Evaluating Readability of LLM-Simplified Texts Using Human Judgments and Automated Metrics

Hatsune Ichidate[1,*], Yiling Dai[2] and Hiroaki Ogata[1]

[1]Kyoto University,Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
[2]Hiroshima University, 1-3-2 Kagamiyama, Higashihiroshima, 739-8511, Japan

Abstract

Extensive reading requires learners to engage with materials that match their proficiency and interests. Recent large language models (LLMs) offer a promising way to rewrite texts for personalized readability adaptation. However, it remains unclear whether improvements reported by text difficulty correspond to pedagogical judgments or learner-perceived appropriateness. To address this issue, we compared the original and LLM-simplified reading materials using an evaluation framework consisting of text difficulty, expert ratings, and learner perceptions. Results showed that LLM-simplified texts were generally rated as simpler by text difficulty and expert ratings, particularly in vocabulary difficulty and syntactic complexity. Learner perceptions showed only weak group differences and weak associations with text difficulty overall, while low-frequency vocabulary showed a relatively large effect size despite the small sample. These findings suggest that evaluating LLM-based readability adaptation should combine objective measures with learner-centered feedback, rather than relying solely on conventional text difficulty.

Keywords

readability, large language models, English as a Foreign Language, text evaluation

1. Introduction

Reading is widely recognized as a key component of language learning because it provides repeated exposure to vocabulary, grammar, and discourse structures [1, 2]. Extensive reading is one of the most powerful approaches for improving reading skills. For it to be successful, learners need access to materials that are both interesting and matched to their reading proficiency [1, 3]. Texts that are too difficult may hinder comprehension, whereas overly easy texts may reduce engagement. Despite its importance, selecting and preparing suitable reading materials for learners remains challenging [1]. Many texts originally written for first-language readers may be too difficult for second-language learners because of advanced vocabulary, complex syntax, and implicit background knowledge [4].

Recent advances in large language models (LLMs) offer a promising way to adapt reading materials to learners' proficiency levels and interests by rewriting difficult texts into more accessible versions while preserving core content [5, 4]. For example, prior work rewrote academic texts using ChatGPT and compared different prompting strategies, finding through LIX-based readability evaluation that prompting can improve the accessibility of academic texts for diverse learners [6]. Other studies used LLMs to estimate and simplify the difficulty of French texts and videos based on learners' proficiency levels and topic interests, reporting that the adapted materials showed lower CEFR difficulty levels after rewriting [7]. These findings suggest that LLMs can support adaptive reading systems through text simplification and difficulty control.

However, an important challenge remains in evaluating LLM-based text simplification for adaptation.

HAI-Agency: Workshop on Orchestrating Human and AI Agency for Proactive and Reflective Learning, co-located with the 27th International Conference on Artificial Intelligence in Education (AIED 2026), Seoul, Korea, June 27 - July 3, 2026.
*Corresponding author.
†These authors contributed equally.
ichidate.hatsune.53r@st.kyoto-u.ac.jp (H. Ichidate); daiyl@hiroshima-u.ac.jp (Y. Dai); ogata.hiroaki.3e@kyoto-u.ac.jp (H. Ogata)
https://researchmap.jp/daiyiling (Y. Dai)
0000-0001-9900-8763 (Y. Dai)

Readability research has long shown that text difficulty cannot be fully captured by surface-level formulas alone [8]. In language learning contexts, perceived difficulty is also influenced by vocabulary familiarity, background knowledge, and individual learner differences. Existing studies often report improvements in automated readability metrics, but it remains unclear whether such improvements correspond to pedagogically meaningful gains recognized by experts or perceived as appropriate by learners[9, 8, 6]. Some prior studies have combined automated metrics with expert ratings when evaluating simplified texts, but learner perceptions remain underexplored [5]. Therefore, relying solely on automated metrics may overlook aspects of readability that matter most in real learning situations.

To address this issue, a multi-perspective evaluation approach is needed. Automated metrics capture linguistic difficulty and surface text features, expert ratings provide pedagogically informed judgments of suitability, and learner perceptions reflect actual reading experiences [10]. Examining where these perspectives align or diverge can provide deeper insight into the strengths and limitations of readability adaptation by LLM-based text simplification.

