Readability Formulas, Systems and LLMs are Poor Predictors of Reading Ease

Keren Gruteke Klein[1], Shachar Frenkel[1], Omer Shubi[1], and Yevgeni Berzak[1]

[1]Faculty of Data and Decision Sciences, Technion – Israel Institute of Technology, Haifa, Israel
{gkeren,fshachar,shubi}@campus.technion.ac.il, berzak@technion.ac.il

November 4, 2025

Abstract

Methods for scoring text readability have been studied for over a century, and are widely used in research and in user-facing applications in many domains. Thus far, the development and evaluation of such methods have primarily relied on two types of offline behavioral data, performance on *reading comprehension* tests and *ratings* of text readability levels. In this work, we instead focus on a fundamental and understudied aspect of readability, real-time *reading ease*, captured with online reading measures using eye tracking. We introduce an evaluation framework for readability scoring methods which quantifies their ability to account for reading ease, while controlling for content variation across texts. Applying this evaluation to prominent traditional readability formulas, modern machine learning systems, frontier Large Language Models and commercial systems used in education, suggests that they are all poor predictors of reading ease in English. This outcome holds across native and non-native speakers, reading regimes, and textual units of different lengths. The evaluation further reveals that existing methods are often outperformed by word properties commonly used in psycholinguistics for prediction of reading times. Our results highlight a fundamental limitation of existing approaches to readability scoring, the utility of psycholinguistics for readability research, and the need for new, cognitively driven readability scoring approaches that can better account for reading ease.

1 Introduction

Over more than a century, researchers have been developing methods for automated scoring of linguistic readability of texts, a research area often referred to as Automatic Readability Assessment (ARA). ARA has been flourishing due to its societal importance in domains such as education, health care, law and media. From its early days, text readability assessment has been of central interest in psychology and education research [35, 19]. Over the past few decades, it has gained considerable traction in Natural Language Processing (NLP), where feature extraction tools and machine learning driven systems for readability scoring have been very active research areas [15, 58]. With the recent rise of user-facing Large Language Models (LLMs), there has also been a growing interest in harnessing the impressive capabilities of such models for readability scoring [57].

While the concept of readability may seem intuitive, or as [35] put it, “We know (in our hearts) what we mean by readability”, it turned out to be challenging to define which theoretical constructs are key to linguistic readability and perhaps even more so, what are its behavioral indexes. One construct that received much attention in readability research is *comprehension ease* [37]. It is typically operationalized by offline comprehension measures: querying text comprehension via reading comprehension tasks. Higher comprehension scores are then taken to indicate higher text readability. Many readability formulas developed in the 20th century took this approach by regressing text properties like sentence length, number of syllables per word, and number of complex words on reading comprehension scores [22, 16, among others]. Some of these formulas are still commonly used today.

A common alternative methodology is human ratings, where readers are asked to judge the readability level of a text according to a given annotation scheme. This approach often takes a more holistic stance towards readability, rather than targeting a specific readability construct. Human annotations of readability became prominent with the rise of modern NLP in the 21st century, in which reliance on human annotations for training text processing systems has become standard practice [58]. Following the recent dramatic advances in NLP with the introduction of LLMs, researchers have also been exploring prompting off-the-shelf LLMs to provide such annotations [57, 21].

Although reading comprehension and annotation based approaches have considerably advanced the study of readability, they do not directly speak to a fundamental aspect of a readable text, namely the degree of *reading ease* while interacting with the text in real time. Reading ease was proposed early on as a key aspect of readability [17, 35]. However, only a handful of readability studies operationalized it [42, 36, 44, 30]. At the same time, over the past few decades, a large body of work in psycholinguistics has studied reading time measures from eye tracking and other behavioral methods, as indexes of real-time processing difficulty [47], a concept which is effectively equivalent to reading ease. Nonetheless, well known linguistic predictors of reading measures [38, 49, among others] have rarely been studied in the context of readability assessment [31, 40].

In this work, we bring psycholinguistic research closer to the study of readability by introducing a *cognitive* framework for evaluating readability scoring methods. This framework focuses on real-time *reading ease* and captures it via behavioral traces of eye movements in reading which are known to index online processing difficulty. Our approach presents four advancements compared to prior work which used eyetracking in the context of readability [42, 36, 44, 30]. First, it introduces the use of eye tracking data for the *evaluation* of readability scoring methods. Second, it focuses on writing *style*, which is central to many definitions of readability [35, 19, 58], and decouples it from text content via the use of manual text simplifications. Third, it presents a *comprehensive* evaluation of a variety of prominent traditional readability formulas, modern NLP based methods, frontier LLMs, and commercial systems used in education. Finally, our study offers *robust* large-scale analyses across L1 and L2 readers, different types of reading interactions, and different textual units.

Our evaluations yield a consequential result. Despite the extensive efforts invested over decades in automated readability scoring methods, as well as the impressive linguistic capabilities of current LLMs, they all have a weak predictive power for reading facilitation in simplification, and by extension reading ease that is associated with writing style. In most cases, they are outperformed by standard word properties from the psycholinguistic literature that are known to be robust predictors of reading times: word entropy, word length, word frequency and surprisal: the negative log probability of a word in context. Surprisal tends to be the most predictive index of readability overall. These results not only call for more caution in the use of existing readability assessment methods and LLMs, but also for new, more cognitively driven methodologies for readability scoring.

2 Experimental Setup

2.1 Eye Tracking Data

Our study relies on OneStop L1&L2, a broad-coverage eye tracking dataset for English with 611 adult participants and over 4 million word tokens for which eye movements were collected. The sample includes both L1 (native) and L2 (non-native) participants from 10 different native language backgrounds. The data includes two reading regimes: ordinary reading for comprehension and information seeking. Crucially, the eye tracking data was collected over a parallel corpus of texts in their original and human-simplified forms, a property which we leverage in this work.

2.2 Online Reading Measures

We use three primary online measures from the psycholinguistic literature: average Total Fixation Duration (TF), Skip Rate (SR) and Regression Rate (RR). All three measures capture reading ease. Longer reading times and less skipping are associated with increased processing difficulty during reading [47], and with lower language proficiency [12, 1]. Increased regression rates were similarly shown to be a marker of processing difficulty and sentence reanalysis [48, 9, 4]. Correspondingly, more readable texts should have shorter Total Fixation times, more word skipping, and fewer regressions.

2.3 Readability Scoring Methods

We evaluate prominent readability scoring methods from the past 80 years. These include 6 traditional readability formulas: Flesch Reading Ease Score [22], Dale Chall Score [16], Gunning Fog Index [27], Automated Readability Index (ARI) [52], Coleman Liau Index [11] and the Flesch Kincaid Grade Level [33], 4 modern NLP-based methods: Coh-Metrix L2 Reading Index (CML2RI) [14], Crowdsourced Algorithm of Reading Comprehension (CAREC) [15], Crowdsourced Algorithm of Reading Speed (CARES) [15] and Sentence-BERT (SBERT) [13], 6 frontier LLMs: GPT-4o, GPT-5, Gemini 2.0-Flash, Gemini 2.5 Pro, Llama 3.3 70B, Claude Sonnet 4.0, and two commercial systems: Lexile (MetaMetrics) [43] and TextEvaluator (Educational Testing Service) [50] used in real-world educational settings. We compare these methods and systems against key processing difficulty measures from the psycholinguistic literature: idea density [34], integration cost [23], embedding depth, pseudo-log-likelihood (PLL) [61], entropy, and the "big three" predictors of reading times [10]: word length, word frequency and word predictability, measured with surprisal [29, 39] using an auto-regressive language model.

2.4 Evaluation Framework

Our evaluation criterion for readability scoring methods is their ability to account for the *reading facilitation* that readers experience as a result of *text simplification.* Importantly, this criterion provides experimental control for the content of the texts. We capture this facilitation effect as the change in reading ease measures between the original and the simplified version of each text:

$$\Delta\text{ReadingEase}_{E,T} = \text{ReadingEase}_{E,T_{\text{original}}} - \text{ReadingEase}_{E,T_{\text{simplified}}}$$

where E is an eye movements measure $\in \{\text{TF}, \text{SR}, \text{RR}\}$, T is a textual item, $\text{ReadingEase}_{E,T_{\text{original}}}$ is the average reading ease according to measure E across participants reading the original version of the text, and $\text{ReadingEase}_{E,T_{\text{simplified}}}$ is the average of the same measure for the participants reading the simplified version of the same text. In our analyses, T can be a sentence or a passage.

We further define a corresponding difference in readability scores for the same text, according to a readability scoring method M:

$$\Delta\text{Score}_{M,T} = \Delta\text{Score}_{M,T_{\text{original}}} - \Delta\text{Score}_{M,T_{\text{simplified}}}$$

To evaluate the quality of a readability scoring method M, we measure the predictivity of $\Delta\text{Score}_{M,T}$ for $\Delta\text{ReadingEase}_{E,T}$ for the same texts using Pearson correlation r:

$$\text{Eval}_{\text{M}} = \text{Pearson}_{\text{corr}}(\Delta\text{Score}_{M,T}, \Delta\text{ReadingEase}_{E,T})$$

3 Results

The results of our main analysis using the OneStopL1&L2 corpus are presented in Figure 1. Depicted are the Pearson correlations for readability scoring methods (ordered chronologically) and psycholinguistic measures with TF, SR and RR, at the sentence and passage levels. The corresponding pairwise statistical comparisons between the correlation coefficients of all the methods and psycholinguistic measures are presented in Appendix Figure A2. The traditional, modern, and LLM-based readability methods, as well as the two commercial systems, tend to have low and, in many cases, non-significant correlations across textual units and reading ease measures. Notably, we do not observe a pattern of improvement in readability scoring method performance over time. Entropy and the big three word properties are on par or better than readability scoring methods in most evaluations. Their main competitors are Dale-Chall and Coleman-Liau. This is likely related to the strong correlation of these two formulas with word length (see Appendix Figure A3). The advantage of the big three and entropy is especially apparent in the *RR* evaluation, where the correlations of nearly all other methods are not significant. The best predictor overall tends to be surprisal.

