PLABA-EVAL: A Multi-Dimensional, In-Context Sentence Readability Dataset for Medical Text

Kexin Bian[1], **Su-Youn Yoon**[1,2], **Mamoru Komachi**[1]
[1]Hitotsubashi University [2]EduLab, Inc.
{kexin, komachi}@scl.sds.hit-u.ac.jp, su-youn.yoon@edulab-inc.com

Abstract

We introduce PLABA-EVAL, a dataset for in-context sentence-level readability assessment in biomedical abstracts and their plain-language adaptations. Participants read biomedical abstracts with full-document access, provide sentence-level ratings of *Processing Ease* and *Perceived Understanding*, and then complete an open-book multiple-choice comprehension check. The dataset comprises 609 sentences from 78 biomedical abstracts and expert plain-language adaptations, each read and rated in the context of its full document by three independent raters, together with responses to 168 manually written open-book MCQs. Our work complements existing medical simplification and readability resources that focus on lexical simplification, single-score readability labels, or decontextualized sentence ratings. Analyses show that ease and understanding can diverge at the sentence level, that simplification yields uneven gains across sentences, and that high perceived understanding does not eliminate comprehension errors. We further provide baseline linguistic analyses and comparisons to existing readability predictors, illustrating how PLABA-EVAL can support work on readability assessment, simplification, and sentence-level difficulty modeling.

Keywords: readability assessment, text simplification, human evaluation

1. Introduction

Biomedical research is increasingly consumed by non-specialists, including patients, caregivers, and the general public (Attal et al., 2023). In these settings, accessibility is not only about scientific accuracy but also about whether readers can extract key claims efficiently and understand them correctly. Readability modeling and text simplification aim to support this goal, yet many practical tools and NLP benchmarks treat difficulty as a single document-level score (Vajjala and Meurers, 2012; Xu et al., 2015; Vajjala and Lučić, 2018). This is a poor fit for how people read, as difficulty is often experienced locally, where a mostly accessible abstract can still contain a handful of sentences that are disproportionately effortful or confusing. For sentence-level methods, difficulty is often estimated from sentences presented in isolation. This setting removes discourse context that readers rely on for interpretation, and can therefore misrepresent the difficulty a sentence poses as part of the full text (Schumacher et al., 2016; Iavarone et al., 2021).

A second gap is that a single "difficulty" judgment conflates distinct aspects of reader experience. Sentences can be easy to process yet remain underspecified or confusing in context, and conversely can require effort but still be understood after integration (Van den Broek et al., 1999). Moreover, perceived understanding can diverge from their actual comprehension, motivating the use of explicit comprehension checks (Cohen et al., 2025; Leroy et al., 2012). Existing resources rarely capture these distinctions, making it difficult to disentangle ease from perceived and actual understanding, and to characterize how simplification affects each dimension across sentences.

To address these gaps, we introduce an in-context, multi-dimensional approach to sentence-level readability assessment for medical text, while combining local ratings with an open-book comprehension check. Our contributions are as follows:

- We introduce **PLABA-EVAL**, a dataset for in-context sentence-level readability assessment in biomedical abstracts and expert plain-language adaptations, with full-document sentence ratings and manually curated MCQs that assess comprehension.[1]
- We present a multi-dimensional view of readability that distinguishes *Processing Ease*, *Perceived Understanding*, and actual comprehension as measured by MCQs.
- We provide analyses and baselines showing that these dimensions are related but not interchangeable, that simplification yields uneven sentence-level gains, and that existing readability predictors capture only part of the human signal.

2. Related Work

Readability is traditionally defined as the ease with which text is read and understood (Dale and

[1]We release PLABA-EVAL under CC BY 4.0 to enable broad reuse and redistribution with attribution.

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 49–60
11 May 2026.

Chall, 1948; Richards and Schmidt, 2013). Cognitive and reading theories suggest, however, that these are not identical: online processing effort and successful meaning construction can diverge (Van den Broek et al., 1999; Just and Carpenter, 1980; van den Broek et al., 2011). Subjective perceptions of difficulty may likewise differ from comprehension performance (Leroy et al., 2010), especially when readers form only shallow or incomplete interpretations (McKoon and Ratcliff, 1992; Trabasso and Van Den Broek, 1985). In health contexts, simplification has been shown to lower perceived difficulty without consistently improving objective information retention (Leroy et al., 2010; Shulman et al., 2020). These distinctions motivate combining subjective ratings with explicit comprehension checks.

These measurement considerations also interact with the choice of annotation unit. Prior work has motivated sentence-level complexity or understandability ratings, but discourse-level factors such as cohesion can shape comprehension across sentences and paragraphs (Snow, 2002). Fully decontextualized sentence judgments may therefore be incomplete, especially for lay readers of technical medical materials, where successful integration depends on domain knowledge and how explicitly relations are stated (Ozuru et al., 2009; Kindig et al., 2004; Berkman et al., 2011). This motivates sentence-level annotation with access to surrounding context.

Existing resources relevant to sentence difficulty emphasize different aspects of the problem. Among those that incorporate human feedback, many collapse complexity into a single scale, such as CEFR-aligned rankings in MedReadMe (Jiang and Xu, 2024) or grade-level targets in general-domain datasets like CLEAR (Crossley et al., 2023). However, CEFR in particular was designed for language-learner proficiency, making it a less natural fit for lay medical comprehension (Council of Europe, 2001; Crossley et al., 2023).