Accordingly, this study compares original and LLM-simplified reading materials from three perspectives: text difficulty, expert ratings, and learner perceptions. Specifically, we address the following research questions:

- **RQ1.** How do LLM-simplified texts differ from original texts in automatically calculated indicators of text difficulty and expert ratings of pedagogical appropriateness?
- **RQ2.** How do learner perceptions of appropriateness relate to simplified texts and text difficulty?

By answering these questions, we aim to provide initial insights into how LLM-simplified texts should be evaluated for learner-centered language learning contexts.

2. Method

2.1. Study Context and Design

This study conducted a secondary data analysis on a dataset collected through a controlled user experiment involving texts generated by an LLM-based reading support system [11]. First, Japanese engineering students who were English L2 learners participated in a reading activity using either original or simplified materials. Second, the texts and learner responses were analyzed from the perspectives of text difficulty, expert rating, and learner perceptions.

The reading support system used Gemini-2.5-flash to simplify materials based on learners' reading levels, combining numeric readability control (Flesch score) [12] qualitative rewriting instructions (Figure 2). In the controlled experiment, 16 university students were randomly assigned to either a control group (original texts) or an experimental group (simplified texts). They first completed a reading-level assessment before the activity. After a 60-minute reading session with a 5-minute break, participants completed a post-questionnaire on the readability of the materials.

As illustrated in Figure 1, two complementary datasets were derived from the study. For RQ1, a paired text dataset consisting of original passages and their LLM-simplified versions was used for text difficulty analysis and expert appropriateness ratings. Two native English-speaking experts evaluated whether each passage was appropriate for extensive reading in EFL contexts. For RQ2, a learner-response dataset linked questionnaire responses with the passages assigned during the reading activity.

2.2. Readability Evaluation Framework

Building on the study design above, we evaluated readability from three complementary perspectives: text difficulty, expert ratings, and learner perceptions. Because readability in learning contexts depends not only on linguistic difficulty but also on pedagogical suitability and learner experience, these perspectives were jointly examined.

Six readability dimensions (shown in Table 1) were selected based on the rewriting prompt shown in Figure 2. Indicators of text difficulty were automatically calculated for dimensions that could

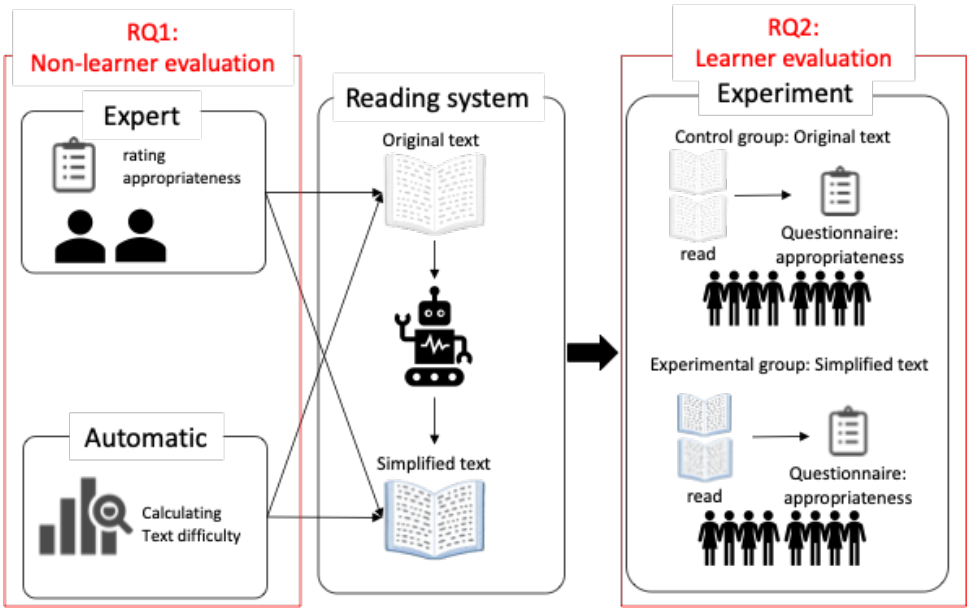

Figure 1: Overview of the study design

Numerous Readability Manipulation:
The Flesch Reading Ease score is a numeric measure of text readability, where higher scores indicate easier readability and lower scores indicate more difficult text.

In this task, we are trying to rewrite a given text into the target Flesch Reading Ease score and preserving the original meaning and information.