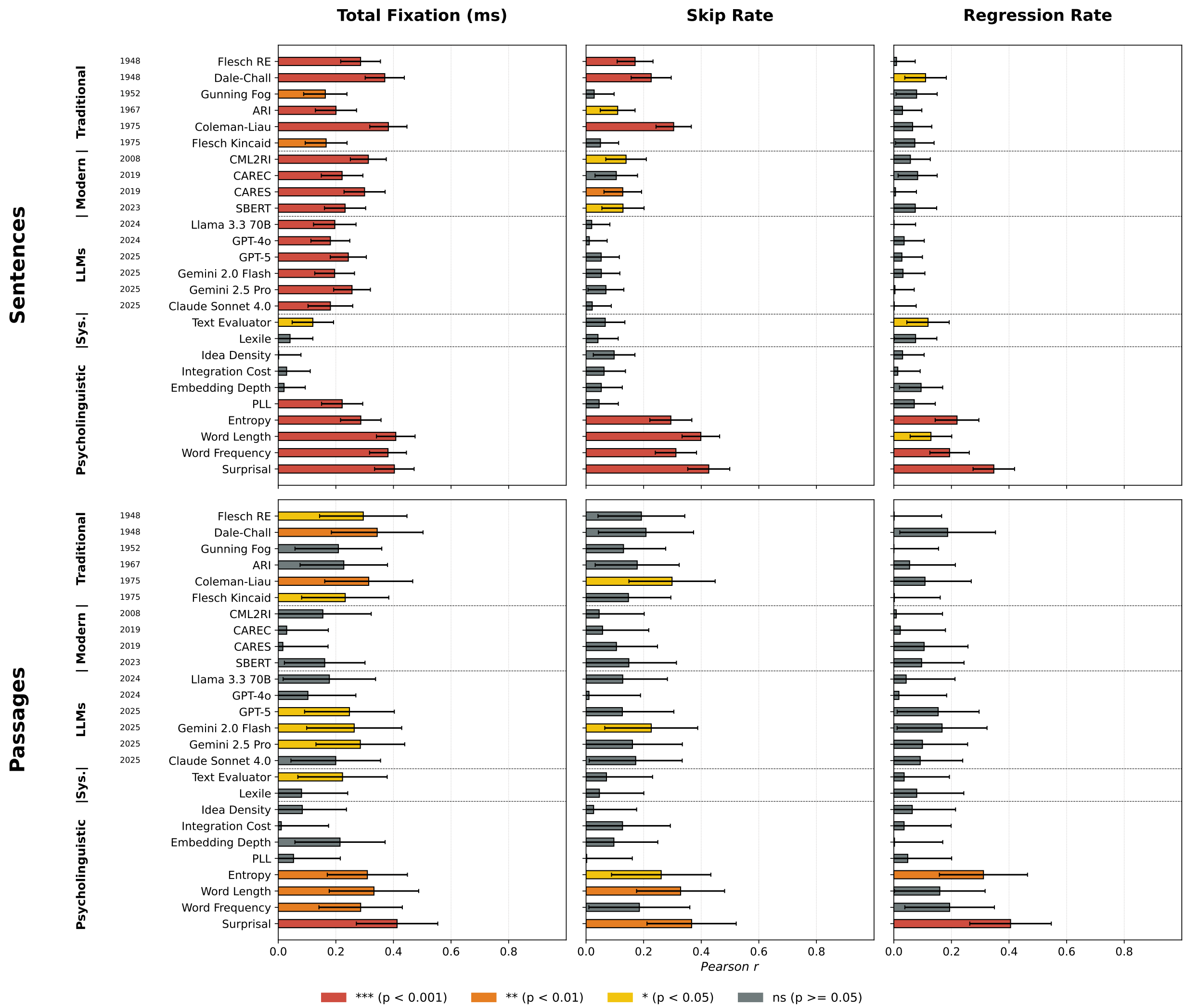

Figure 1: **Evaluation of readability scoring methods and psycholinguistic measures using their predictivity of reading ease.** Years denote the publication year of each method. The evaluations use three eye movement measures of reading ease: TF (the sum of all fixation durations on fixated words), SR (the fraction of skipped words), and RR (the number of backward saccades per word). Each bar depicts the Pearson correlation r coefficient of $\Delta\text{Score}_{M,T}$ with $\Delta\text{ReadingEase}_{E,T}$, where $\text{Score}_{M,T}$ is the readability score difference between pairs of original and simplified texts T according to method M, and $\Delta\text{ReadingEase}_{E,T}$ is the average reading ease difference according to measure E on the same pairs of texts across participants. Error bars are 95% confidence intervals. Colors represent the statistical significance level of the correlation.

3.1 Analyses of Generality

We provide additional analyses that test the robustness and generality of the main analysis results.

Reader groups and reading regimes

Prior work has shown differences in reading patterns across L1 and L2 [e.g. 62, 12, 1] and across different reading regimes [e.g. 28, 51]. Here we examine whether such differences are consequential for our evaluation. In Appendix Figure A4 we present the evaluations separately for the L1 and L2 participants, were we observe largely similar results across the two groups. Similarly, consistent results are obtained when separately examining ordinary reading for comprehension and information seeking in Appendix Figure A5. These evaluations suggest a key strength of our proposed reading ease focused evaluation framework: it is largely invariant both across different reader groups and across different reading regimes.

Surprisal as a readability measure

Are the surprisal results sensitive to the choice of language model from which the surprisal values were obtained? To answer this question, in Appendix F, we perform the main analysis with 31 additional public language models from 10 language model families. We find that the correlations are stable across models, with very moderate increases as a function of model perplexity for TF and SR and no increase for RR.

Annotation prompts for LLMs

In Appendix G we use different readability level annotation schemes and instructions provided to the LLMs via the prompt. The results are largely consistent for the different prompts.

Correlation measure

Appendix Figure A9 presents the main analysis using Spearman ρ. The results are qualitatively similar to those obtained with Pearson r.

Additional reading ease measures

Finally, in Appendix I, we perform the main analysis with 8 additional online reading measures, and the offline measure of reading speed. The main results largely hold for these additional measures.

3.2 Comparison to reading ease without control for text content

In Appendix Figure A13 we present the main evaluation, along with an evaluation of the direct prediction of reading ease measures for all the texts in the OneStopL1&L2 data. We observe that the lack of control for content in the latter evaluation does affect the evaluation outcomes: for most methods, and especially the traditional formulas, the content-controlled evaluations yield lower correlations compared to the corresponding evaluations without such control. This suggests that these methods are substantially affected by the text topic, i.e. by factors that go beyond writing style. The big three and entropy, on the other hand, tend to obtain stable results across both evaluation methods.

4 Discussion

Departing from a century-old tradition in ARA which relies on reading comprehension outcomes and readability level annotations, we focus on the construct of reading ease, and introduce reading facilitation as a result of simplification as a cognitive benchmark for the evaluation of readability scoring methods. We use this framework to evaluate a wide range of prominent readability scoring approaches, as well as key psycholinguistic measures that have been previously linked to online processing difficulty. Based on this analysis, we find that existing methods are poor predictors of reading ease, outperformed by entropy and the "big three" text properties, word length, frequency and especially surprisal.

Our results have several important implications for readability research and applications. First, they provide empirical evidence for the drawbacks of existing readability methodologies in capturing the real-time experience of readers with texts, and call for caution in the adoption of existing methods in high-stakes settings. This empirical evidence reflects both conceptual and practical limitations of these approaches in quantifying readability, which we discuss below.

While comprehension ease is indeed central to readability, its practical assessments have a number of limitations. First, comprehension scores depend not only on the difficulty of the text but also on the difficulty of the reading comprehension questions, which typically differ when comparing the readability levels of different texts. Further, with a limited number of questions, it is impossible to probe reading comprehension exhaustively, and highly challenging to do so reliably. Fundamentally, it is a measure that cannot fully account for readability: differences in reading and comprehension ease do not have to lead to differences in reading comprehension performance, and vice versa, differences in reading comprehension performance do not necessarily stem from differences in text readability. Indeed, several studies found that substantial reading ease differences in adult L1 speakers, translate to negligible differences in reading comprehension performance [60, 26].

The alternative, human readability labeling approach includes absolute and comparative labeling of the readability level of different texts (e.g., graded readers and pairwise annotation of relative readability). Level annotations have been obtained both from experts, such as teachers, and from non-expert annotators, for example, via crowdsourcing. While such annotations have been widely adopted for training and evaluating readability scoring systems in NLP, they too have fundamental drawbacks [58]. First, labeling is operationalized via highly non-trivial annotation tasks, whose inter-annotator agreement is largely unknown.The texts often differ in content, which, according to Vajjala "... leads us to question what the ARA models learn – is it a notion of text complexity, or topical differences among texts?"[58].

Eye tracking enables a viable alternative to the reading comprehension and labeling approaches, that obviates many of their limitations. Perhaps its key advantage is that differently from comprehension questions and human judgments, which are both *offline* behavioral signals, it provides *online* measures of how readers experience the text. While in this work we highlight reading ease, such measures are also intimately related to online language comprehension processes and, as such, provide a multifaceted measure for readability.

Eye tracking based methodologies, however, also have limitations. First, they are not the only source of cognitive information that is relevant for readability. For example, post-reading text recall [41] may also be of major interest. More broadly, reading ease is best viewed as an additional tool, thus far underexplored, which complements existing approaches. Eye movements also pose practical challenges. They are inherently noisy, and there is a great deal of variability in eye movement patterns between readers. In this work, we tackle this challenge by computing summary statistics and averaging over a large number of participants. This requires collecting data from many participants, which is not always feasible. Summary statistics also involve the loss of potentially valuable information from the full eye movement trajectories.

The results of this study support a broader adoption of surprisal for readability assessment. We note that surprisal is conceptually related to left-to-right cloze [53]. While the current primary use of cloze is estimation of subjective probabilities in psycholinguistics, its bidirectional version was originally proposed as a method for quantifying readability [56] and bidirectional cloze tests were used for fitting the regression coefficients of several traditional readability formulas [52, 33, 11]. Differently from cloze, surprisal is not limited by the need to collect large-scale data from human participants, and can be computed automatically using readily available language models. More broadly, the results underscore the need for the development of new methods of readability scoring that will better account for reading ease.