Many resources focus on identifying difficult terms and how they should be simplified (Ondov et al., 2026; Xia et al., 2025), which is highly relevant in medical settings, where lexical accessibility is a major source of difficulty. But lexical difficulty alone does not fully determine sentence difficulty in context. Sentence-difficulty resources that manipulate local contextual windows (Schumacher et al., 2016; Iavarone et al., 2021) show that surrounding context can change perceived difficulty, but they typically model context through limited local windows rather than full-document reading context. Finally, a smaller line of work on health-information readability and simplification has also incorporated comprehension questions to distinguish perceived from actual difficulty (Leroy et al., 2010; Guidroz et al., 2025). However, these studies do not typically pair comprehension checks with in-context sentence-level ratings. Our work aims to bridge these gaps and provide a more granular view of how readers process and understand medical text.

	Original		Simplified	
	Mean	SD	Mean	SD
Sent per doc	6.97	1.81	8.64	2.96
Word per doc	152.36	14.47	174.49	38.31
MedReadMe	4.67	0.40	4.44	0.39

Table 1: Descriptive statistics for original and simplified documents.

3. Data Collection

We construct PLABA-EVAL by sampling 39 PubMed abstracts and their corresponding expert adaptations from the PLABA corpus (Attal et al., 2023). This results in a parallel set of 78 documents, totaling 609 sentences (272 in the original abstracts and 337 in the adaptations). We collect 3,654 sentence-level Ease/Understanding ratings from three independent raters per document, along with 468 document-level ratings. To evaluate comprehension, we additionally develop 168 open-book MCQs answerable from both variants.

3.1. Source stimuli

We draw our source stimuli from the Plain Language Adaptation of Biomedical Abstracts (PLABA) dataset (Attal et al., 2023), which pairs PubMed abstracts with manual plain-language adaptations written to answer high-frequency consumer health questions from MedlinePlus query logs.

We applied filters to retain abstracts between 120 and 180 words and applied stratified sampling over document-level readability, computed as the mean MedReadMe sentence score per abstract (Jiang and Xu, 2024) [2], to cover a range of baseline difficulty. To encourage topical diversity, we limited the sample to at most two abstracts per PLABA source question, yielding 39 source–adaptation pairs. Table 1 reports summary statistics for the sampled subset.

3.2. Text Difficulty Rating Procedure

Figure 1 summarizes the procedure. Participants were instructed to read each abstract naturally, as if

[2]We use the authors' released checkpoint: https://huggingface.co/chaojiang06/medreadme_medical_sentence_readability_prediction_CWI

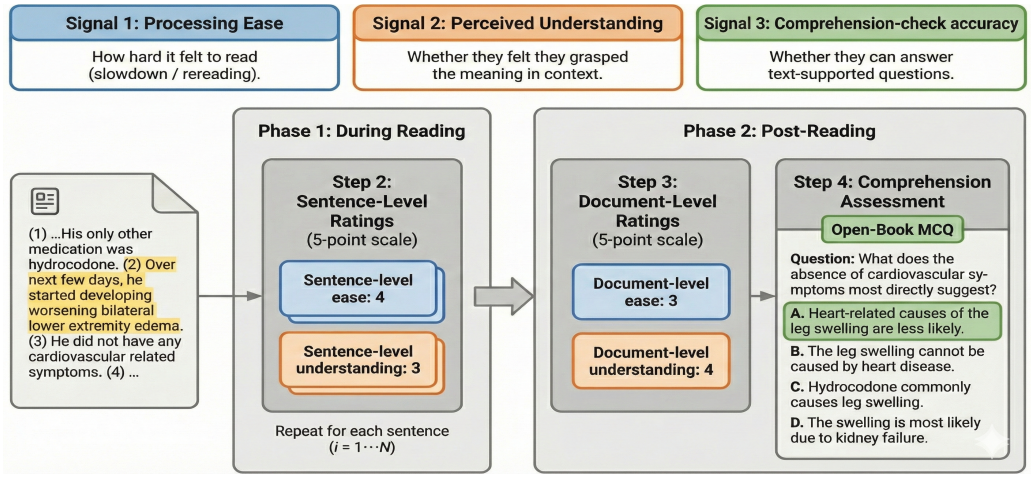

Figure 1: Difficulty rating framework. Participants read each document with full-text access and provide sentence-level ratings in sequence on two 5-point scales. After reading, they provide document-level ratings on the same scales and complete an open-book multiple-choice comprehension check.

looking for important medical information for themselves or a family member. They then completed a short onboarding tutorial to familiarize them with the interface, rating scales, and study flow.

We elicited text-difficulty judgments at two points: during reading and post-reading, followed by an open-book comprehension check. During reading, participants rated each sentence in sequence on two 1–5 scales and provided brief diagnostic feedback, with the full document visible at all times. After completing all sentence-level ratings for a document, participants provided corresponding document-level ratings. To characterize potential variation in reader background, we also collected self-reported topic familiarity and familiarity with academic-style writing. Finally, participants answered text-supported multiple-choice questions (MCQs) with the document still visible; correct answers were revealed after submission to avoid leaving participants misinformed by any study item.

3.3. Measures

Subjective ratings We operationalize perceived difficulty using two complementary subjective signals collected on 1–5 scales (higher is better): *Processing ease* ("While reading, how easy was this sentence to read?"), capturing experienced workload (e.g., slowdown or rereading); *Perceived understanding* ("How confident are you that you understood what this sentence means in this context?"), capturing metacognitive confidence that the sentence "made sense" given the surrounding text.

Diagnostic feedback After rating each sentence, participants also provided diagnostic feedback on why a sentence felt difficult. In the first half of data collection, raters could optionally leave brief free-text comments (e.g., to flag confusing terms or ask questions). In the second half, raters selected up to three difficulty factors from a fixed set of *issue tags* derived from earlier free-text feedback (e.g., vocabulary, structure/density, unnatural wording, unclear takeaway, unclear logic/connection).