Given the original draft (Flesch Reading Ease = {original_fresch}):
[TEXT START]
{text}
[TEXT END]
Rewrite the above text to the difficulty level of:
Flesch Reading Ease = {target_flesch}

Qualitative Readability Control
· Content preservation
- Maintain the original meaning faithfully.
- Do not add new information.
- Do not remove important information.
- Do not introduce interpretations or opinions that are not present in the original text.

· Clarity & Naturalness Refinement
- Rewrite the text in clear, modern English. Remove archaic expressions and unnatural or outdated syntax typical of older texts at all levels.
- Minimize figurative language, idioms, and expressions whose meanings are not directly inferable.
- Use simple and direct sentence structures.
- Prefer familiar, high-frequency vocabulary.
- Avoid jargon; if technical terms are necessary, explain them clearly in simple language.
- Maintain a natural flow of reading rather than a sequence of evenly shortened sentences. Reduce average sentence length while avoiding uniform or fixed-length sentences. Introduce natural variation in sentence length to maintain rhythm and avoid mechanical patterns.

Formatting Constraints:
· Scope of rewriting
- Rewrite ONLY the main body text.
- Completely EXCLUDE titles, headings, chapter labels, author names, source information, footnotes, annotations, and introductions.
- Do NOT include any text other than the rewritten main body under any circumstances.
· Structure and formatting
- Preserve the original paragraph structure of the main text.
[Omitted]
- Do not include [TEXT START] and [TEXT END] in the output.

Figure 2: Prompt for simplifying reading materials.

be objectively quantified through natural language processing methods. Expert ratings and learner perceptions were collected using the same six dimensions to enable comparison across perspectives. This design enabled both passage-level and experience-level evaluation. Because these three perspectives differed in evaluation granularity, direct alignment across text difficulty, expert ratings, and learner perceptions should be interpreted as approximate rather than exact.

Table 1
Overview of the readability evaluation framework.

Readability dimension	Text difficulty	Expert	Learner
Amount of low-frequency vocabulary	✓	✓	✓
Amount of technical terminology		✓	✓
Frequency of archaic expressions		✓	✓
Frequency of figurative expressions		✓	✓
Sentence length	✓	✓	✓
Syntactic complexity	✓	✓	✓

2.3. Analysis Methods

Based on the framework above, we operationalized automatically measurable dimensions as indicators of text difficulty. We focused on sentence length, syntactic complexity, and low-frequency vocabulary, computed in Python using the Natural Language Toolkit (NLTK) [13].

Sentence length was measured using *average sentence length* and *Flesch score* [12]. *Syntactic complexity* was measured using *clauses per sentence*, d*ependent clause ratio*, *relative clause density*, and *complex nominal density*. To estimate *low-frequency vocabulary*, we used the CEFR-J Wordlist, a CEFR-aligned lexical resource developed for English education in Japan [14]. Words were assigned to four levels (A1, A2, B1, B2), and words outside the CEFR-J Wordlist were treated as low-frequency items.

For RQ1, original and LLM-simplified passages were compared using text difficulty indicators and expert ratings, and associations between these perspectives were examined. For RQ2, learner questionnaire responses were compared between groups and related to text difficulty indicators of the passages read.

3. Result

3.1. RQ1

3.1.1. Text Difficulty

Using the paired text dataset ($N = 8$ text pairs), we compared original and LLM-simplified texts using Wilcoxon signed-rank tests. Table 2 summarizes the results, and Figures 3 and 4 visualize text-level changes. Most indicators showed significant differences between the two versions.

For *sentence length, average sentence length* decreased and *Flesch score* increased, indicating shorter and easier texts after simplification. All text pairs showed this overall pattern.

For *syntactic complexity*, most indices significantly decreased, suggesting simpler sentence structures in the LLM-simplified texts. However, *relative clause density* did not change significantly ($p = .250$), indicating that some syntactic features were less consistently modified.

For *vocabulary difficulty*, lexical profiles shifted toward easier CEFR levels. The proportions of *A1*- and *A2*-level words increased, whereas the proportion of out-of-list words(*low-frequency vocabulary*) decreased. *B1*-level vocabulary showed no significant change. Overall, these results suggest that simplification mainly reduced *sentence length*, *syntactic complexity*, and *low-frequency vocabulary*.