Finally, [35] asked in the mid 70s "Who cares about readability?". Their answer, based on the analysis of publications from 1941 to 1973, was "Psychologists don't, not anymore". Over the past decades, the psychology of reading and psycholinguistics have greatly advanced our understanding of reading processes and their behavioral traces. Although much of this research is highly relevant to readability assessment and to applications of readability in education and NLP, it has not sufficiently impacted these areas. Our study highlights some of the potential benefits of increasing the interaction of psycholinguistics with the study of readability and the application of psycholinguistic methods and insights in applied and educational settings.

5 Materials and Methods

We use OneStopL1&L2, a combination of two datasets, OneStop [3], which is publicly available, and OneStopL2 data, which we newly collected. Both are broad-coverage datasets of eye movements in English reading with identical textual materials and an identical experimental design, collected with an EyeLink 1000 Plus eye tracker (SR Research). Crucially, these datasets use a parallel corpus of texts in their original and simplified forms, which makes them uniquely suited for our study.

5.1 Textual Materials

OneStopL1&L2 uses textual materials from the OneStopQA dataset [2, 59]. They consist of 30 Guardian news articles with 4-7 paragraphs (162 paragraphs in total) from the News Lessons section of the English language-learning portal onestopenglish.com by Macmillan Education. Each article was simplified by a staff member of onestopenglish.com from its original "Advanced" version to a simplified "Elementary" version. The simplification method is "intuitive", relying on experience and subjective judgment. Each paragraph has three multiple-choice reading comprehension questions. The questions and answers are identical for both difficulty levels of each paragraph. To support the current study, we manually created an additional sentence-level segmentation and alignment between the two versions of each paragraph. This alignment enables analyses not only at the passage level, but also for individual sentences. Appendix Table A2 presents summary statistics of the textual materials across the two text difficulty levels.

5.2 Participants

OneStopL1&L2 has 611 adult participants, 360 L1 participants in OneStopL1 and 251 L2 participants in OneStopL2. 329 of the participants read texts for comprehension. The remaining 282 participants read in an information seeking regime described below. The mean participant age of L1 participants is 22.8, and their mean English age of acquisition (AoA) is 0.4. The mean participant age of L2 participants is 29.6 and their mean English AoA is 9.6. The L2 participants have the following 10 L1 backgrounds, with the number of participants per L1 in parentheses: Arabic (24), Chinese (36), French (13), Hebrew (30), Japanese (16), Korean (27), Portuguese (28), Russian (35), Spanish (36), and Vietnamese (6).

5.3 Eye Movement Data

Each participant is assigned to one of three 10-article batches (54 paragraphs) in one of two between-subject reading regime conditions, ordinary reading for comprehension or information seeking. The texts are presented paragraph by paragraph. After each paragraph, the participant has to answer one of the three multiple-choice reading comprehension questions for the paragraph on a new screen, without the ability to return to the paragraph. In the information seeking regime, the question is also presented before reading the paragraph. Each paragraph in a given article is shown to the participant in either the original or simplified version, selected at random.

The data is counterbalanced such that each participant reads 27 original and 27 simplified paragraphs overall and approximately the same number of original and simplified paragraphs within each article. On average, in each reading regime (reading for comprehension, information seeking), each paragraph is read by 204 participants, 102 in the original difficulty level and 102 in the simplified level. Overall, the eye tracking data contains 4,289,977 word tokens over which eye movement data was collected, 2,372,088 for the original texts and 1,917,889 for their simplified versions.

5.4 Simplification Effects in the Data

An important prerequisite for the applicability of reading facilitation-centered evaluations is the presence of simplification effects on reading measures in the dataset. This question was investigated for L1 readers in OneStop by [26], who found a robust effect of simplification on reading times (12ms in TF) and significant reading speed increases in 42% of the participants. The effects are even more pronounced in OneStopL2 where 66% of the participants show significant reading speed increases and an overall effect of 32ms in TF. These results suggest that both the L1 and the L2 parts of the data are suitable for the development of reading facilitation-centered evaluations.

5.5 Eye Movement Measures

In the main analysis, we use three key summary measures of the eye movement trajectory.

1. **Total Fixation Duration (TF)** TF is the sum of all the fixation durations on a word. We analyze average TF, taking into consideration only words that were not skipped.

2. **Skip Rate (SR)** The fraction of words that were skipped (i.e. were not fixated).

3. **Regression Rate (RR)** The number of backward saccades per word.

In Appendix I we provide analyses for additional online measures: First Fixation (FF), Mean Fixation Duration (FD), number of Fixations (NF), first pass measures of Gaze Duration (fpGD), Skip Rate (fpSR) and Regression Rate (fpRR), Gaze Duration (GD), and higher pass Fixation Duration (hpFD). We further include reading speed, an offline measure that can be obtained without eye tracking. Appendix I includes the definitions of all the measures.

5.6 Readability Scoring Methods

We evaluate 18 widely used traditional and modern readability scoring methods, LLMs and commercial systems.

Traditional Formulas

We use 6 prominent methods from the 20th century. These are linear regression formulas that use a small set of word property features, with coefficients fitted using reading comprehension data from English L1 speakers. Two commonly used features in these formulas are word length, a heuristic for measuring word complexity, and sentence length, a heuristic measure of grammatical complexity. Appendix Table A3 presents the formulas and their interpretations. While these formulas were originally developed for passages, they are all applicable to single sentences. We used the formula implementations in the `textstat` library version 0.7.4.

Modern Methods and Systems

Over the past two decades, the rise of NLP has led to the introduction of readability scoring methods and systems that take advantage of automated linguistic analysis of texts and supervised machine learning. Similarly to traditional formulas, modern scoring methods were typically developed based on data from passages, but can be applied to single sentences.

- *Coh-Metrix L2 Reading Index (CML2RI)* [14] A linear regression formula with three features, word frequency, sentence syntactic similarity and word overlap between adjacent sentences. The regression coefficients of CML2RI were fitted using L2 reading comprehension (cloze) scores.

- *Crowdsourced Algorithm of Reading Comprehension (CAREC)* [15] A linear regression formula from 13 linguistic features to L1 speakers' pairwise judgments of text difficulty ("Which text is easier to understand"), for pairs of different texts.

- *Crowdsourced Algorithm of Reading Speed (CARES)* [15] A linear regression formula from 9 linguistic features to L1 speakers' pairwise judgments of reading speed ("Which text did you read more quickly"), for pairs of different texts.
- *Sentence-BERT (SBERT)* [13] A transformer-based readability assessment model that predicts the readability score of texts, based on a corpus spanning grades 3–12.

We use the implementation of the above methods in the Automatic Readability Tool for English (ARTE).

Large Language Models

LLMs have revolutionized NLP, and among their many uses, they are increasingly used in readability research and applications [57, 21, among others]. Here, we evaluate 6 frontier LLMs through their APIs. The first model is open-weights and the remaining 5 are closed-source.

- Llama 3.3 70B Versatile (Meta, 2024).
- *GPT-4o*: `gpt-4o-2024-08-06` (OpenAI, 2024).
- *GPT-5*: `gpt-5-2025-08-07` (OpenAI, 2025).
- *Gemini 2.0 Flash*: `gemini-2.0-flash-001` (Google, 2025).
- *Gemini 2.5 Pro*: `gemini-2.5-pro` (Google, 2025).
- *Claude Sonnet 4*: `claude-sonnet-4-20250514` (Anthropic, 2025).

We evaluate these models using two different annotation frameworks provided in the model prompt:

1. *Grade-level*: predict a school grade level (1–12). School grades are commonly used for readability level annotation [58].
2. *Score*: assign a readability score from 1 (easy) to 100 (hard), following prior work in [57].

Following [57], we also experiment with a second variant for each of the prompts above, which further includes additional guidance to consider sentence structure, discourse structure, vocabulary, and clarity when scoring the text. For the main analysis, we use the *Grade-level* prompt with additional guidance. The other prompts yield similar results. The prompts and the results for all four prompt variants are provided in Appendix G.

Commercial Systems

We consider two prominent commercial systems for readability scoring used in educational settings.

- *Lexile Text Analyzer* A commercial readability assessment system developed by MetaMetrics https://hub.lexile.com/text-analyzer/. In the spirit of traditional readability formulas, Lexile readability scoring is based on regression from word frequency ("semantic component") and sentence length ("syntactic component") to reading comprehension outcomes [43]. Lexile scores are commonly used in the US K-12 educational system https://metametricsinc.com/products/state-eogeoc-assessments/. Lexile scores were obtained using the Lexile API.

- *TextEvaluator* A commercial readability scoring system by the Educational Testing Service (ETS) https://textevaluator.ets.org/TextEvaluator/. The system uses a variety of textual features extracted with NLP tools, regressed against human annotations of text readability level [50]. TextEvaluator scores were obtained by manually querying the TextEvaluator web interface.

5.7 Psycholinguistic Measures

[31] introduced the idea of using theoretically motivated measures from the psycholinguistic literature for predicting text difficulty. They used several variants of 4 such measures for the classification of sentence difficulty level for pairs of original and simplified sentences. Here we use similar features:

- *Idea Density* The ratio of ideas or propositions to words [34]. This ratio is expected to be lower in more readable texts. To compute idea density, we use the Computerized Propositional Idea Density Rater (CPIDR) 3.2, implemented in the `pycpidr` library version 0.3.0.

- *Integration Cost* This measure is rooted in the dependency locality theory [23, 24], which ties processing difficulty to the distance between syntactic heads and dependents. More readable texts should have a lower integration cost. We use `icy-parses`, to compute sentence integration cost, and evaluate the average integration cost measure.

- *Embedding Depth* Syntactic embedding depth is a measure that reflects expected processing difficulty due to memory load. We examine average embedding depth across words. We implement this measure using syntactic parses obtained with the small English `spaCy` model `en_core_web_sm-3.8.0`.

- *Surprisal* Surprisal theory [29, 39] relates processing difficulty to the predictability of the word in its context. Surprisal is defined as $-\log_2(p(w|context))$, where *context* are the words preceding the current word w in the word sequence *seq*. The average surprisal for a word sequence is: $\text{surp\_avg} = \frac{1}{|\text{seq}|}\sum_{w\in\text{seq}} -\log_2(p(w|context))$. In the main analysis, we use the Pythia-70M language model [5] whose surprisals correlate well with reading times [25].