Comprehension check (MCQs) After rating each document, participants answered an open-book comprehension check consisting of multiple-choice questions designed to test text-supported understanding. Items targeted information likely to matter to lay readers and avoided questions that hinge solely on highly technical details. The open-book format reduces memory demands and shifts the task toward locating and using textual evidence, so errors more directly reflect misinterpretation or failure to apply relevant evidence (Durning et al., 2016).

3.4. MCQ Curation

We manually wrote MCQs and used the same question set for both versions within each original–simplified abstract pair. To support this paired design, items targeted claims preserved under simplification, and we aimed to minimize lexical overlap between item text and either abstract version to avoid version-specific cues.

Because the task was open-book, stems and correct answers were often paraphrased to discourage keyword matching. Distractors were designed as plausible alternatives, drawing on common sources of confusion such as superficial lexical similarity, reasonable but incorrect interpretations, or scope/magnitude changes. All items were manually drafted and reviewed to ensure answerability from the abstract and exactly one correct option.

LLMs were used as a brainstorming aid for a subset of items (e.g., candidate stems or distractors); all suggestions were treated as drafts and manually edited and verified against the text.

Quality check We further validated MCQ quality using GPT-5.2, prompting the model to answer using only the provided abstract or simplification. The model agreed with the manually verified gold answers on 314/316 item–version instances (99.37%). The two discrepant cases, both on simplified abstracts, were traced to subtle interpretive differences in paraphrasing. After manual review, these items were excluded from the final dataset and all subsequent analyses.

Critical sentence Following evidence-linking practices in MCQ reading comprehension (e.g., STARC (Berzak et al., 2020)), each item was anchored to a single critical *evidence sentence*. When evidence was distributed or required cross-sentence integration, we selected the sentence that most explicitly supported the correct option. We recorded sentence indices in the original abstracts and mapped them to the simplified versions using PLABA sentence alignments.

3.5. Participants and Setup

To approximate reading by non-expert consumers of health information, we recruited native English speakers without screening for biomedical background. To characterize potential variation in expertise, we collected self-reported topic familiarity (per abstract) and familiarity with academic-style writing (frequency of reading academic texts). Because topic familiarity depends on document topic, we treat it as descriptive and not directly comparable across topics. Overall, a majority of raters reported that most or almost all information in the text was new to them, consistent with a largely non-expert reader pool. Some raters reported relevant personal experience (e.g., being diagnosed with the condition discussed); we retain these cases as plausible in real-world health-information seeking.

Implementation Participants were recruited via Prolific with eligibility restricted to first-language English speakers residing in US/UK/CA/AU and approval rate > 0.99, but did not otherwise control for socio-professional background. The study was implemented as a web-based annotation interface. Each document was annotated by three distinct workers, recruited independently for each task instance (i.e., no worker overlap across documents). Compensation was £1.50 for an estimated completion time of 8 minutes; the median observed completion time was 10.50 minutes (IQR: 6.26).

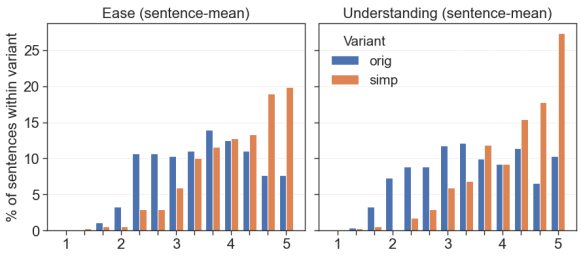

Figure 2: Sentence-level distributions of subjective ratings by variant (mean of three rater judgments). Bars show the percentage of sentences in each bin for the original (n=272) and simplified (n=337) variants.

3.6. Data Quality

Because the dataset intentionally captures subjective reading experience, we incorporated design-time safeguards to improve comparability across raters, texts, and variants. Instructions framed the task as goal-oriented health-information reading (Sabou et al., 2014), and the 1–5 rubrics for ease and understanding were anchored with behavioral examples (Melchers et al., 2011). Workers completed a short practice round before the main task (Knowles and Lo, 2025). We excluded low-quality submissions using completion-time thresholds and response-pattern checks (e.g., straightlining).

Reliability / Agreement We estimate inter-rater reliability using Gwet's AC2 with quadratic weights, which is comparatively robust to prevalence effects (Gwet, 2014). Computed per document and summarized across documents, median AC2 is 0.53 for Ease and 0.52 for Understanding, with substantial between-text variability (Ease IQR $= 0.40$, Understanding IQR $= 0.38$). This level of agreement is consistent with a subjective judgment task and with heterogeneity across biomedical texts. We find no evidence that agreement differs between original and simplified variants ($p = 0.145$), suggesting that the protocol yields comparable reliability across conditions.

4. Behavioral Findings

4.1. Subjective Rating Distributions

Sentence-level rating distributions Figure 2 shows sentence-level ratings by variant (mean over three raters per sentence). For the original abstracts, ratings span most of the 1–5 scale but are concentrated in the mid-to-high range, with relatively few sentences receiving very low scores on either dimension. Simplification shifts the distributions rightward on both axes (Ease $3.53 \pm 0.87 \rightarrow 4.08 \pm 0.79$; Understanding $3.46 \pm 0.97 \rightarrow 4.21 \pm 0.77$,

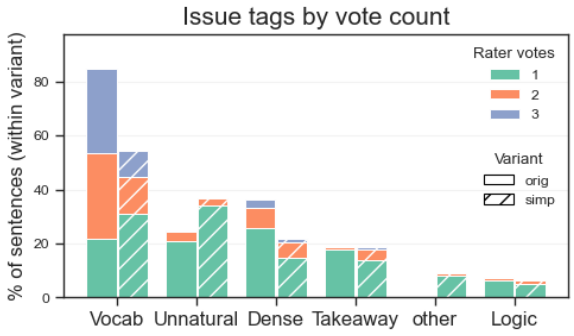

Figure 3: Issue tag prevalence by variant, shown as the percentage of sentences (within each variant) with 1, 2, or 3 raters endorsing each tag.