3.1.2. Expert Ratings

Experts evaluated each text on a five-point scale from 1 (very inappropriate) to 5 (very appropriate) for L2 learners. Figure 5 shows that LLM-simplified texts generally received higher ratings than the

Table 2
Comparison of original and LLM-simplified texts across readability metrics

Metric	Orig. Mean (SD)	Gen. Mean (SD)	Diff	W	p	r
Sentence length						
Sentence Length	25.11 (6.79)	12.97 (2.33)	-12.15	0.0	.003	0.89
Flesch score	46.41 (9.05)	63.17 (10.68)	+16.76	0.0	.004	0.89
Syntactic Complexity						
Clauses per sentence	3.15 (1.37)	2.09 (0.48)	-1.06	4.0	.027	0.73
Dependent clause ratio	1.00 (0.21)	0.66 (0.15)	-0.34	0.0	.004	0.89
Relative clause density	0.19 (0.17)	0.13 (0.07)	-0.06	12.0	.250	0.41
Complex nominal density	1.59 (0.55)	0.76 (0.33)	-0.84	0.0	.004	0.89
vocabulary difficulty						
A1 ratio	0.40 (0.07)	0.45 (0.07)	+0.06	3.0	.020	0.77
A2 ratio	0.14 (0.03)	0.17 (0.03)	+0.03	0.0	.004	0.89
B1 ratio	0.12 (0.02)	0.12 (0.02)	+0.00	17.0	.570	0.22
B2 ratio	0.09 (0.03)	0.08 (0.02)	-0.01	5.0	.039	0.69
out of list ratio	0.25 (0.08)	0.18 (0.07)	-0.07	0.0	.004	0.89

Note. W = Wilcoxon signed-rank statistic. r = effect size.

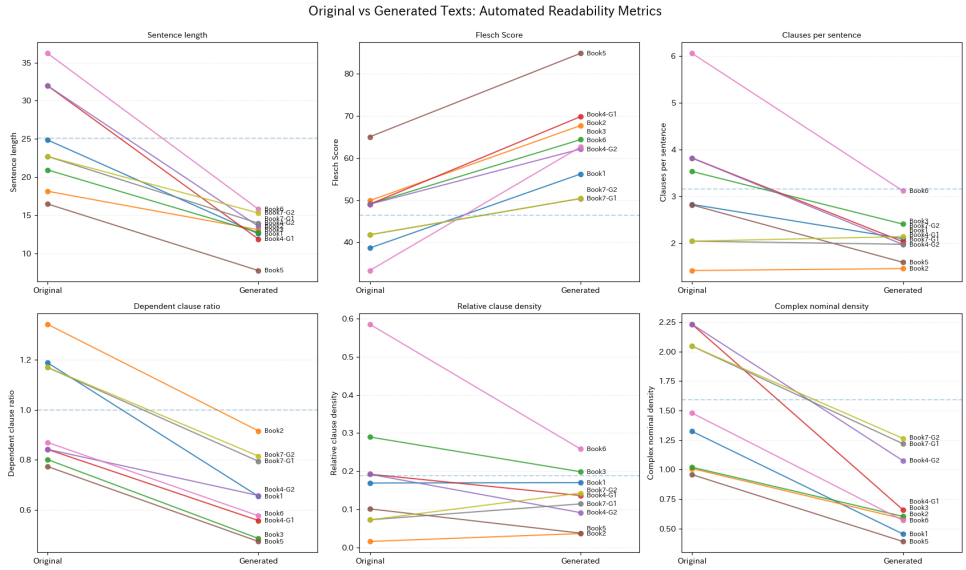

Figure 3: Changes in sentence length and syntactic complexity features

original texts across several dimensions, particularly for *low-frequency vocabulary*, *syntactic complexity*, and *archaic expressions*.

Table 3 reports inter-rater agreement using quadratic weighted Cohen's kappa. For inter-rater agreement analysis, only items rated by both evaluators were included. Agreement was highest for *amount of low-frequency vocabulary* ($\kappa = .591$), indicating moderate consistency between raters. Moderate agreement was also observed for *technical terminology*, *archaic expressions*, and *syntactic complexity*. In contrast, *figurative expressions* ($\kappa = .149$) and *sentence length* ($\kappa = .270$) showed relatively lower agreement, suggesting that these dimensions may have involved greater ambiguity or variability in interpretation across raters.