- *Entropy* of the next word given a context is another information-theoretic measure that was linked to processing difficulty.
  (Shannon) Entropy is defined as $-\sum_{w\in seq} p(w|context)\log_2(p(w|context))$. We use mean per-word entropy. This measure was not used by [31], but we include it as it has been shown to contribute to reading times above and beyond surprisal [7, 46].

- *Pseudo Log Likelihood (PLL)* A method for scoring words under bidirectional Masked Language Models (MLMs) [61]. This method is related to surprisal, and is conceptually similar to bidirectional cloze tasks which were suggested for readability scoring [56]. We use *PLL-word-l2r* [32], with `minicons` https://github.com/kanishkamisra/minicons, with the best performing model from [32], `roberta-large`.

In addition to the above measures, similarly to [31], we include average word length and average word frequency. Together with predictability (measured using surprisal), these measures form the "big three" word property predictors of reading times [10].

- *Word Length* is not only a strong predictor of reading times, but also a common predictor used in traditional readability measures (see Appendix Table A3). We measure word length in characters, excluding punctuation.

- *Word Frequency* is similarly a robust predictor of reading times. We use frequency counts from Wordfreq [54], coded as unigram surprisal $-\log_2(p(w))$.

We use `text-metrics library` version 1.1.7 for calculating per-word surprisal, length, and frequency. Additional details on the libraries used for computing readability measures are provided in Appendix M.

5.8 Language Models for Surprisal Estimation

The estimation of surprisal in our main analysis is conducted using the Pythia 70m model. To analyze the robustness of surprisal as a readability measure, in Appendix F, we use 31 additional publicly available language models from the GPT-2, GPT-J, GPT-Neo, Pythia, OPT, Mistral, Gemma, Llama-2, RWKV, and Mamba families, ranging from 70 million to 13 billion parameters. The complete list of models is provided in Appendix Table A1.

5.9 Statistical Tests

The main results in Figure 1 are reported with 95% confidence intervals computed using bootstrap over texts (200 resamples with replacement). Pairwise comparisons of correlations between readability scoring methods, reader groups and reading conditions are performed with Steiger's (1980) two-sided test for dependent overlapping correlations. Additional details are provided in Appendix A.

5.10 Code and Data Availability

The code for this paper is publicly available at https://github.com/lacclab/Readability-Evaluation-Using-Reading-Ease. OneStop [3] is publicly available at https://osf.io/2prdq/. OneStopL2 will be made publicly available upon paper publication.

References

[1] Yevgeni Berzak and Roger Levy. Eye movement traces of linguistic knowledge in native and non-native reading. *Open Mind*, 7:179–196, 2023.

[2] Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.507. URL `https://aclanthology.org/2020.acl-main.507`.

[3] Yevgeni Berzak, Jonathan Malmaud, Omer Shubi, Yoav Meiri, Ella Lion, and Roger Levy. Onestop: A 360-participant english eye-tracking dataset with different reading regimes. *PsyArXiv preprint*, 2025.

[4] Klinton Bicknell and Roger Levy. Why readers regress to previous words: A statistical analysis. In *Proceedings of the annual meeting of the Cognitive Science Society*, volume 33, 2011.

[5] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

[6] Cati Brown, Tony Snodgrass, Susan J Kemper, Ruth Herman, and Michael A Coving-

ton. Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior research methods*, 40(2):540–545, 2008.

[7] Benedetta Cevoli, Chris Watkins, and Kathleen Rastle. Prediction as a basis for skilled reading: Insights from modern language models. *Royal Society open science*, 9(6):211837, 2022.

[8] Joon Suh Choi and Scott A Crossley. Advances in readability research: A new readability web app for english. In *2022 International Conference on Advanced Learning Technologies (ICALT)*, pages 1–5. IEEE, 2022.

[9] Charles Clifton Jr, Adrian Staub, and Keith Rayner. Eye movements in reading words and sentences. *Eye movements*, pages 341–371, 2007.

[10] Charles Clifton Jr, Fernanda Ferreira, John M Henderson, Albrecht W Inhoff, Simon P Liversedge, Erik D Reichle, and Elizabeth R Schotter. Eye movements in reading and information processing: Keith rayner's 40 year legacy. *Journal of Memory and Language*, 86:1–19, 2016.

[11] Meri Coleman and Ta Lin Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283, 1975.

[12] Uschi Cop, Denis Drieghe, and Wouter Duyck. Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. *PloS one*, 10(8): e0134008, 2015.

[13] Scott Crossley, Joon Suh Choi, Yanisa Scherber, and Mathis Lucka. Using large language models to develop readability formulas for educational settings. In *International Conference on Artificial Intelligence in Education*, pages 422–427. Springer, 2023.

[14] Scott A Crossley, Jerry Greenfield, and Danielle S McNamara. Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3):475–493, 2008.

[15] Scott A Crossley, Stephen Skalicky, and Mihai Dascalu. Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, 42(3-4):541–561, 2019.

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

[17] Edgar Dale and Jeanne S. Chall. The concept of readability. *Elementary English*, 26(1): 19–26, 1949.

[18] Birk Diedenhofen and Jochen Musch. cocor: A comprehensive solution for the statistical comparison of correlations. *PloS one*, 10(4):e0121945, 2015.

[19] William H. DuBay. The principles of readability. 2004.

[20] Olive Jean Dunn and Virginia Clark. Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association*, 64(325):366–377, 1969.

[21] Asma Farajidizaji, Vatsal Raina, and Mark Gales. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339, Torino, Italia, May 2024. ELRA and ICCL. URL `https://aclanthology.org/2024.lrec-main.815/`.

[22] Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

[23] Edward Gibson. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76, 1998.

[24] Edward Gibson. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain/MIT Press*, 2000.

[25] Keren Gruteke Klein, Yoav Meiri, Omer Shubi, and Yevgeni Berzak. The effect of surprisal on reading times in information seeking and repeated reading. In Libby Barak and Malihe Alikhani, editors, *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 219–230, Miami, FL, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.conll-1.17. URL `https://aclanthology.org/2024.conll-1.17/`.

[26] Keren Gruteke Klein, Omer Shubi, Shachar Frenkel, and Yevgeni Berzak. The effect of text simplification on reading fluency and reading comprehension in l1 english speakers, Feb 2025. URL `osf.io/preprints/psyarxiv/dhk8c_v1`.

[27] Robert Gunning. The technique of clear writing, 1952.

[28] Michael Hahn and Frank Keller. Modeling task effects in human reading with neural network-based attention. *Cognition*, 230:105289, 2023.

[29] John Hale. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*, 2001.

[30] Nora Hollenstein, Itziar Gonzalez-Dios, HiTZ Center-IXA, Lisa Beinborn, and Lena Jäger. Patterns of text readability in human and predicted eye movements. *CogALex-VII*, page 1, 2022.

[31] David M Howcroft and Vera Demberg. Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 958–968, 2017.

[32] Carina Kauf and Anna Ivanova. A better way to do masked language model scoring. *arXiv preprint arXiv:2305.10588*, 2023.

[33] J Peter Kincaid, RP Fishburne, RL Rogers, and BS Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Chief of Naval Technical Training*, 1975.

[34] Walter Kintsch. Notes on the structure of semantic memory. In Endel Tulving and W. Donaldson, editors, *Organization of Memory*. Academic Press, 1972.

[35] Walter Kintsch and Douglas Vipond. Reading comprehension and readability in educational practice and psychological theory. In *Perspectives on Memory Research*, pages 329–365. Psychology Press, 1985.

[36] George R. Klare, Earl H. Shuford, and William H. Nichols. The relationship of style difficulty, practice, and ability to efficiency of reading and to retention. *Journal of Applied Psychology*, 41(4):222–226, 1957.

[37] George R Klare et al. *The measurement of readability*. Iowa State University Press Ames, 1963.

[38] Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. Length, frequency, and predictability effects of words on eye movements in reading. *European journal of cognitive psychology*, 16(1-2):262–284, 2004.

[39] Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008.

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

[41] Eric Martínez, Francis Mollica, and Edward Gibson. Poor writing, not specialized concepts, drives processing difficulty in legal language. *Cognition*, 224:105070, 2022.

[42] Howard Y. McClusky. A quantitative analysis of the difficulty of reading materials. *Journal of Educational Research*, 28:276–282, 1934. URL `https://api.semanticscholar.org/CorpusID:145566694`.

[43] MetaMetrics. Lexile® framework for reading development and validity evidence. Technical Manual, January 2022. URL `https://partnerhelp.metametricsinc.com/image/Lexile%20Framework%20for%20Reading%20Validity%20Evidence_2022.pdf#page=10.44`.

[44] Shingo Nahatame. Text readability and processing effort in second language reading: A computational and eye-tracking investigation. *Language learning*, 71(4):1004–1043, 2021.

[45] Byung-Doh Oh and William Schuler. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350, 2022. URL `https://api.semanticscholar.org/CorpusID:255096260`.

[46] Tiago Pimentel, Clara Meister, Ethan G Wilcox, Roger P Levy, and Ryan Cotterell. On the effect of anticipation on reading times. *Transactions of the Association for Computational Linguistics*, 11:1624–1642, 2023.

[47] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998.

[48] Keith Rayner, Kathryn H Chace, Timothy J Slattery, and Jane Ashby. Eye movements as reflections of comprehension processes in reading. *Scientific studies of reading*, 10(3): 241–255, 2006.

[49] Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121, 2024.

[50] Kathleen M Sheehan, Irene Kostin, Diane Napolitano, and Michael Flor. The textevaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal*, 115(2):184–209, 2014.

[51] Omer Shubi and Yevgeni Berzak. Eye movements in information-seeking reading. In *Proceedings of the annual meeting of the cognitive science society*, volume 45, 2023.

[52] Edgar A Smith and RJ Senter. *Automated readability index*, volume 66. Aerospace Medical Research Laboratories, Aerospace Medical Division, 1967.

[53] Nathaniel Smith and Roger Levy. Cloze but no cigar: The complex relationship between

cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

[54] Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. Luminosoinsight/wordfreq: v2.2, October 2018. URL https://doi.org/10.5281/zenodo.1443582.

[55] James H Steiger. Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245, 1980.