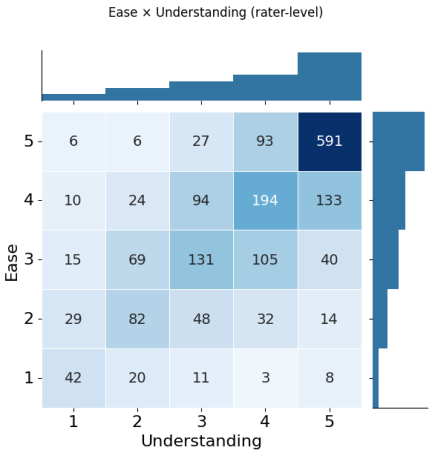

Figure 4: Joint distribution of rater-level Ease and Understanding scores (pooled across variants). Cells show counts of individual judgments; marginal bars show the corresponding one-dimensional distributions, highlighting ceiling effects.

orig→simp), and increases near-ceiling ratings, most notably for Understanding (with a clear mass at $= 5$).

Issue tag prevalence Figure 3 summarizes issue-tag prevalence by variant, stratified by the number of raters endorsing each tag. Vocabulary-related issues are the most frequent in both variants, with substantially stronger multi-rater agreement in the original text (a larger 2–3 vote component) than in the simplified text. Sentence density is flagged more often in the original variant, whereas Unnaturalness/Redundancy is flagged more often in the simplified variant. For most non-vocabulary tags (e.g., LOGIC and OTHER), endorsements are dominated by single-rater votes, indicating weaker consensus.

Document-level rating distributions Overall, post-reading judgments on documents are more moderate. At the document level, mean ratings are less concentrated at the higher end than at the sentence level, and occupy a broader range of values. By variant, document-level averages are also higher in the simplified condition (Ease $2.90 \pm 1.04 \rightarrow 3.47 \pm 1.12$; Understanding $3.15 \pm 1.09 \rightarrow 3.89 \pm 1.06$, orig→simp), but the ceiling effect is substantially weaker than for sentence-level Understanding.

We also quantify within-abstract variation by summarizing the spread of sentence-level mean ratings across sentences (P90–P10) within each document. The resulting ranges are typically 1.2–1.5 points on a 5-point scale, indicating that most abstracts contain a mix of locally easier and more challenging sentences rather than a uniform difficulty level. Document-level Ease and Understanding align strongly with the mean sentence rating (Spearman $\rho \approx 0.77$–0.83), suggesting that global ratings reflect an integrated impression across sentences.

4.2. Relationship between Ease and Understanding

Figure 4 shows the joint distribution of rater-level Ease and Understanding judgments. Ratings concentrate along the diagonal, with a strong overall association (rater-level Spearman $\rho = 0.696$; sentence-mean $\rho = 0.772$), indicating that sentences judged easier are typically also judged better understood. The positive association also holds within each variant (orig $\rho = 0.684$; simp $\rho = 0.656$).

However, the measures are not interchangeable. A nontrivial share of judgments falls off the diagonal, with large discrepancies persisting (orig: 12.3%, simp: 9.5% differ by ≥ 2 points in absolute value). Qualitatively, divergences often reflect a mismatch between decoding load and inferential load. For instance, brief clinical statements (e.g., “She was intubated for airway protection.”) are rated easy to read but poorly understood without domain knowledge, whereas sentences that add explanatory scaffolding (e.g., “hyperferritinemia (high levels of ferritin – an iron-containing protein)”) are rated harder to read yet better understood. These divergences reflect distinct sources of difficulty, which call for different interventions, motivating multi-signal evaluation rather than a single readability score.

4.3. Effects of Simplification on Sentence Ratings and Issue Tags

Ratings We estimate variant differences using between-condition comparisons on sentence alignments between the original and simplified variants ($n = 265$). Alignments include one-to-many cases arising from sentence splitting; in these cases, the

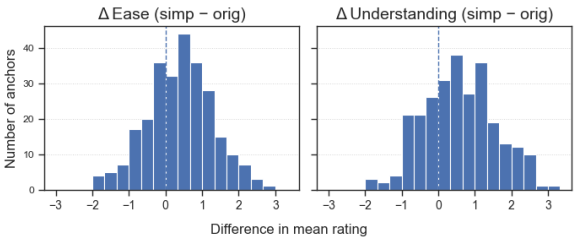

Figure 5: Alignment-based rating changes under simplification. Histograms show Δ = simp − orig in sentence-mean Ease and Understanding across sentence alignments between the original and simplified variants ($n = 265$)

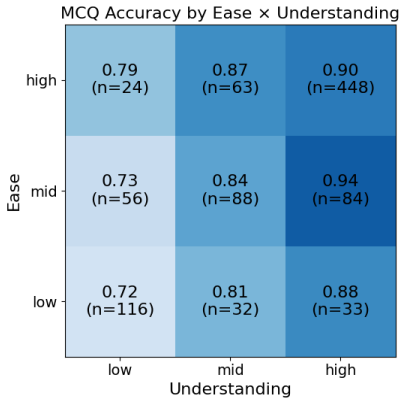

Figure 6: MCQ accuracy by evidence-sentence ratings. Cells show $P(\text{correct})$ (and n) by binned Understanding × Ease (low=1–2, mid=3, high=4–5).

simplified score is averaged across the corresponding split sentences.

