assessment. Classic approaches to quantifying syntactic complexity have often relied on T-Units (Hunt, 1965). A T-Unit is a main clause plus any subordinate clauses that may be attached to it. Other features of syntactic complexity that correlate with superior writing quality include the limited use of finite verbs, finite subordinates, and coordinate clauses (Myhill, 2008), more words before the main verb (McNamara et al. 2010), and the use of fewer simple declarative sentences and longer noun phrases.

Another measure of text quality explored in prior research is textual cohesiveness, referring to the extent to which paragraphs in a text are linked to each other (Delu & Rushan, 2021; Halliday & Hasan, 1976), and research indicates a clear link between global cohesive devices and text quality (Neuner, 1987; Tabari & Johnson, 2023). Modern indices of global textual cohesion tend to focus on measuring lexical and semantic overlaps across paragraphs (e.g., Crossley et al., 2011b; McNamara et al., 2013).

The objective of this study is to assess these linguistic features and investigate their consistency across various genres of reading passages using quantitative analysis. In doing so, this research aims to provide insight into how well the passages align with the intended proficiency levels for the exam and to identify potential areas where the linguistic features may be misaligned, thus affecting the fairness and effectiveness of the assessment. This analysis also contributes to the broader field of language testing by providing valuable evidence that can be used for improving the design and development of reading comprehension sections on the G-TELP Level 2 and other English proficiency exams.

Method and Materials

The materials for this study include thirty-six sets of reading passages from retired G-TELP Level 2 tests used for assessment sessions between August 2021 and December 2024. Level 2 has been selected for this study because it is the most popular level of the exam; thus, more reading passages for this level exist and are readily available, and the results of this analysis will be most enlightening. Each set contains four passages, 275–390 words each, that are specifically created for the exam: a biography article, a magazine article, an encyclopedia article, and a business letter, yielding 144 total passages for analysis.

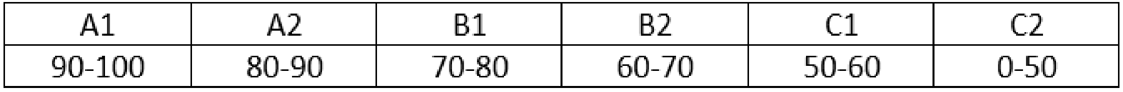
Automated text processing was performed using a suite of freely available tools. Readability was assessed using Flesch Reading Ease and Flesch-Kincaid Grade Level measures generated by ARTE (described in Choi & Crossley, 2022); lexical diversity was measured by vocabulary level according to the Common European Framework of Reference (CEFR) and also by type-token ration (TTR) values generated using TAALED (Kyle et al. 2021). TAASSC (Kyle, 2016) was used to measure the mean length of T-Units as an index of syntactic sophistication. Finally, TAACO (Crossley et al., 2019) was used to generate Word2Vec similarity scores in order to assess global cohesiveness.

Results and Discussion

The Flesch Reading Ease index values for all passages fall between 34.2 and 72.2, without any statistical outliers (i.e., values greater than three standard deviations from the group mean), and with a clear modal range of 50–60 for all genres. As can be seen in Table 1 below, this puts each genre and the passages as a whole within the C1–C2 range of the CEFR (Natova, 2019; Cherian & Jha, 2024), which is the mid- to upper end of the intended range for G-TELP Level 2 reading passages. The distributions for the encyclopedia article and business letter passages are approximately normal, as are those for the biography passages, with the exception of a spike at the lower end of the scale. A similar, but more pronounced pattern is also observed in the magazine article data. Although the mean for the magazine articles falls within the expected range, this appears to be due to an unexpected bimodal distribution with peaks above and below the mean at 40–45 and 60–65. This unexpected distribution is also borne out in a relatively higher standard deviation of 9.7 compared to an average of 7.4 for the non-magazine article genres.