[56] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

[57] Sean Trott and Pamela Rivière. Measuring and modifying the readability of English texts with GPT-4. In Matthew Shardlow, Horacio Saggion, Fernando Alva-Manchego, Marcos Zampieri, Kai North, Sanja Štajner, and Regina Stodden, editors, *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 126–134, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.tsar-1.13. URL https://aclanthology.org/2024.tsar-1.13/.

[58] Sowmya Vajjala. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, 2022.

[59] Sowmya Vajjala and Ivana Lučić. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In Joel Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis, editors, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0535. URL https://aclanthology.org/W18-0535.

[60] Sowmya Vajjala and Ivana Lučić. On understanding the relation between expert annotations of text readability and target reader comprehension. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 349–359, 2019.

[61] Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*, 2019.

[62] Veronica Whitford and Debra Titone. Second-language experience modulates first-and second-language word frequency effects: Evidence from eye movement measures of natural paragraph reading. *Psychonomic bulletin & review*, 19(1):73–80, 2012.

Appendix

A Pairwise Statistical Comparisons

To evaluate whether two correlations obtained from the same set of texts differ significantly, we use the *test for overlapping correlations based on dependent groups* implemented in the `cocor.dep.groups.overlap` function of the `cocor` package [18]. The description of the test in the official documentation of `cocor`:

> Performs a test of significance for the difference between two correlations based on dependent groups (e.g., the same group). The two correlations are overlapping, i.e., they have one variable in common. The comparison is made between r_{jk} and r_{jh}. The function tests whether the correlations between j and k (r_{jk}) and between j and h (r_{jh}) differ in magnitude.

Following this definition, we denote:

- r_{jk}: correlation between variables j and k
- r_{jh}: correlation between variables j and h
- r_{kh}: intercorrelation between k and h
- n: number of paired observations (texts)

We employ Steiger's (1980) two-sided test [55], which provides a modification of Dunn and Clark's z test [20]. In all analyses, the unit of observation is the **text**, with $n = 162$ for passage-level and $n = 790$ for sentence-level analyses.

A.1 Comparison between readability formulas

We assess whether two readability formulas M_1 and M_2 differ in how strongly their scores correlate with observed reading-ease differences across texts. Formally,

$$\begin{aligned} r_{jh} &= \mathrm{Eval}_{M_1} = \mathrm{Pearson}_{\mathrm{corr}}\Big(\Delta\mathrm{Score}_{M_1,T}, \Delta\mathrm{ReadingEase}_{E,T}\Big) \\ r_{jk} &= \mathrm{Eval}_{M_2} = \mathrm{Pearson}_{\mathrm{corr}}\Big(\Delta\mathrm{Score}_{M_2,T}, \Delta\mathrm{ReadingEase}_{E,T}\Big) \\ r_{hk} &= \mathrm{Pearson}_{\mathrm{corr}}\Big(\Delta\mathrm{Score}_{M_1,T}, \Delta\mathrm{Score}_{M_2,T}\Big) \end{aligned}$$

where $\Delta\mathrm{ReadingEase}_{E,T}$ is the difference in reading-ease measure E (e.g., Total Fixation Duration) between the original and simplified versions of text T, and $\Delta\mathrm{Score}_{M,T}$ is the difference in readability scoring method M between the same pair of text T.

A.2 Comparison between L1 and L2 readers

We use the same test to compare the correlations obtained from two dependent groups that read the same set of texts: L1 and L2 readers. Specifically,

$$
\begin{aligned}
r_{jh} &= \text{Eval}_M^{(\text{L1})} = \text{Pearson}_{\text{corr}}\Big(\Delta\text{Score}_{M,T}, \Delta\text{ReadingEase}_{E,T}^{(\text{L1})}\Big) \\
r_{jk} &= \text{Eval}_M^{(\text{L2})} = \text{Pearson}_{\text{corr}}\Big(\Delta\text{Score}_{M,T}, \Delta\text{ReadingEase}_{E,T}^{(\text{L2})}\Big) \\
r_{hk} &= \text{Pearson}_{\text{corr}}\Big(\Delta\text{ReadingEase}_{E,T}^{(\text{L1})}, \Delta\text{ReadingEase}_{E,T}^{(\text{L2})}\Big)
\end{aligned}
$$

The correlations are considered dependent because both groups were assessed on the same texts, and the readability scoring method serves as the overlapping variable.

A.3 Comparison between reading regimes

Finally, we test whether the relationship between text readability scores and reading ease differs across reading regimes (*ordinary reading* for comprehension vs. *information-seeking*). Here,

$$
\begin{aligned}
r_{jh} &= \text{Eval}_M^{(\text{Ordinary})} = \text{Pearson}_{\text{corr}}\Big(\Delta\text{Score}_{M,T}, \Delta\text{ReadingEase}_{E,T}^{(\text{Ordinary})}\Big) \\
r_{jk} &= \text{Eval}_M^{(\text{InfoSeek})} = \text{Pearson}_{\text{corr}}\Big(\Delta\text{Score}_{M,T}, \Delta\text{ReadingEase}_{E,T}^{(\text{InfoSeek})}\Big) \\
r_{hk} &= \text{Pearson}_{\text{corr}}\Big(\Delta\text{ReadingEase}_{E,T}^{(\text{Ordinary})}, \Delta\text{ReadingEase}_{E,T}^{(\text{InfoSeek})}\Big)
\end{aligned}
$$

B Main Analysis: Pairwise Comparisons of the Reading Ease Predictivity of Readability Scoring Methods and Psycholinguistic Measures

In Figure A2, each cell (i, j) compares the evaluation of readability formula M_i with the evaluation of readability formula M_j. Specifically, the cell is colored according to the statistical significance of comparing $\text{Eval}_{M_i} = \text{Pearson}_{\text{corr}}\Big(\Delta\text{Score}_{M_i,T}, \Delta\text{ReadingEase}_{E,T}\Big)$ with $\text{Eval}_{M_j} = \text{Pearson}_{\text{corr}}\Big(\Delta\text{Score}_{M_j,T}, \Delta\text{ReadingEase}_{E,T}\Big)$, where $\Delta\text{ReadingEase}_{E,T}$ denotes the difference in the reading ease measure E averaged across participants, between pairs of original and simplified versions of text T, and $\Delta\text{Score}_{M,T}$ is the corresponding difference of readability scores from method M.

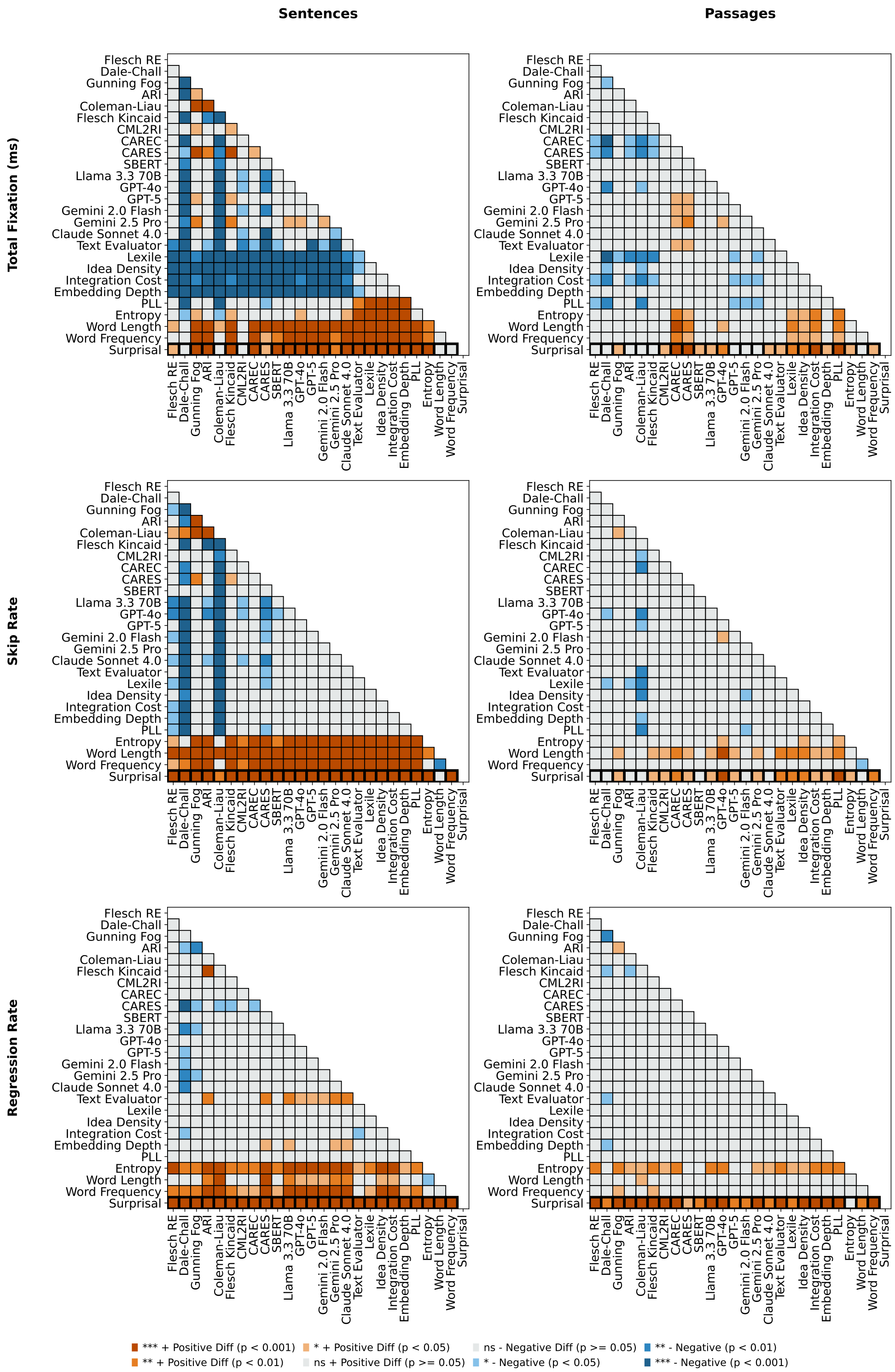

Figure A2: **Pairwise statistical comparisons of the reading ease predictivity of readability scoring methods and psycholinguistic measures.** Each cell (i, j) compares the evaluation of readability formula M_i with the evaluation of readability formula M_j. Significance is assessed using Steiger's (1980) test for dependent overlapping correlations (see Section A.1). Cell color indicates the p value of the test, with "Positive" and "Negative" denoting higher or lower correlations for the row measure relative to the column measure, respectively.