Figure 5 shows the distribution of alignment-based changes in sentence-mean ratings (Δ = simp − orig). Both Ease and Understanding exhibit a clear positive shift, with median $\Delta = 2/3$ (mean $= 0.54$ for Ease and 0.72 for Understanding), indicating larger gains for Understanding. Large joint improvements are frequent (100/265; 37.7%), where both dimensions improve by at least $2/3$ ($\Delta_{\text{Ease}} \geq 2/3$ and $\Delta_{\text{Und}} \geq 2/3$). Among one-dimensional improvements (one dimension $\geq 2/3$ while the other below threshold), shifts are more common for Understanding than for Ease (14.0% vs. 6.4%). Notably, simplification is not uniformly beneficial: 27.2% show no meaningful change on either dimension, and 14.7% worsen on at least one dimension. This heterogeneity suggests caution in treating human-written simplifications as uniformly improved gold references for sentence-level evaluation.

Issue tags To characterize *what* changes when ratings shift, we analyze how issue-tag endorsements change under simplification for sentence alignments from the second half of data collection ($n = 123$), when raters selected up to three tags from a fixed set. For each tag, we compare the number of raters endorsing it in the original and simplified versions of the same aligned sentence and summarize the change as Δvotes = simp − orig.

Overall, simplification most consistently reduces terminology-related difficulty. UNCLEAR TERMS shows a large negative shift in endorsement (mean Δvotes $= -0.98$, computed over the 109 aligned pairs where the tag appears in either variant). SENTENCE DENSITY also tends to diminish, but less uniformly (mean Δvotes $= -0.40$; $n = 50$). In contrast, simplification more often introduces phrasing-level costs: UNNATURAL/WORDY PHRASING shifts upward (mean Δvotes $= +0.18$; $n = 61$). TAKEAWAY CLARITY is comparatively unstable, showing near-symmetric increases and decreases (mean Δvotes $= +0.14$; $n = 43$). These shifts are illustrated by paired examples in Table 2, highlighting how simplification can remove shared lexical barriers while leaving (or occasionally introducing) more heterogeneous clarity or coherence concerns.

4.4. Objective Comprehension (MCQ)

Across all MCQ responses (N=942), accuracy is high at 0.857. This ceiling-adjacent performance limits sensitivity to small comprehension differences, since multiple-choice correctness can reflect partial knowledge, elimination strategies, or guessing. However, MCQ accuracy still provides an objective check on self-reports and supports within-item comparisons across variants.

Relationship between subjective ratings and MCQ accuracy Figure 6 relates MCQ accuracy to workers' Ease and Understanding ratings for the corresponding evidence sentence. Accuracy increases monotonically with Understanding across Ease bins, supporting perceived understanding as a meaningful subjective proxy for text-supported comprehension. While the heatmap suggests a small advantage for higher Ease, we find no clear additional benefit of Ease beyond Understanding.

Echoing prior findings that subjective difficulty is not equivalent to comprehension performance (Leroy et al., 2013), we find that even when raters report high perceived understanding, MCQ errors persist. We verified that these errors are not driven by a single unusually difficult question or a small subset of workers. Qualitative inspection suggests that many such errors occur on implication-style items rather than fact-retrieval questions. One possible explanation is that perceived understanding and MCQ accuracy diverge when readers form a good-enough local interpretation without fully encoding the precise relation, scope, or implication re-

Original	Simplified	Issue Tag votes
Mutations in an enzyme, such as PAH, are recessive since one functioning enzyme with the wild-type allele is sufficient.	Changes to DNA of a protein, such as PAH, are made in both copies of the gene that is altered, because one working copy of the gene allows the protein to function.	Vocab 1 (-2); Dense 1 (± 0); Unnatural 1 ($+1$); Takeaway 2 ($+2$);
Once the AF frequency has been estimated and tracked by a hidden Markov model approach, the resulting trend is analyzed for the purpose of detecting and characterizing the presence of circadian variation.	When the atrial fibrillation frequency is estimated and tracked by signal processing tools, the information is further reviewed to detect and describe the presence of circadian variation.	Vocab 1 (-2); Dense 1 (-2); Unnatural 1 (± 0); Reference 1 ($+1$)

Table 2: Aligned sentence pairs illustrating how simplification changes *what readers flag as difficult*. The right column reports issue-tag votes (out of three raters) for the simplified sentence, with change $\Delta = \text{simp} - \text{orig}$; negative values indicate fewer raters endorsing the issue after simplification.

quired by the question (McKoon and Ratcliff, 1992). More targeted follow-up designs, such as collecting self-explanations or think-aloud reports, could help better characterize when subjective understanding aligns with, or diverges from, successful comprehension.

Effect of simplification on MCQ accuracy We estimate simplification effects on comprehension via within-item comparisons of MCQ accuracy, using the same question set for both variants of each base text. Overall, accuracy is modestly higher in the simplified condition (orig: 0.84 vs. simp: 0.88). At the item level, we pair 154 questions that appear in both variants and compare per-item accuracy (mean over three respondents per version). 89 items (58%) show no change in accuracy, partly due to ceiling effects. Among the remainder, accuracy improves for 40 items; 9 show large gains (i.e., $\geq$2 additional correct responses out of 3). Many of these large-gain items involve correctly identifying specific terms or relations stated in the text, suggesting that simplification helps most when comprehension is limited by lexical/terminological load. 24 items show a drop in accuracy under simplification. In several cases, the simplified wording appears to broaden or shift key terms, making the evidence sentence a less direct cue for the intended inference and leaving more room for distractors, even though the question remains answerable.

For each question, we also collected two 1–5 self-reports of perceived ease of answering and confidence in the selected answer. Beyond correctness, participants report higher ease and confidence when answering the MCQs after reading simplified texts (Δquiz-ease $\approx +0.19$; Δconfidence $\approx +0.42$, simp–orig), consistent with prior work that simplification also improves subjective experience during comprehension tasks (Guidroz et al., 2025).