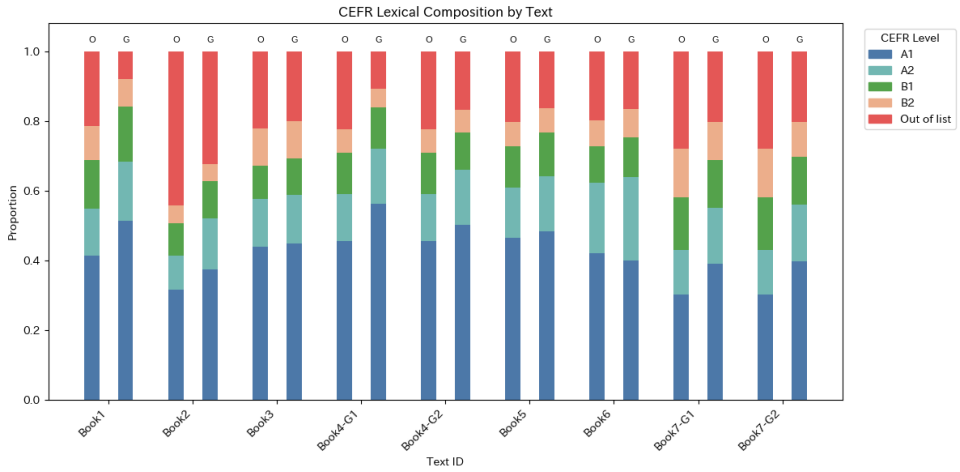

Figure 4: CEFR-based lexical composition

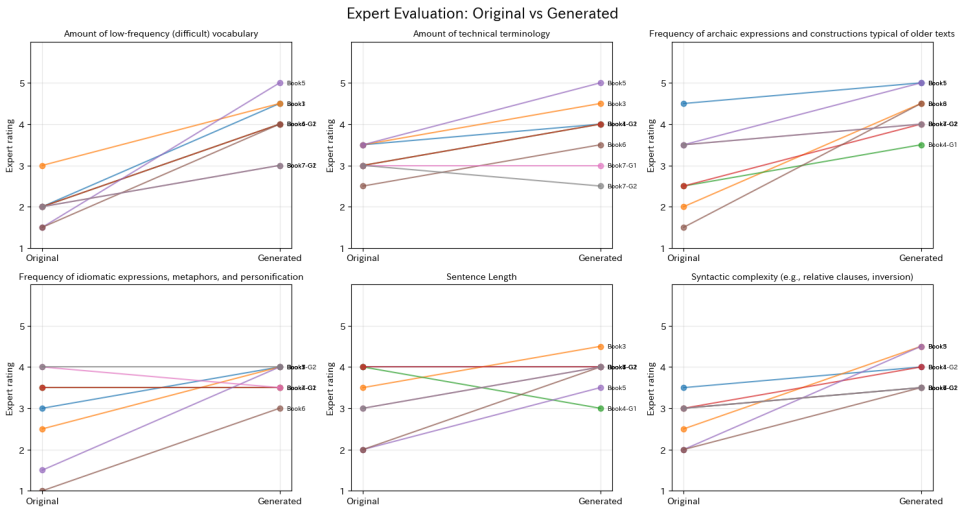

Figure 5: Results of expert ratings

Table 3
Inter-rater agreement for expert evaluations using quadratic weighted Cohen's κ

Metric	N	Weighted κ
Low-frequency vocabulary	16	.591
Technical terminology	16	.459
Archaic expressions	16	.465
Figurative expressions	15	.149
Sentence length	16	.270
Syntactic complexity	16	.364

3.1.3. Alignment Between Text Difficulty and Expert Ratings

To examine whether automated indicators of text difficulty aligned with expert judgments of pedagogical appropriateness, we calculated Spearman correlations between corresponding dimensions (Table 4). Vocabulary ratings were compared with the *out-of-list ratio*, syntactic ratings with four syntactic complexity indices, and sentence-length ratings with *average sentence length* and *Flesch score*.