Table 1

CEFR Vocabulary level by Flesch Reading Ease score

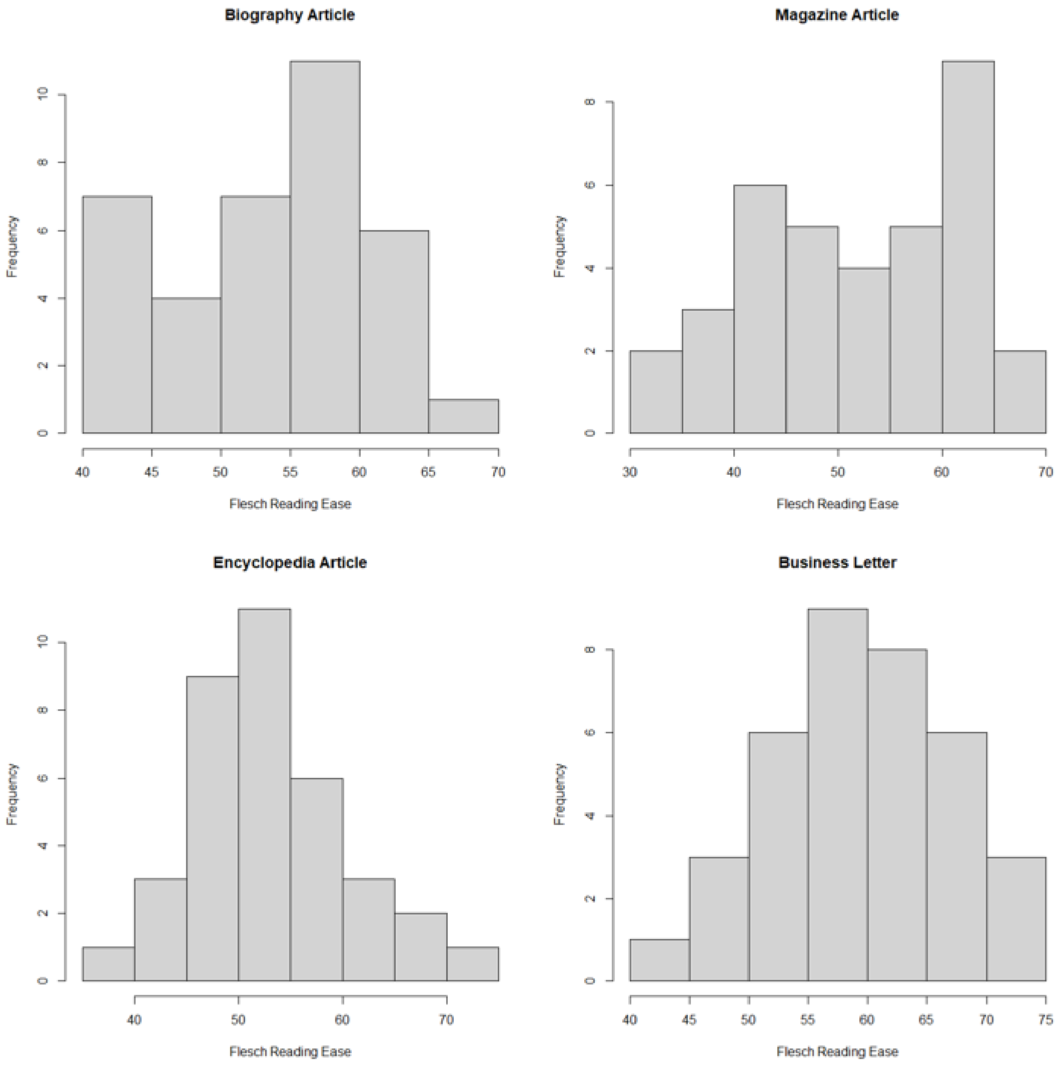


A1	A2	B1	B2	C1	C2
90-100	80-90	70-80	60-70	50-60	0-50

It is worth noting here that the guidance given to G-TELP exam writers indicates that magazine articles are intended to be slightly more conversational in tone, thus using less formal language, than biography articles. This difference in guidelines may provide at least a partial explanation for the peak around 60–65 observed in the Flesch Reading Ease data (see Figure 1 below). Since Reading Ease and Flesch-Kincaid Grade Level are roughly inversely correlated—higher Reading Ease scores tend to imply a lower Grade Level—this particular Reading Ease peak likely corresponds to the peak around 9–10 in the Flesch-Kincaid Grade Level data. The second peak then, occurring at 40–45 in the Reading Ease graph, may be the result of difficulty consistently adhering to this guidance.