C Correlations between Different Readability Scoring Methods and Psycholinguistic Measures

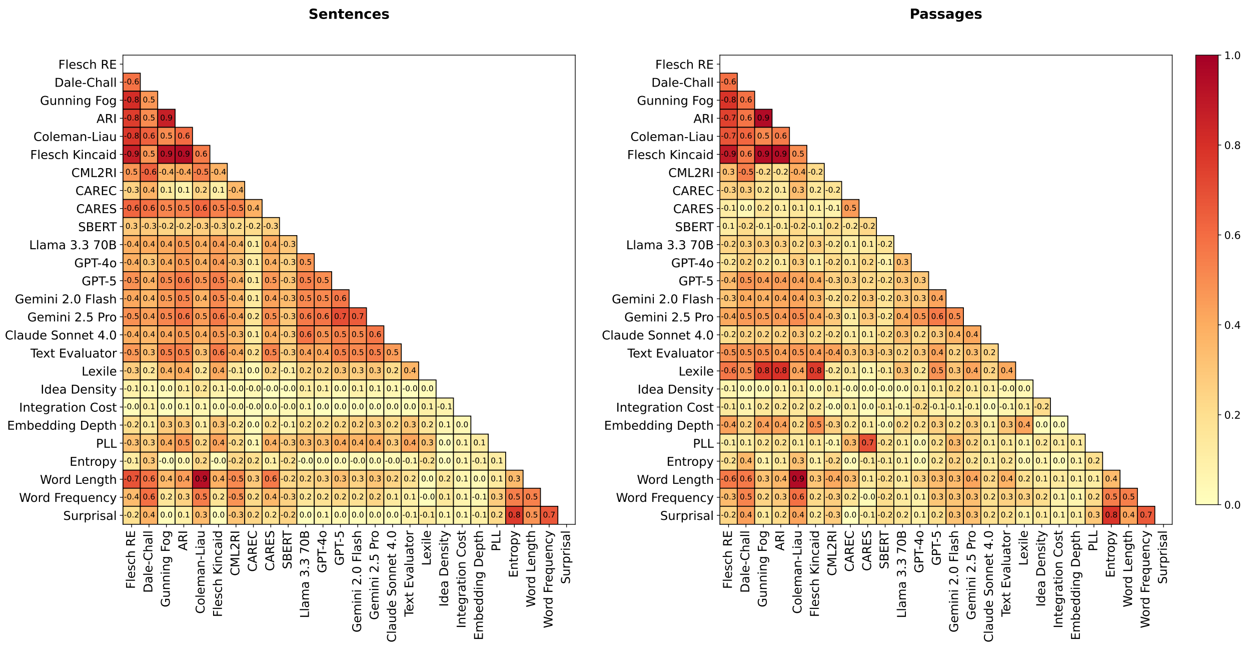

Figure A3: **Pairwise Pearson correlation between different readability scoring methods and psycholinguistic measures**. Each cell (i, j) is colored by the Pearson r correlation between the readability score differences $\Delta\text{Score}_{M_i,T}$ and $\Delta\text{Score}_{M_j,T}$ produced by methods M_i and M_j, measured across all the textual items T in OneStopL1&L2.

D Analysis of Generality: Results across Different Reader Groups: L1 and L2

Figure A4 presents the main analysis separately for English L1 and English L2 speakers.

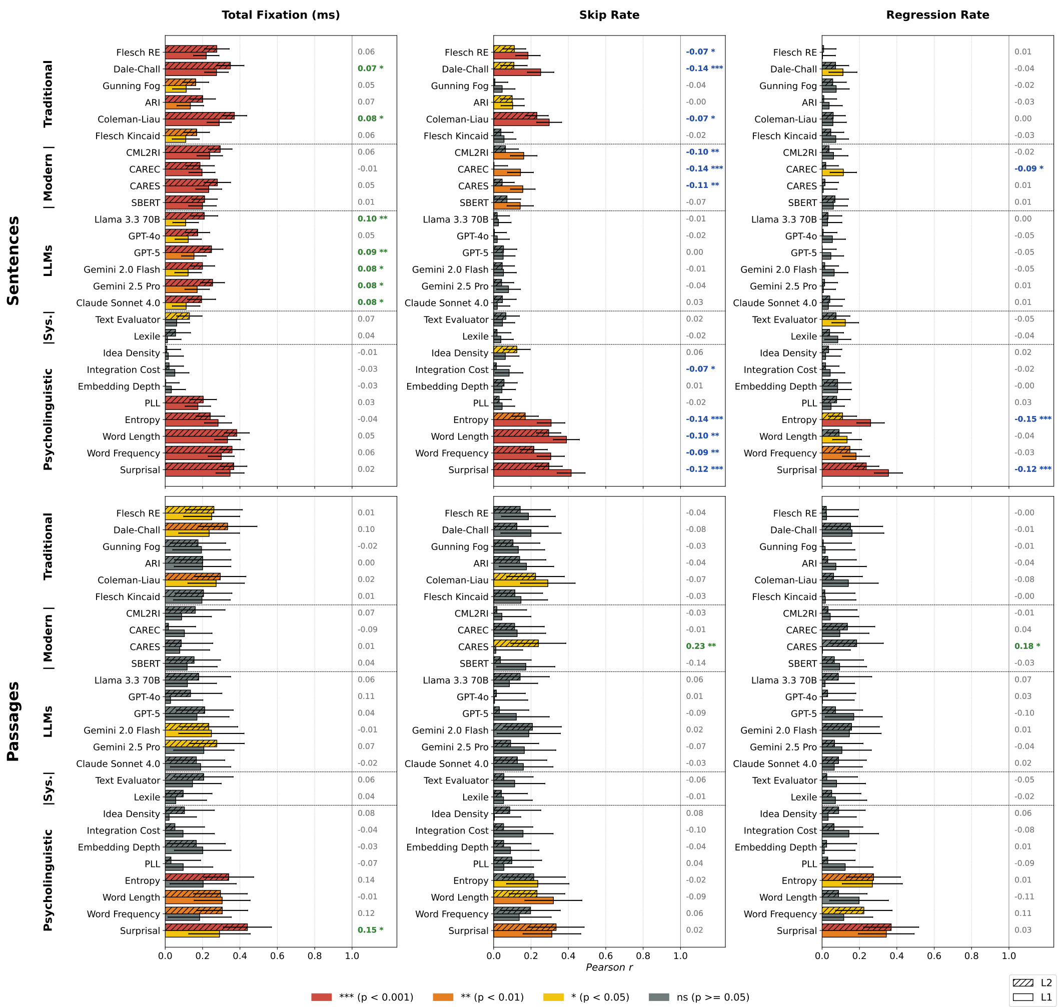

Figure A4: **Reading ease predictivity of readability scoring methods and psycholinguistic measures for *L1* and *L2* speakers.** Each pair of bars shows the Pearson correlation r for *L2* readers (top, striped) and *L1* readers (bottom). To the right of each pair, we display the difference in correlations (L2-L1) together with its statistical significance, assessed using a Steiger correlation test (see Section A.2). Significant differences are colored in green if *L2* > *L1* and in blue if *L1* > *L2*. Bar colors indicate the statistical significance level of the correlation.

E Analysis of Generality: Results for Different Reading Regimes

Figure A5 presents the main analysis separately for two between subjects reading regimes: ordinary reading for comprehension and information seeking (where the participants read the question before reading the passage).

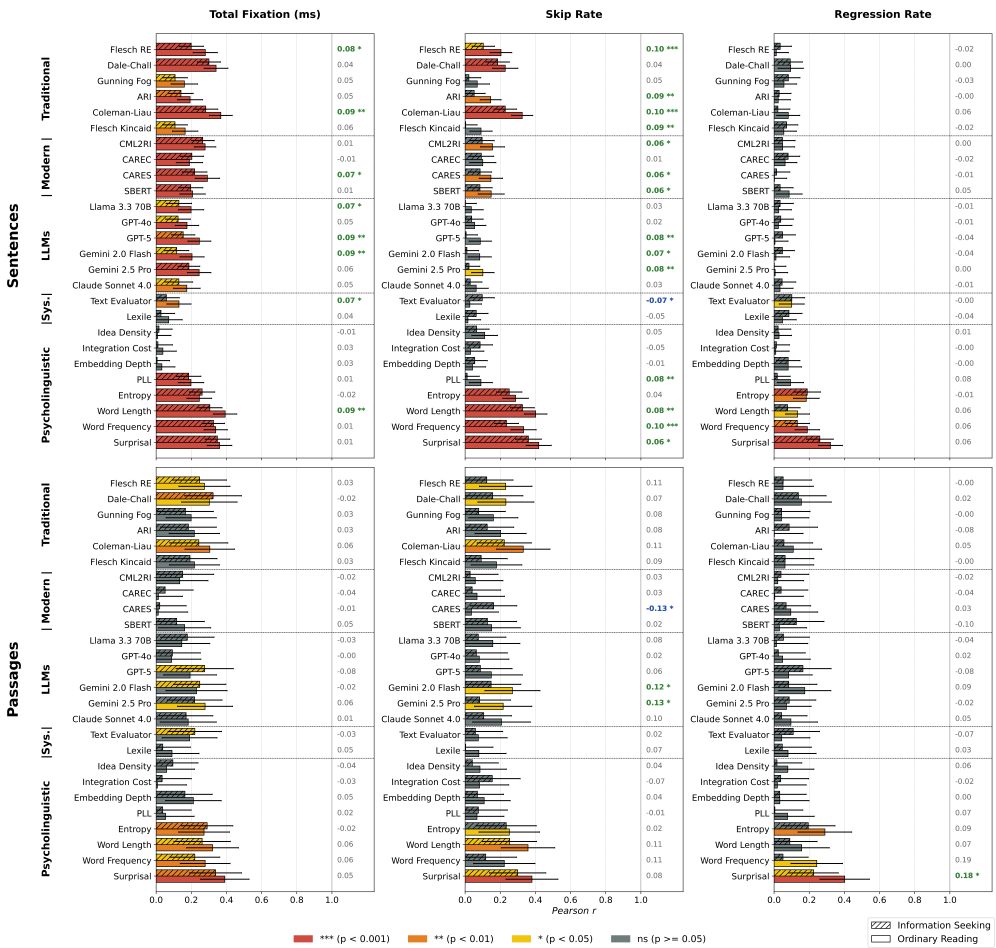

Figure A5: **Reading ease predictivity of readability scoring methods and psycholinguistic measures for *ordinary reading* and *information seeking* reading regimes.** Each pair of bars shows the Pearson correlation r for *information seeking* readers (top, striped) and *ordinary reading* readers (bottom). To the right of each pair, we display the difference in correlations (ordinary reading − information seeking) together with its statistical significance, assessed using a Steiger correlation test (see Section A.3). Significant differences are colored in green if ***ordinary reading > information seeking*** and in blue if ***information seeking > ordinary reading***. Bar colors indicate the statistical significance level of the correlation.