For vocabulary, expert ratings were significantly negatively correlated with the *out-of-list ratio* ($\rho = -.670, p = .004$), indicating that texts containing more low-frequency vocabulary were judged

Table 4
Spearman correlations between text difficulty and expert ratings

Feature	ρ	p
Vocabulary		
Out-of-list ratio	-.670	.004
Syntactic complexity		
Clauses per sentence	-.608	.012
Dependent clause ratio	-.628	.009
Relative clause density	-.240	.370
Complex nominal density	-.593	.015
Sentence length (Expert Evaluation)		
Flesch score	0.115	.670
Average sentence length	-0.210	.434

Note. ρ = Spearman's rank correlation coefficient.

as less appropriate. For syntactic complexity, significant negative correlations were found for *clauses per sentence* ($\rho = -.608, p = .012$), *dependent clause ratio* ($\rho = -.628, p = .009$), and *complex nominal density* ($\rho = -.593, p = .015$), suggesting that structurally more complex texts were also judged as less appropriate. *Relative clause density* was not significantly associated with expert ratings.

For sentence length, no significant correlations were observed. Although shorter sentences showed a weak tendency to receive higher appropriateness ratings, neither *average sentence length* nor *Flesch score* was significantly related to expert judgments.

3.2. RQ2

3.2.1. Group Differences in Learner Perceptions

Using learner responses from 16 participants (control $n = 8$, experimental $n = 8$), we compared post-activity ratings of perceived appropriateness between learners who read original texts and those who read LLM-simplified texts.

There were no statistically significant differences in table 5 experimental group tended to report higher appropriateness ratings for *sentence length*, *figurative expressions*, and *low-frequency vocabulary* than the control group. In particular, medium effect sizes were observed for *figurative expressions* ($r = .42$) and *low-frequency vocabulary* ($r = .86$), suggesting that simplified texts may have been perceived as more accessible in these respects despite the lack of statistical significance.

Overall, these results suggest that while learners did not report significantly different perceptions between conditions, the simplified texts showed a tendency toward more favorable readability judgments.

Table 5
Comparison of learner perceptions between control and experimental groups

Feature	Control group Mean (SD)	Experimental group Mean (SD)	Test statistic	p	Effect size
Sentence length	3.38 (0.74)	3.75 (0.71)	$U = 21.5$	.199	.28
Amount of technical terminology	2.38 (0.92)	2.38 (0.52)	$U = 31.0$	.953	.03
Syntactic complexity	3.25 (0.89)	3.25 (0.89)	$U = 32.0$	1.000	.00
Frequency of figurative expressions	2.63 (1.06)	3.50 (0.53)	$U = 16.0$	.085	.42
Amount of low-frequency vocabulary	2.38 (0.92)	3.13 (0.83)	$t = -1.71$	.109	.86
Frequency of archaic expressions	3.13 (1.13)	3.50 (0.53)	$U = 28.0$	.686	.11

Note. The control group read original texts, whereas the experimental group read LLM-generated simplified texts. Independent-samples t-tests were used when both groups satisfied normality assumptions; otherwise Mann–Whitney U tests were applied. Effect size is reported as Cohen's d for t-tests and rank-biserial correlation for Mann–Whitney tests.

3.2.2. Associations with Text Difficulties

Using the learner-response dataset ($N = 16$), we examined whether learner perceptions of appropriateness were associated with automated indicators of text difficulty. For each learner, all passages read during the activity were concatenated into a single text, and the same automated indicators used in RQ1 were computed from this aggregated reading input.

Table 6 presents the Spearman correlations between learner perceptions and automated readability indices. No statistically significant correlations were observed for any dimension. For vocabulary, *out-of-list ratio* showed a moderate negative correlation with learner perceptions ($\rho = -.414, p = .111$), suggesting that learners tended to perceive texts with fewer out-of-list words as more appropriate. For the evaluation dimension of *sentence length*, *average sentence length* showed a moderate negative tendency ($\rho = -.376, p = .152$), suggesting that shorter sentences may have been perceived as more appropriate. *Flesch score* was not associated with learner perceptions.

Table 6
Spearman correlations between automated readability metrics and learner perceptions

Category	Feature (Statistic)	ρ	p
Vocabulary (Subjective Fit)			
out of list ratio		-0.414	.111
Syntactic complexity (Subjective Fit)			
Clauses per sentence		0.147	.587
Depth clause ratio		-0.083	.760
Relative clause density		0.013	.963
Complex nominal density		0.173	.523
Sentence length (Subjective Fit)			
Flesch score		-0.018	.947
Average sentence length		-0.376	.152

Note. ρ = Spearman's rank correlation coefficient.