Figure 1

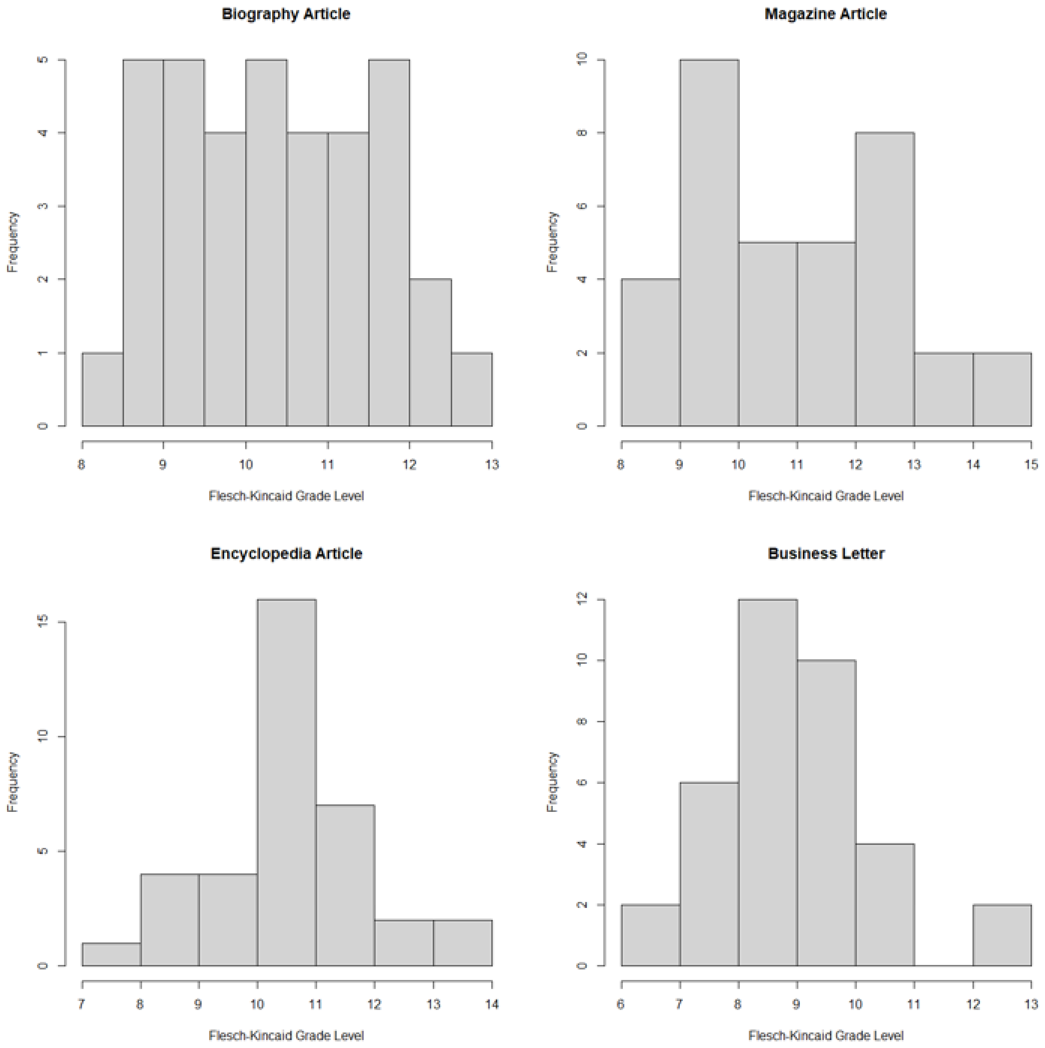
Flesch Reading Ease



The Flesch-Kincaid Grade Level measurements for all passages fall between 6.7 and 14.4 without any statistical outliers, and the mean across all genres is 10.2 with a 1.4 standard deviation. The same basic pattern observed for the Flesch Reading Ease data is also observed in these results (see Figure 2 below): the encyclopedia articles and business letters are again approximately normal, while the magazine articles again exhibit a bimodal distribution, with peaks above and below the all-passage means and an accompanying standard deviation approximately 30% larger than the average of the non-magazine articles. Although business letters have a relatively normal distribution, they also have a lower mean grade level at 8.9 compared to 10.6 for the non-business letter genres. This appears to be due at least in part to how the grade level is calculated, as will be discussed in more detail in the syntactic complexity section below. Regarding grade level, the biography articles have a broader profile than the other genres, with over 40% of passages falling greater than 1 standard deviation away from the mean value.