F Surprisal as a Readability Measure: Robustness of Reading Ease Predictivity to the Choice of Language Model

In Figures A6 and A7, right panels, we compare the predictivity of surprisal for reading ease using our proposed content-controlled evaluation across different language models, as a function of model log perplexity. We use Pythia 70M from the main analysis, and 31 additional publicly available language models (see list in Table A1). At the sentence level, we observe very moderate predictivity increases as function of model log perplexity for TF ($\beta = 0.0024$, $p < 0.001$) and SR ($\beta = 0.0009$, $p < 0.001$), and no significant correlation for RR ($\beta = -0.0001$, $p > 0.05$), using the linear model $r \sim$ Perplexity. Similar results are obtained at the passage level with TF ($\beta = 0.0019$, $p < 0.001$), SR ($\beta = 0.0024$, $p < 0.001$), and RR ($\beta = -0.0009$, $p > 0.05$). This suggests that surprisal is a robust predictor of reading ease, irrespective of the language model.

Note that previous work has shown that recent language models with lower perplexity exhibit reduced predictive power for reading measures [45, 49, 25, among others]. We reproduce this result in the left panels, where reading ease is measured across all texts in the OneStop corpus, without controlling for text content. Taken together with the result above, these outcomes further highlight an advantageous consequence of our content-controlled evaluation setting.

#	Model Family	Model Name	Model Size	PPL	Model Identifier
1	Pythia	Pythia 70M	70M	48.99	EleutherAI-pythia-70m
2		Pythia 160M	160M	32.87	EleutherAI-pythia-160m
3		Pythia 410M	410M	22.61	EleutherAI-pythia-410m
4		Pythia 1B	1B	19.43	EleutherAI-pythia-1b
5		Pythia 1.4B	1.4B	17.70	EleutherAI-pythia-1.4b
6		Pythia 2.8B	2.8B	15.83	EleutherAI-pythia-2.8b
7		Pythia 6.9B	6.9B	14.63	EleutherAI-pythia-6.9b
8	GPT-2	GPT-2 117M	117M	28.29	gpt2
9		GPT-2 345M	345M	21.35	gpt2-medium
10		GPT-2 774M	774M	18.76	gpt2-large
11		GPT-2 1558M	1558M	16.90	gpt2-xl
12	GPT-J	GPT-J 6B	6B	14.77	EleutherAI-gpt-j-6B
13	GPT-Neo	GPT-Neo 125M	125M	33.20	EleutherAI-gpt-neo-125M
14		GPT-Neo 1.3B	1.3B	19.58	EleutherAI-gpt-neo-1.3B
15		GPT-Neo 2.7B	2.7B	17.53	EleutherAI-gpt-neo-2.7B
16	Llama-2	Llama-2 7B	7B	9.05	meta-llama/Llama-2-7b-hf
17		Llama-2 13B	13B	8.40	meta-llama/Llama-2-13b-hf
18	OPT	OPT 350M	350M	25.54	facebook/opt-350m
19		OPT 1.3B	1.3B	18.32	facebook/opt-1.3b
20		OPT 2.7B	2.7B	16.74	facebook/opt-2.7b
21		OPT 6.7B	6.7B	15.08	facebook/opt-6.7b
22	Mistral	Mistral-v0.1 7B	7B	9.21	mistralai/Mistral-7B-v0.1
23		Mistral-v0.3 7B	7B	9.31	mistralai/Mistral-7B-v0.3
24	Gemma	Gemma 7B	7B	11.55	google/gemma-7b
25		Gemma-2 9B	9B	12.08	google/gemma-2-9b
26		Recurrent-Gemma 9B	9B	11.77	google/recurrentgemma-9b
27	RWKV	RWKV-4 169M	169M	28.06	RWKV/rwkv-4-169m-pile
28		RWKV-4 430M	430M	21.54	RWKV/rwkv-4-430m-pile
29	Mamba	Mamba 370M	370M	19.92	state-spaces-mamba-370m-hf
30		Mamba 790M	790M	17.35	state-spaces-mamba-790m-hf
31		Mamba 1.4B	1.4B	15.92	state-spaces-mamba-1.4b-hf
32		Mamba 2.8B	2.8B	14.42	state-spaces-mamba-2.8b-hf

Table A1: Language models used for extracting surprisal: family, model name, model size, and perplexity (PPL) measured on the OneStop Eye Movements dataset, with their corresponding Hugging Face (`https://huggingface.co/`) identifiers. The main analysis uses Pythia 70M.

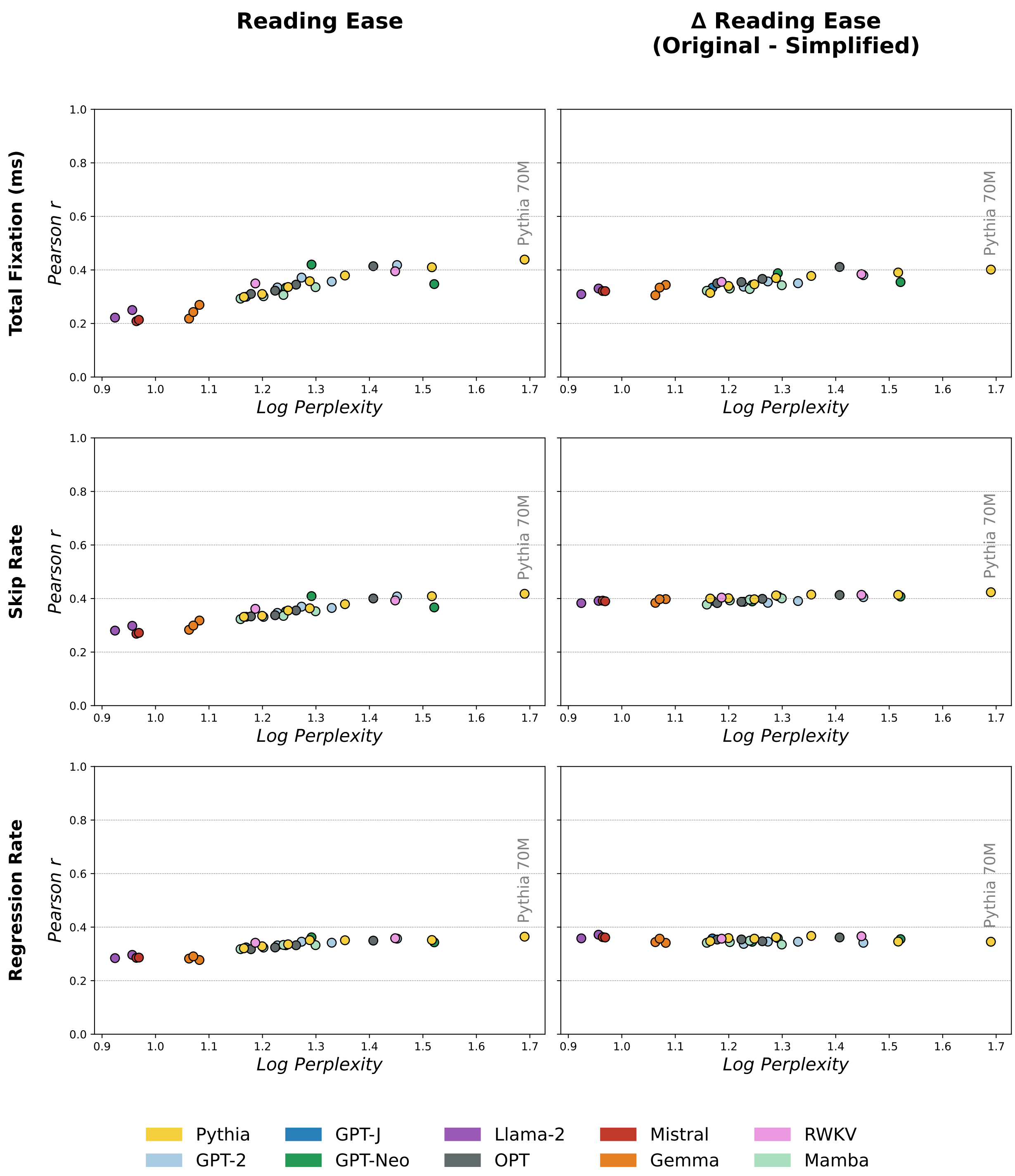

Figure A6: **Surprisal as a readability measure: robustness of reading ease predictivity to the choice of language model.** Analysis at the ***sentence*** level. Right: our evaluation which controls for text content, $\text{Eval}_{\text{Surprisal},\text{LM}_i} = \text{Pearson}_{\text{corr}}\left(\Delta\text{Surprisal}_{\text{LM}_i,T}, \Delta\text{ReadingEase}_{E,T}\right)$. Left prediction of reading ease without such control, $\text{Eval}_{\text{Surprisal},\text{LM}_i} = \text{Pearson}_{\text{corr}}\left(\text{Surprisal}_{\text{LM}_i,T}, \text{ReadingEase}_{E,T}\right)$ where Surprisal are mean surprisal values per textual unit according to language model LM_i. Colors represent the model family. Model sizes range from 70 million to 13 billion parameters. The main analysis uses Pythia 70M.