4. Discussion and Conclusion

4.1. Interpretation and Insights

This study examined LLM-based text simplification from the perspectives of text difficulty, expert appropriateness ratings, and learner perceptions in order to better understand how readability improvements should be evaluated in language learning contexts. For RQ1, LLM-simplified texts were generally found to be easier than the original texts according to automated indicators of text difficulty and expert ratings. For RQ2, learner perceptions did not show statistically significant differences between groups and were only weakly related to automated indicators. Together, these findings suggest that reducing text difficulty alone may not be sufficient to produce readability improvements that are meaningfully perceived as appropriate by learners. Future readability adaptation should therefore integrate linguistic simplification with learner-centered and personalized evaluation.

A closer examination of individual dimensions helps explain why reductions in text difficulty did not always translate into learner-perceived appropriateness. Among the evaluated dimensions, vocabulary-related indicators showed the most consistent alignment across automated metrics, expert ratings, and learner tendencies. This suggests that lexical simplification may be a particularly robust and interpretable target for LLM-based readability adaptation. Prior research has also emphasized vocabulary knowledge as a central predictor of L2 reading comprehension and text difficulty [15, 16]. Replacing rare or unfamiliar words with more frequent and level-appropriate alternatives may therefore provide relatively reliable benefits for diverse learners.

In contrast, *syntactic complexity* showed a different pattern. Automated syntactic indicators showed

that simpler generated texts were also judged as more appropriate by experts. However, learner perceptions did not clearly reflect these differences. One possible interpretation is that syntactic difficulty can be identified analytically but is less consciously noticeable to learners. Prior research has shown that syntactic complexity is associated with text readability, particularly for second-language readers [10]. Learners may experience smoother comprehension without explicitly attributing that experience to grammatical structure. Another possibility is that subjective questionnaire ratings are less sensitive to structural changes than direct comprehension measures. In either case, the results suggest that some readability improvements may operate implicitly rather than explicitly.

The findings also suggest limits of traditional readability formulas when used alone. Although the *Flesch score*, a widely used readability indicator, improved substantially after rewriting, it showed little association with learner perceptions in this study. This may suggest that broad readability formulas capture some useful surface-level features, but may only partially reflect how second-language learners experience difficulty, particularly when evaluating simplified texts at the sentence level [10, 8]. Readability judgments may also depend on more specific factors such as word familiarity, conceptual clarity, and local processing demands.

Some dimensions were difficult even for human evaluators to assess consistently. For example, *figurative expressions* showed relatively low inter-rater reliability, suggesting that judgments about figurativeness may be subjective and strongly context dependent [17]. Similarly, learner perceptions did not strongly differentiate *technical terminology* or *archaic expressions*. Learners may have found it difficult to explicitly judge these categories, particularly when unfamiliar items were perceived simply as unknown words regardless of their linguistic type [15, 16]. These findings suggest that readability adaptation should focus not only on linguistic categories themselves, but also on whether learners actually experience the text as understandable and appropriate. This mismatch may be especially relevant when document-level readability criteria are applied to learner-centered sentence rewriting tasks.

4.2. Limitations

Several limitations should be noted. First, this study was exploratory and involved a relatively small sample size, particularly for learner perceptions. As a result, some non-significant findings may reflect limited statistical power rather than the absence of meaningful effects. Second, learner perceptions were measured using post-activity questionnaire ratings rather than direct comprehension, behavioral logs, or think-aloud data. Additional measures may provide a more fine-grained understanding of how simplification affects reading processes. Third, the generated materials were produced using a single LLM configuration. Different models, prompting strategies, or rewriting constraints may yield different patterns of alignment across automated readability metrics, expert ratings, and learner perceptions.