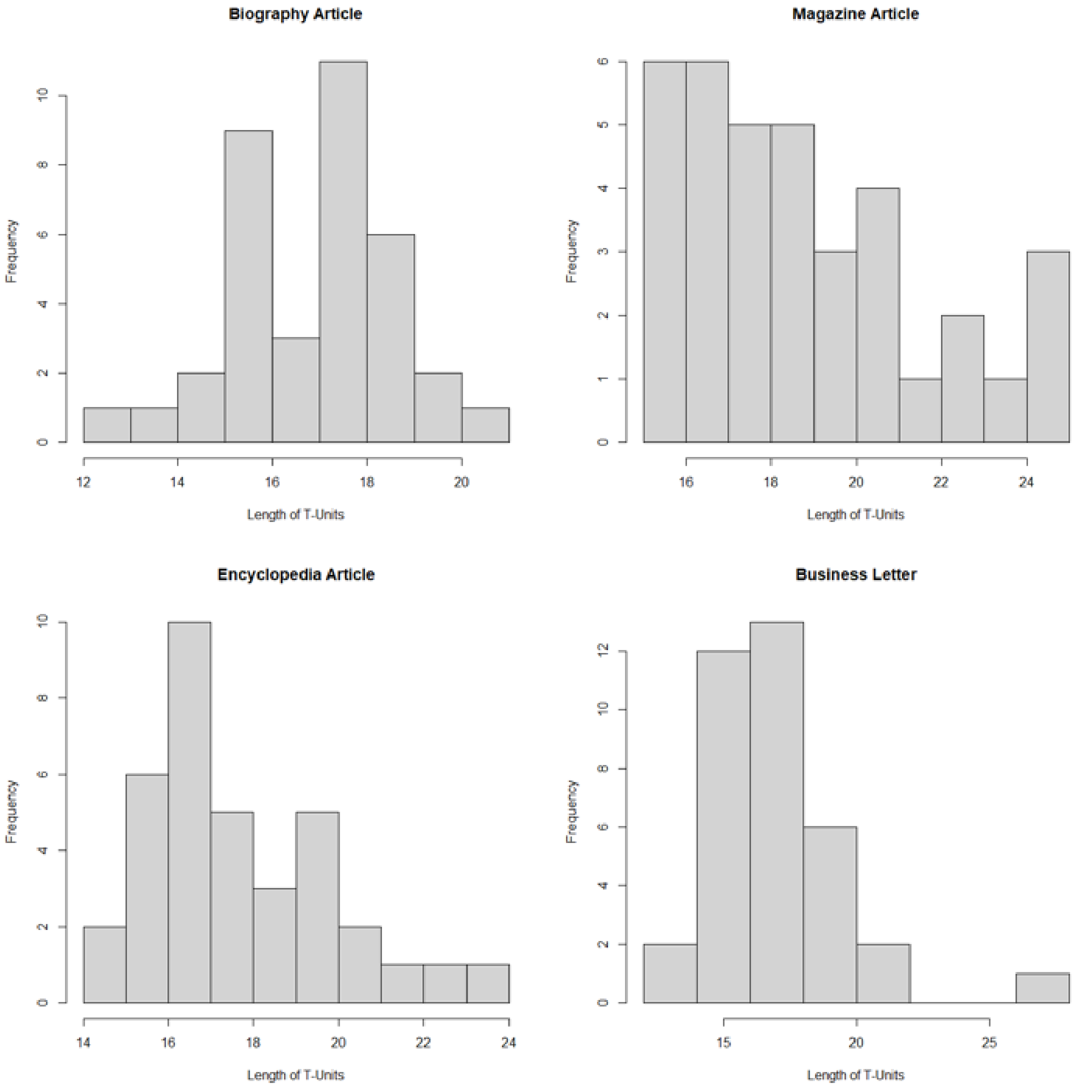
Figure 2

Flesch-Kincaid Grade Level



As shown in Figure 3, the length of T-Units across all passages fell between 12.6 and 26.5, with an all-genre mean of 17.6. A single, high-end outlier was observed in each of two groups: the magazine article and business letter passages. Examining the mean length of T-Units for each of these genres helps to clarify why the business letters displayed relatively higher reading ease and lower grade level in the readability statistics above. While the readability statistics were calculated based on sentence length rather than T-Units strictly, T-Units tend to be complete sentences; thus, the fact that business letters have, on average, T-Units shorter than the non-business letters—16.7 versus 17.9—seems to have also affected their readability statistics. However, unlike the readability results, none of these distributions are particularly normal, with the magazine passages being the most skewed toward the lower end of their range. Nevertheless, the mean length for all genres of passages falls consistently between 16 and 19.