5. Correlates and Predictors of Difficulty

We examine how readers' sentence-level ratings of Ease and Perceived Understanding relate to both linguistic features and existing readability predictors. For the feature-based analyses, we extract variables from established toolkits (e.g., LFTK (Lee and Lee, 2023), LCA (Lu, 2012), SCA (Lu, 2010)) as well as coherence- and cohesion-oriented measures linking each sentence to its context. We exclude features with low coverage or negligible variation and use FDR correction within each outcome (Benjamini and Hochberg, 1995). For each linguistic feature, we run a separate regression predicting the sentence's mean rating, including variant and document fixed effects; features are z-scored and standard errors clustered by document.

5.1. Sentence-level Linguistic Features

Table 3 summarizes representative correlates of sentence-level Ease and Perceived Understanding after redundancy pruning. Overall, our linguistic correlates are broadly consistent with previous findings in that **lexical accessibility** emerges as the strongest signal across both outcomes (Xia et al., 2025; Jiang and Xu, 2024). In particular, *jargon count* is among the clearest correlates of lower Ease and Perceived Understanding, suggesting that biomedical terminology is a major driver of reader-perceived difficulty in our data. Several top predictors reflect local lexical bottlenecks, such as low *minimum word frequency* or high *maximum Age of Acquisition*, indicating that sentence difficulty in our data is sensitive not only to aggregate complexity but also to especially demanding lexical items.

Beyond vocabulary, features related to **information packaging** and **surface form**, such as *nominal density* and *long-distance syntactic dependencies*, are also associated with lower ratings. This is consistent with greater difficulty when in-

Table 3: Representative Linguistic Features with Effect Sizes. Coefficients are per one-standard-deviation increase in the feature, with document and variant fixed effects and document-clustered standard errors.

Feature	β_{ease}	$\beta_{und.}$
Num. of jargons[a]	-0.437	-0.372
Words $\geq$3 syllables	-0.373	-0.305
Min word frequency	0.324	0.284
Avg word frequency	0.288	0.282
Lexical sophistication (LS2)	-0.275	-0.254
Max AoA	-0.288	-0.218
Max dep. link length	-0.302	-0.209
Complex nominals / T-unit	-0.267	-0.184
Mean length of clause	-0.216	-0.141
Punctuations / sentence	-0.295	-0.206

[a] Jargon spans are extracted using the released complex-span identification checkpoint from (Jiang and Xu, 2024): https://huggingface.co/chao06/medreadme_medical_complex_span_identification_CWI.

formation is packed into complex noun phrases or spread across long spans (Pitler and Nenkova, 2008). *Punctuation density* shows a similar pattern, plausibly reflecting parentheticals or stacked clauses that fragment the sentence and increase reading difficulty.

Despite substantial overlap in their strongest correlates, Ease and Understanding differ in how they relate to specific linguistic features. We additionally fit a stacked regression with a feature$\times$outcome interaction (Ease vs. Understanding), and treat cross-outcome differences as reliable only when the interaction survives FDR correction. Several of the clearest differences involve **reference and grounding**. *Third-person singular pronoun incidence* is positively associated with both outcomes but more strongly with Ease ($\Delta = \beta_{\text{und}} - \beta_{\text{ease}} \approx -0.09$). This is in line with prior work showing that pronouns can reduce local processing cost relative to repeated names when the referent is already discourse-salient (Gordon et al., 1993), even if resolving referents can still leave uncertainty about meaning. In contrast, higher *named-entity density* is linked to lower Understanding than Ease ($\Delta \approx -0.09$), suggesting that unfamiliar entities can reduce confidence in understanding even when decoding is not especially difficult. We also find systematic differences for **lexical variability** and **syntactic packaging**. For example, high *textual lexical diversity* (McCarthy and Jarvis, 2010) is associated with lower ratings on both outcomes, but more strongly with Ease than with Understanding ($\Delta \approx +0.09$), suggesting that variability primarily increases perceived reading effort. Similarly, longer clauses and deeper dependency structures track Ease more strongly than Understanding ($\Delta \approx +0.08$), consistent with dense syntactic integration increasing processing burden without proportionally reducing confidence in comprehension.

5.2. Context / Coherence Features

We next examine whether a sentence's relationship to its surrounding context predicts perceived Ease or Understanding. We replicate the approach of (Cohen et al., 2025), considering (i) language-model chain predictability (the predictability of a sentence from its preceding context) and (ii) embedding-based proximity, operationalized as similarity to adjacent sentences or to document centroids. We additionally include (iii) cohesion indices from TAACO (Crossley et al., 2019), capturing lexical overlap with the preceding sentence, connectives, and semantic similarity.

Using the same feature-wise regression setup as above, many context metrics show negative associations with Ease/Understanding in baseline screens, but these patterns largely reflect sentence-length confounding: proximity and overlap measures are strongly length-correlated, and coefficients attenuate sharply after adding word count. After length adjustment, no coherence/proximity effects remain robust for Understanding (all small; none survive FDR), and overlap-based TAACO effects similarly disappear. The main exception is a small but stable positive association between Ease and explicit additive/connective markers (e.g., conjunctions, addition), consistent with overt discourse cues modestly improving perceived readability. Overall, context/cohesion features add limited signal beyond basic length effects in these abstracts, underscoring the importance of length control when interpreting context-based metrics. We note that the embedding-based similarity measures were computed following Cohen et al. (2025) without additional whitening or mean-centering; such post-processing may reduce hubness effects in sentence embeddings, and could be explored in future work.