4.3. Implications

These findings provide several implications for the design of adaptive reading systems using LLMs. The present results suggest that reducing text difficulty alone may be insufficient when generating simplified materials, because learner-perceived appropriateness did not always align with improvements in automated indicators. Rather than treating text simplification as a complete solution, adaptive systems should integrate multiple perspectives, particularly learner experience, consistent with learner-centered principles in adaptive educational systems [18]. In addition, vocabulary control may be a practical target for readability adaptation, as prior studies have also highlighted lexical adaptation as an effective strategy for language learners [7]. Because learner perceptions varied substantially, future systems may also benefit from personalization based on proficiency estimation rather than one-size-fits-all simplification. Rather than generating a single fixed version of a text, systems could adapt dimensions such as vocabulary difficulty, sentence length, or degree of explanation based on learner feedback and reading behavior. Overall, these findings suggest that adaptive text simplification should balance objective indicators of text difficulty with learner-perceived appropriateness.

5. Acknowledgments

This work was partly supported by the following grands: JST BOOST AC7266300002, Council for Science, Tehcnology and Innovation (CSTI), Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Early-Career Scientists JP24K20902, Cross-ministrerial Strategic Innovation Promotion Program (SIP) the 3rd perionde of SIP Creation of new ways of leaning and working in a post-COVID-19 era, and Grant Number JPJ012347 (Funding agency: JST).

6. Appendices

Declaration on Generative AI

During the preparation of this work, the author used GPT-5 in order to perform text translation, paraphrasing and rewording, writing improvement, grammar and spelling checking, and peer review simulation. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] R. R. Day, J. Bamford, W. A. Renandya, G. M. Jacobs, V. W.-S. Yu, Extensive reading in the second language classroom, Relc Journal 29 (1998) 187–191.

[2] S. Krashen, The Power of Reading: Insights from the Research, Libraries Unlimited, 2004.

[3] R. Day, J. Bamford, Top ten principles for teaching extensive reading, Reading in a Foreing Language, 2002.

[4] J. Arora, I. Elgort, J. Zhao, Content adaptation for language learning: A hybrid ai approach, The EuroCALL Review 32 (2025) 169–182.

[5] Y. Shi, K. Yu, Y. Dong, F. Chen, Large language models in education: a systematic review of empirical applications, benefits, and challenges, Computers and Education: Artificial Intelligence 10 (2026) 100529.

[6] E. Hedlin, L. Estling, J. Wong, C. Demmans Epp, O. Viberg, Got it! prompting readability using chatgpt to enhance academic texts for diverse learning needs, in: Proceedings of the 15th International Learning Analytics and Knowledge Conference, LAK '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 115–125.

[7] H. Jamet, M. Manderlier, Y. R. Shrestha, M. Vlachos, Evaluation and simplification of text difficulty using llms in the context of recommending texts in french to facilitate language learning, in: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 987–992.

[8] S. Vajjala, D. Meurers, Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications, International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification 165 (2014). doi:10.1075/itl.165.2.04vaj.

[9] F. Alva-Manchego, L. Martin, C. Scarton, L. Specia, EASSE: Easier automatic sentence simplification evaluation, in: S. Padó, R. Huang (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 49–54. URL: https://aclanthology.org/D19-3009/. doi:10.18653/v1/D19-3009.

[10] S. A. CROSSLEY, J. GREENFIELD, D. S. McNAMARA, Assessing text readability using cognitively based indices, TESOL Quarterly 42 (2008) 475–493.

[11] Y. Dai, T. Terao, Llm-driven text simplification and its effects on extensive reading in efl learners, in: International Conference on Artificial Intelligence in Education, Springer, 2026.

[12] R. Flesch, A new readability yardstick, Journal of Applied Psychology 32 (1948) 221–233.
[13] E. Loper, S. Bird, NLTK: The natural language toolkit, in: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 63–70.
[14] Y. Tono, The cefr-j and its impact on english language teaching in japan, in: JACET international convention selected papers, volume 4, 2016, pp. 31–52.
[15] P. Nation, Learning vocabulary in another language, volume 10, Cambridge university press Cambridge, 2001.
[16] N. Schmitt, X. Jiang, W. Grabe, The percentage of words known in a text and reading comprehension, The Modern Language Journal 95 (2011) 26 – 43. doi:10.1111/j.1540-4781.2011.01146.x.
[17] C. Cacciari, S. Glucksberg, Understanding figurative language, Handbook of Psycholinguistics (1994).
[18] O. Zawacki-Richter, V. Marín, M. Bond, F. Gouverneur, Systematic review of research on artificial intelligence applications in higher education -where are the educators?, International Journal of Educational Technology in Higher Education 16 (2019) 1–27. doi:10.1186/s41239-019-0171-0.