Figure 3
Length of T-Units



The most striking pattern in the lexical diversity data is that the distributions are fairly normal across all genres, with narrowly clustered peaks falling between 0.53 and 0.55, and small, essentially identical standard deviations of 0.03 (see Figure 4). No outliers were observed and, indeed, all passages fall within a relatively narrow type-token ratio (TTR) range of 0.45 to 0.62. Of the linguistic features examined here, lexical diversity is easily the clearest marker of textual consistency observed across genres and across passages as a whole.

Figure 4

Type-Token Ratio

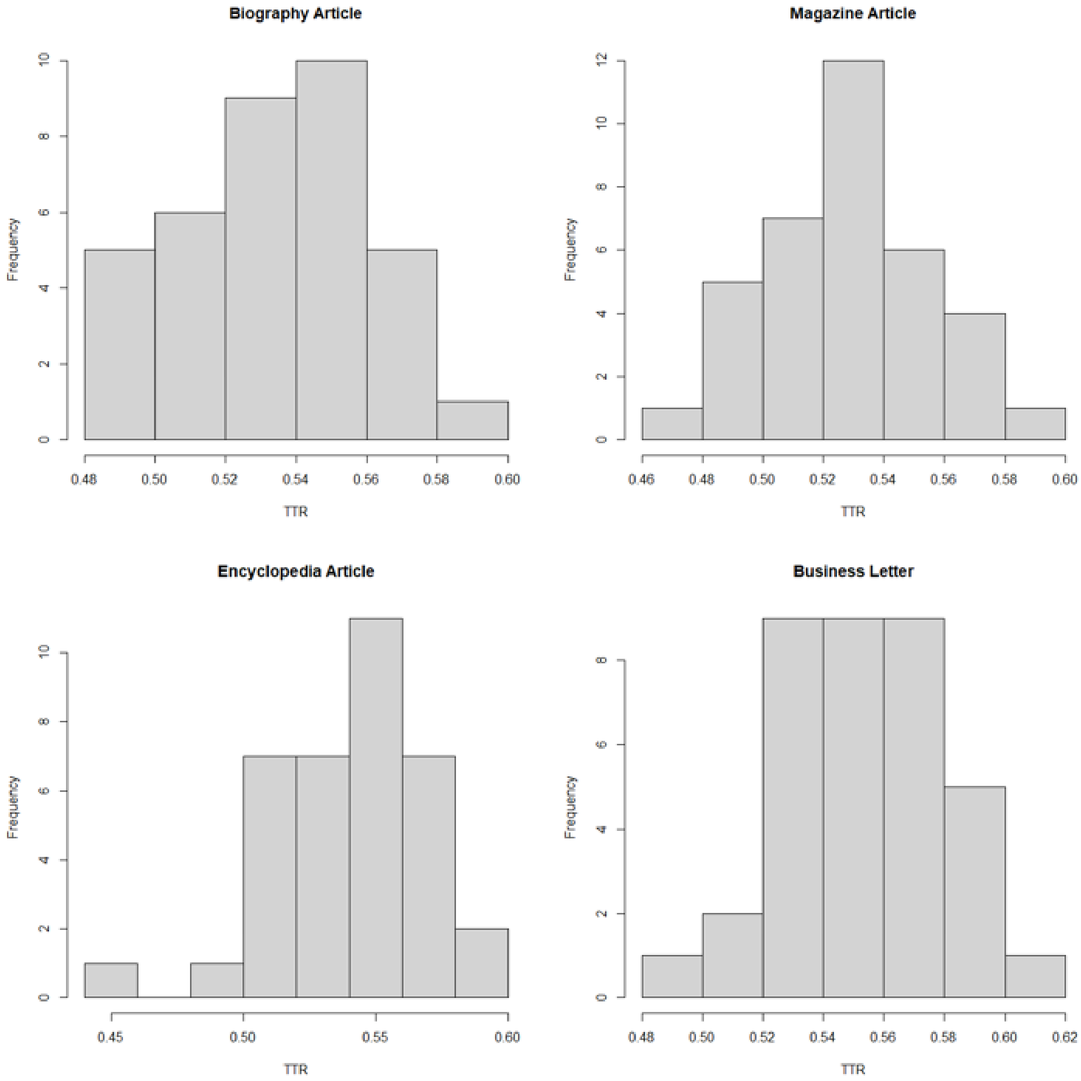
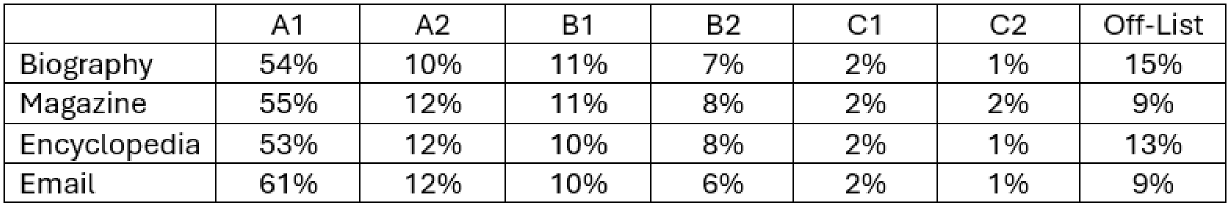