5.3. Existing Readability Predictors

We lastly examine how well sentence-averaged difficulty can be predicted using existing readability predictors. We consider four commonly used **unsupervised metrics**: FKGL (Kincaid et al., 1975), ARI (Smith and Senter, 1967), SMOG (Mc Laughlin, 1969), and RSRS (Martinc et al., 2021). We also evaluate **jargon-enhanced variants** of these metrics proposed by Jiang and Xu (2024), which incorporate a weighted count of medical jargon spans into the original formulas[3]. We further

[3]We use the pre-tuned weights (α) reported in Jiang and Xu (2024).

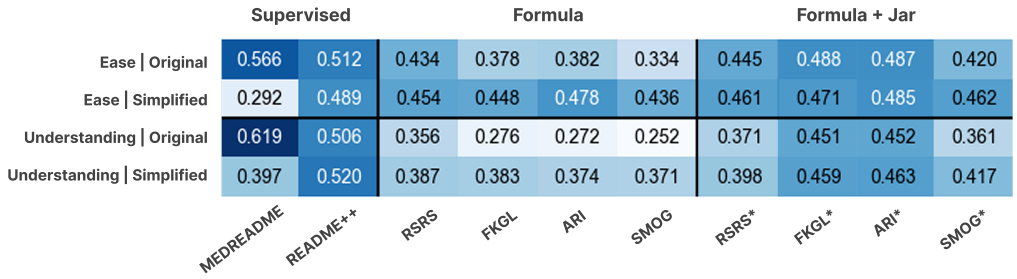

Figure 7: Alignment between readability predictors and human ratings by variant and dimension, measured with Spearman's ρ

compare against **supervised readability predictors** fine-tuned on existing readability datasets: MedReadMe (Jiang and Xu, 2024) and, as a broader-domain contrast, ReadMe++ (Naous et al., 2024). [4]

Figure 7 summarizes alignment between existing readability predictors and human ratings across both dimensions and variants. MedReadMe shows the strongest alignment on the original texts, particularly for Understanding, but its performance drops substantially on the simplified variants. We hypothesize that this reflects a transfer gap between MedReadMe's training distribution and PLABA, likely related to differences in the source and rewriting conventions underlying the simplified texts. By contrast, ReadMe++ is weaker on the originals but more stable across variants.

The impact of adding a weighted jargon-count term varies significantly by formula type. For traditional metrics (FKGL, ARI, SMOG), adding a jargon term provides a vital semantic signal that their base formulas lack, leading to substantial gains in alignment. Conversely, the neural RSRS metric see only marginal gains from explicit jargon modeling, likely because RSRS is based on word-likelihood, and already implicitly penalizes technical jargon as low-probability tokens, making manual jargon counts redundant.

The predictors also differ systematically across the two human dimensions. The base formula-based metrics align more strongly with Ease than with Understanding, consistent with their reliance on surface proxies such as length and word-form complexity. Adding jargon counts narrows this gap across all four formulas, with larger gains for Understanding than for Ease. By contrast, the MedReadMe predictor aligns more strongly with Understanding, whereas ReadMe++ is more balanced across Ease and Understanding. Overall, these patterns suggest that existing readability predictors capture part of the signal in PLABA-EVAL, but do not collapse Processing Ease and Perceived Understanding into a single construct.

[4]We use the authors' released checkpoints.

6. Conclusion

In this work, we introduced **PLABA-EVAL**, an in-context dataset and protocol for sentence-level readability assessment of biomedical abstracts and expert plain-language adaptations. By pairing sequential sentence ratings of *Processing Ease* and *Perceived Understanding* with an open-book MCQ check, the dataset supports a more fine-grained view of how readers process and understand medical text than single-score readability labels alone. Our analyses show that ease and understanding are strongly related but not interchangeable, and that existing readability predictors align inconsistently with these human judgments across dimensions and text variants. We hope PLABA-EVAL will support future work on medical readability and simplification, while encouraging more fine-grained, multi-dimensional evaluation of how texts are read and understood.

7. Limitations

A first limitation is the modest size of PLABA-EVAL. In addition, each document is annotated by only three raters, with no overlap across documents, which limits how precisely rater-specific tendencies can be separated from item difficulty. The dataset is also limited to a single genre and participant setting, so its generalizability to other medical domains, discourse structures, and reader populations remains to be established. In addition, while open-book MCQs provide an objective anchor, they capture only one form of comprehension and may miss more open-ended forms of understanding.

Finally, recent extensions to PLABA through the TREC shared task add fine-grained annotations for identifying difficult terms and characterizing replacement strategies (Ondov et al., 2026). Aligning PLABA-EVAL with these newer term-level resources would be a valuable next step for relating reader judgments to expert-annotated difficult terms and simplification strategies.

8. Acknowledgments

This work was supported by a commissioned research project from the National Institute of Information and Communications Technology (NICT), titled “Research and Development of Externally Controllable Modeling of Multimodal Information for Improving Machine Translation Accuracy” and by a Grant-in-Aid for Scientific Research from JSPS (KAKENHI Grant Number 25K03178).

9. Lay Summary

People often rely on medical texts to understand health information and make decisions, so it is important to measure whether and how a text is difficult. In practice, however, medical texts are often judged as simply “easy” or “hard.” But a sentence can be easy to read while still being hard to understand, and a sentence that takes effort to read can still make sense once the reader works through it. Moreover, people can feel that they understand a text while still missing or misunderstanding what it implies. These differences matter because they may call for different kinds of improvement, such as replacing jargon, breaking down dense sentences, clarifying how ideas are connected, or making important implications more explicit.

In this study, we created a new dataset, PLABA-EVAL, to capture these different aspects of reading difficulty in biomedical abstracts and their plain-language adaptations. Readers rated each sentence for both Processing Ease and Perceived Understanding, and also answered open-book multiple-choice questions to test what they actually understood. Our results show that ease, understanding, and comprehension do not always align, and that simplification does not improve them uniformly in every sentence. We hope this resource will help researchers build better ways of evaluating and improving medical texts so that they are not only easier to read, but also easier to truly understand.