Table 2 below displays what percentage of vocabulary within these passages falls at each CEFR vocabulary level, divided by genre. Only 1-2% of the vocabulary in any of the passages rises above the C1 level, which aligns with the intended vocabulary level range of the G-TELP Level 2. While these profiles look much the same across passage types, a notable exception is that business letter passages have a higher proportion of words at the A1 level and, correspondingly, a relatively low proportion of off-list vocabulary, which would include proper nouns (Owen et al., 2021). This discrepancy is intentional on the part of G-TELP writers, as business letters are the shortest, simplest passages of the four parts. They are also the only passage of the four to include a fictional scenario, absent of high-level vocabulary that is often present in nonfiction texts. Nevertheless, the business letters themselves maintain consistency across multiple test iterations.

Table 2

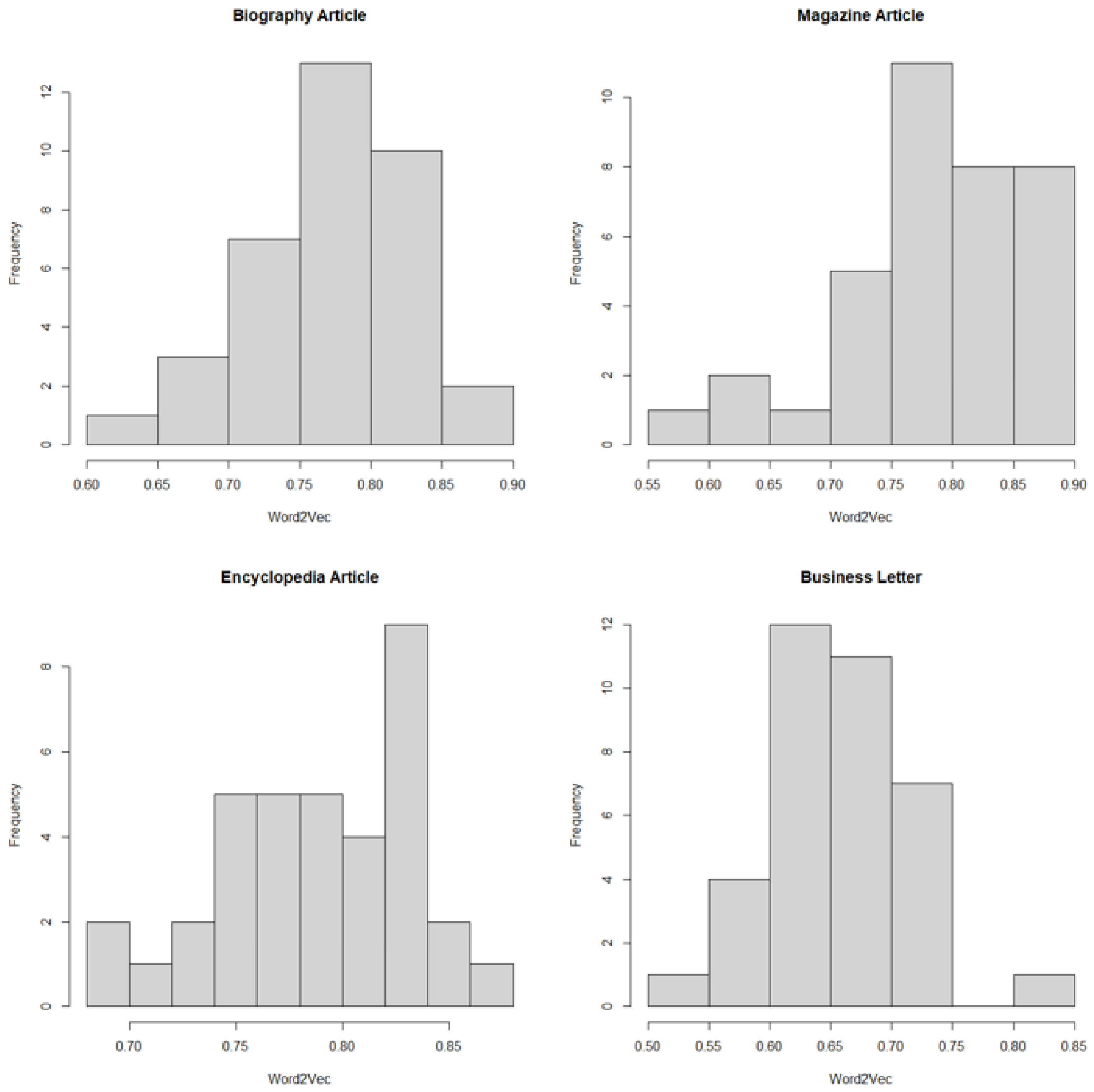
Vocabulary Distribution



	A1	A2	B1	B2	C1	C2	Off-List
Biography	54%	10%	11%	7%	2%	1%	15%
Magazine	55%	12%	11%	8%	2%	2%	9%
Encyclopedia	53%	12%	10%	8%	2%	1%	13%
Email	61%	12%	10%	6%	2%	1%	9%

Text global cohesion measured by Word2Vec similarity (see Figure 5) within an adjacency span of two paragraphs reveals that similarity ratings for all genres fall between 0.51 and 0.91, with an all-genre average of 0.75 and a standard deviation of 0.06. A single, low-end outlier occurred in the business letter data.

Figure 5
Word2Vec Similarity


Biography Article

Magazine Article

Encyclopedia Article

Business Letter

The modal values for biography and magazine articles fall near the all-genre mean, but only the biography articles approximate normal distribution, with the magazine articles being skewed toward higher values of similarity. The encyclopedia articles have an unexpected spike around 0.83, which is reflected in the somewhat higher group mean. The business letter passages have a lower value of similarity, which could again be due to the general brevity of business letters as a genre, resulting in shorter paragraphs. Due to the way that the Word2Vec algorithm operationalizes similarity—based on lexical similarity between paragraphs—shorter paragraphs typically correlate with lower scores.

Conclusion

This study examined the consistency of key linguistic features—readability, lexical diversity, syntactic sophistication, and global textual cohesion—across 144 reading passages from the G-TELP Level 2 reading exam, contributing to the broader understanding of text quality in language proficiency assessments. The analysis revealed that while certain aspects of the passages, such as lexical diversity and global cohesion, demonstrated consistent patterns across genres, other features showed more variability. Despite this variation across text genres, there was substantial consistency among the passages within each genre.

The Flesch-Kincaid Grade Level and Flesch Reading Ease indices largely aligned with the intended proficiency level of the reading passages, though the magazine articles presented an unexpected bimodal distribution for readability, which could signal a misalignment in complexity. Lexical diversity exhibited the highest consistency across all genres, reinforcing the reliability of vocabulary measures in these passages. However, syntactic sophistication, particularly the length of T-Units, varied across genres, with the business letters unsurprisingly exhibiting shorter T-Units and, consequently, lower readability scores. The global cohesion analysis showed that while biography articles and magazine articles had similarity measures close to the overall average, business letter passages demonstrated lower cohesion, likely due to their shorter paragraph structures. Taken as a whole, these results indicate that G-TELP Level 2 reading passages examined here are highly consistent in terms of linguistic features associated with lexical diversity, lexical sophistication, syntactic complexity, and global cohesion. These are elements associated with high-quality text, indicating that the test development processes used for the G-TELP Level 2 reading test consistently produce high-quality content. Results also highlight areas for writers and editors to target to make the passages even more consistent.

These findings illustrate how consistency of linguistic features can be measured to assess the quality of texts used in reading exams. While, overall, the linguistic features of the G-TELP Level 2 reading passages are largely consistent with expected CEFR levels, the variability observed in certain genres—especially magazine article and business letter passages—raises important questions about genre-specific text complexity and its impact on assessment outcomes.

Despite the insights provided, this study has its limitations. First, the analysis focused solely on G-TELP Level 2 reading passages, which means these findings may not generalize to other levels of the exam or other English proficiency exams. It does, however, demonstrate how one might approach an investigation of passage consistency. Additionally, the study relied on relatively coarse-grained data gleaned from automated text analysis tools, which, while efficient, may not capture finer nuances of text quality or context. Human judgments on factors such as textual coherence, the appropriateness of complexity for the target audience, and the interpretability of certain syntactic constructions may offer valuable insights that quantitative measures cannot fully replicate (McNamara et al. 2010).