10. Bibliographical References

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Nancy D Berkman, Stacey L Sheridan, Katrina E Donahue, David J Halpern, Anthony Viera, Karen Crotty, Audrey Holland, Michelle Brasure, Kathleen N Lohr, Elizabeth Harden, et al. 2011. Health literacy interventions and outcomes: an updated systematic review. *Evidence report/technology assessment*, (199):1–941.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735.

Trevor Cohen, Weizhe Xu, Yue Guo, Serguei Pakhomov, and Gondy Leroy. 2025. Coherence and comprehensibility: Large language models predict lay understanding of health-related content. *Journal of biomedical informatics*, 161:104758.

Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2023. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55(2):491–507.

Scott A Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods*, 51(1):14–27.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Steven J Durning, Ting Dong, Temple Ratcliffe, Lambert Schuwirth, Anthony R Artino Jr, John R Boulet, and Kevin Eva. 2016. Comparing open-book and closed-book examinations: a systematic review. *Academic medicine*, 91(4):583–599.

Peter C Gordon, Barbara J Grosz, and Laura A Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive science*, 17(3):311–347.

Theo Guidroz, Diego Ardila, Jimmy Li, Adam Mansour, Paul Jhun, Nina Gonzalez, Xiang Ji, Mike Sanchez, Sujay Kakarmath, Mathias MJ Bellaiche, et al. 2025. LLM-based text simplification and its effect on user comprehension and cognitive load. *arXiv preprint arXiv:2505.01980*.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the*

extent of agreement among raters. Advanced Analytics, LLC.

Benedetta Iavarone, Dominique Brunato, and Felice Dell'Orletta. 2021. Sentence complexity in context. In *Proceedings of the workshop on cognitive modeling and computational linguistics*, pages 186–199.

Chao Jiang and Wei Xu. 2024. MedReadMe: A systematic study for fine-grained sentence readability in medical domain. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2024, page 17293.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: from eye fixations to comprehension. *Psychological review*, 87(4):329.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report.

David A Kindig, Allison M Panzer, and Lynn Nielsen-Bohlman. 2004. *Health Literacy: A Prescription to End Confusion*. National Academies Press.

Rebecca Knowles and Chi-kiu Lo. 2025. Calibration and context in human evaluation of machine translation. *Natural Language Processing*, 31(4):1017–1041.

Bruce W Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19.

Gondy Leroy, James E Endicott, Obay Mouradi, David Kauchak, and Melissa L Just. 2012. Improving perceived and actual text difficulty for health information consumers using semi-automated methods. In *AMIA Annual Symposium Proceedings*, volume 2012, page 522.

Gondy Leroy, Stephen Helmreich, and James R Cowie. 2010. The influence of text characteristics on perceived and actual difficulty of health information. *International journal of medical informatics*, 79(6):438–449.

Gondy Leroy, David Kauchak, and Obay Mouradi. 2013. A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *International journal of medical informatics*, 82(8):717–730.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

G Harry Mc Laughlin. 1969. SMOG grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Philip M McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Gail McKoon and Roger Ratcliff. 1992. Inference during reading. *Psychological review*, 99(3):440.

Klaus G Melchers, Nadja Lienhardt, Miriam Von Aarburg, and Martin Kleinmann. 2011. Is more structure really better? a comparison of frame-of-reference training and descriptively anchored rating scales to improve interviewers'rating quality. *Personnel Psychology*, 64(1):53–87.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266.

Brian Ondov, William Xia, Kush Attal, Ishita Unde, Jerry He, and Dina Demner-Fushman. 2026. Lessons from the TREC plain language adaptation of biomedical abstracts (PLABA) track. *Journal of Biomedical Informatics*, page 104983.

Yasuhiro Ozuru, Kyle Dempsey, and Danielle S McNamara. 2009. Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and instruction*, 19(3):228–242.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195.

Jack C Richards and Richard W Schmidt. 2013. *Longman dictionary of language teaching and applied linguistics*. Routledge.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 859–866.

Elliot Schumacher, Maxine Eskenazi, Gwen Frishkoff, and Kevyn Collins-Thompson. 2016. Predicting the relative difficulty of single sentences with and without surrounding context. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1871–1881.

Hillary C Shulman, Graham N Dixon, Olivia M Bullock, and Daniel Colón Amill. 2020. The effects of jargon on processing fluency, self-perceptions, and scientific engagement. *Journal of Language and Social Psychology*, 39(5-6):579–597.

EA Smith and RJ Senter. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (US)*, pages 1–14.

Catherine Snow. 2002. *Reading for understanding: Toward an R&D program in reading comprehension*. Rand Corporation.

Tom Trabasso and Paul Van Den Broek. 1985. Causal thinking and the representation of narrative events. *Journal of memory and language*, 24(5):612–630.

Sowmya Vajjala and Ivana Lučić. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173.

Paul van den Broek, Catherine M Bohn-Gettler, Panayiota Kendeou, Sarah Carlson, and Mary Jane White. 2011. When a reader meets a text. In *Text Relevance and Learning from Text*, pages 123–139. Emerald Publishing Limited.

Paul Van den Broek, Michael Young, Yuhtsuen Tzeng, Tracy Linderholm, et al. 1999. The landscape model of reading: Inferences and the online construction of a memory representation. *The construction of mental representations during reading*, pages 71–98.

William Xia, Ishita Unde, Brian David Ondov, and Dina Demner-Fushman. 2025. Jebs: A fine-grained biomedical lexical simplification task. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17654–17666.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.