A natural extension of this research paradigm would be to examine the text quality and consistency not only of reading passages, but also of G-TELP listening scripts. An additional venue of inquiry would be to look beyond an investigation of the consistent results of current test development practices to explore how those processes have improved the quality of text since their adoption. Such a study could easily expand to include not only the small handful of linguistic features examined here, but also the hundreds of other more sophisticated indices that can be measured by the linguistic analysis tools used in this present study. Another direction for future exploration could be to incorporate test-taker performance data, examining how the linguistic features discussed here correlate with actual reading comprehension score outcomes. This would help further validate the impact of readability, lexical diversity, syntactic sophistication, and global cohesion on test-takers' ability to comprehend and engage with the texts.

In conclusion, this study offers a valuable quantitative assessment of linguistic features in the G-TELP Level 2 reading passages. Further research is needed to refine our understanding of how these features interact with test-taker performance and to ensure that English proficiency exams are effective in assessing candidates' language abilities.

References

Cherian, G. & Jha, S. K. (2024). Impact of Readability and CEFR Levels on EFL Materials. *Journal of Emerging Technologies and Innovative Research, 11*(6), 44–52.

Choi, J. S., & Crossley, S. A. (2022). Advances in readability research: A new readability Web app for English. In *2022 International Conference on Advanced Learning Technologies (ICALT)* (pp. 1–5). IEEE. https://doi.org/10.1109/icalt55010.2022.00007

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*, 14–27. https://doi.org/10.3758/s13428-018-1142-4

Crossley, S. A., Roscoe, R., & McNamara, D. S. (2011). Predicting human scores of essay quality using computational indices of linguistic and textual features. *In Artificial Intelligence in Education: 15th International Conference, AIED 2011*, Auckland, New Zealand, June 28–July 2011 15 (pp. 438–440). Springer. https://doi.org/10.1007/978-3-642-21869-9_62

Crossley, S. A., Weston, J. L., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, *28*(3), 282–311. https://doi.org/10.1177/0741088311410188

Delu, Z., & Rushan, L. (2021). *New research on cohesion and coherence in linguistics*. Routledge. https://doi.org/10.4324/9781003190110-11-15

Douglas, S. R. (2013). The lexical breadth of undergraduate novice level writing competency. *Canadian Journal of Applied Linguistics*, *16*(1), 152–170.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman. https://doi.org/10.4324/9781315836010

Hunt, K. W. (1965). *Grammatical structures written at three grade levels.* NCTE Research Report No. 3. National Council of Teachers of English.

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*(1), 160–212. https://doi.org/10.1075/bpa.13.03jeo

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

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*, 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1080/15434303.2020.1844205

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1), 57–86. https://doi.org/10.1177/0741088309351547

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*, 499–515. https://doi.org/10.3758/s13428-012-0258-1

Mesmer, H. A., Cunningham, J. W., & Hiebert, E. H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading Research Quarterly*, *47*(3), 235–258. https://doi.org/10.1002/rrq.019

Myhill, D. (2008). Towards a linguistic model of sentence development in writing. *Language and Education*, *22*(5), 271–288. https://doi.org/10.2167/le775.0

Natova, I. (2021). Estimating CEFR reading comprehension text complexity. *The Language Learning Journal*, *49*(6), 699–710. https://doi.org/10.1080/09571736.2019.1665088

Neuner, J. L. (1987). Cohesive ties and chains in good and poor freshman essays. *Research in the Teaching of English*, *21*(1), 92–105. https://doi.org/10.58680/rte198715592

Owen, N., Shrestha, P., & Bax, S. (2021). *Researching lexical thresholds and lexical profiles across the Common European Framework of Reference for Languages (CEFR) levels assessed in the APTIS test*. AR-G/2021/1. ARAGs Research Reports Online, British Council.

Tabari, M. A., & Johnson, M. D. (2023). Exploring new insights into the role of cohesive devices in written academic genres. *Assessing Writing*, *57*. https://doi.org/10.1016/j.asw.2023.100749

Wright, T. S., & Cervetti, G. N. (2017). A systematic review of the research on vocabulary instruction that impacts text comprehension. *Reading Research Quarterly*, *52*(2), 203–226. https://doi.org/10.1002/rrq.163