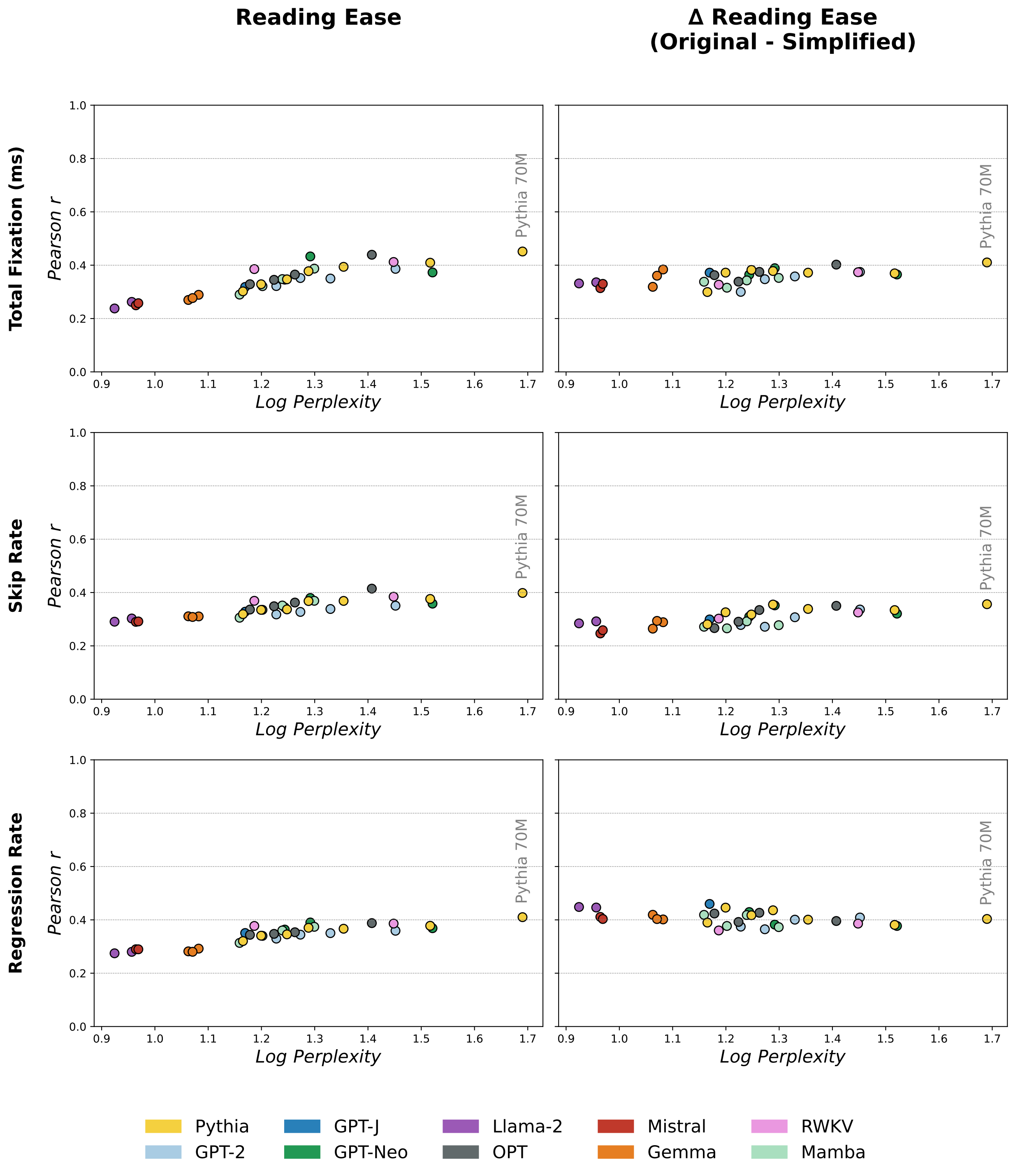

Figure A7: **Surprisal as a readability measure: robustness to the choice of language model.** Analysis at the ***passage*** level. Right: our evaluation which controls for text content, $\mathrm{Eval}_{\mathrm{Surprisal,LM}_i} = \mathrm{Pearson}_{\mathrm{corr}}\left(\Delta\mathrm{Surprisal}_{\mathrm{LM}_i,T}, \Delta\mathrm{ReadingEase}_{E,T}\right)$. Left prediction of reading ease without such control, $\mathrm{Eval}_{\mathrm{Surprisal,LM}_i} = \mathrm{Pearson}_{\mathrm{corr}}\left(\mathrm{Surprisal}_{\mathrm{LM}_i,T}, \mathrm{ReadingEase}_{E,T}\right)$ where Surprisal are mean surprisal values per textual unit according to language model LM_i. Colors represent the model family. Model sizes range from 70 million to 13 billion parameters. The main analysis uses Pythia 70M.

G Analysis of Generality: Robustness to Prompt Variants for LLMs

Figure A8 presents evaluations for LLMs using the following four prompt variants. The first two prompts are similar to the prompts used in [57]. The third and fourth prompts are introduced in this work and use similar wording but with a different output range of school grades, common in the readability literature.

- *Score*:

  Read the text below.

  Then, indicate the readability of the text, on a scale from 1 (very easy to read and understand) to 100 (very difficult to read and understand).

  Please answer with a single number in the range 1 to 100.

  <Text>

- *Score + Criteria*:

  Read the text below.

  Then, indicate the readability of the text, on a scale from 1 (very easy to read and understand) to 100 (very difficult to read and understand).

  To determine your score, consider factors such as the complexity of sentence structure, the complexity of discourse structure, the vocabulary used, and the overall clarity of the text.

  Please answer with a single number in the range 1 to 100.

  <Text>

- *Grade*:

  Read the text below.

  Then, indicate the readability level of the text by specifying the school grade level (1–12) for which the text would be most appropriate.

  Please answer with a single number in the range 1 to 12.

  <Text>

- *Grade + Criteria*:

    Read the text below.

    Then, indicate the readability level of the text by specifying the school grade level (1–12) for which the text would be most appropriate.

    To determine your score, consider factors such as the complexity of sentence structure, the complexity of discourse structure, the vocabulary used, and the overall clarity of the text.

    Please answer with a single number in the range 1 to 12.

    <Text>

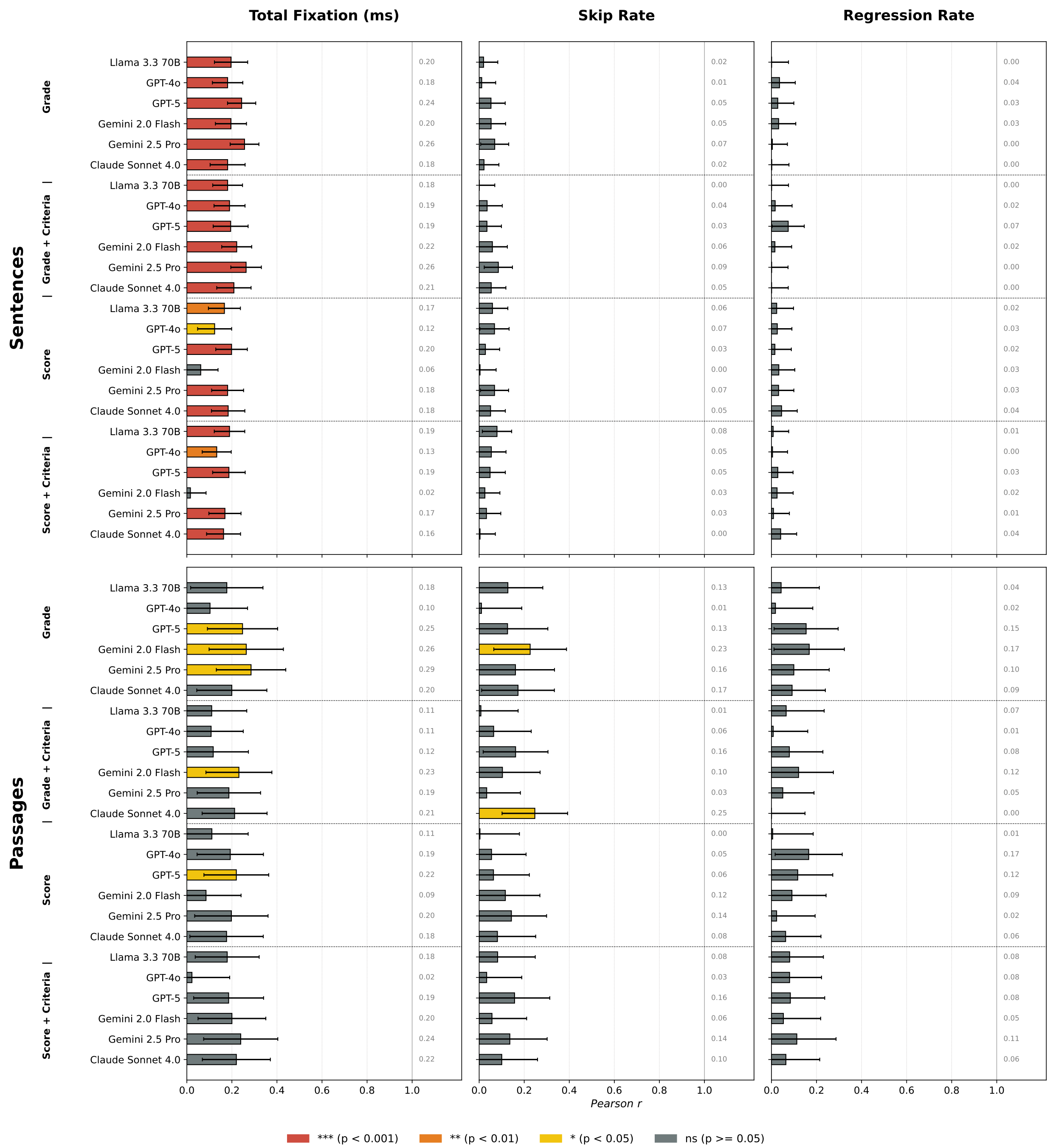

Figure A8: **Reading ease predictivity of LLMs using different prompts.** Presented are Pearson correlation r coefficients. Error bars are 95% confidence intervals. Colors represent the statistical significance level of the correlation.

H Analysis of Generality: Correlation Measure

In Figure A9 we present side by side the results of the main analysis which uses Pearson correlation r and the same analysis using the **Spearman** ρ rank correlation coefficient.

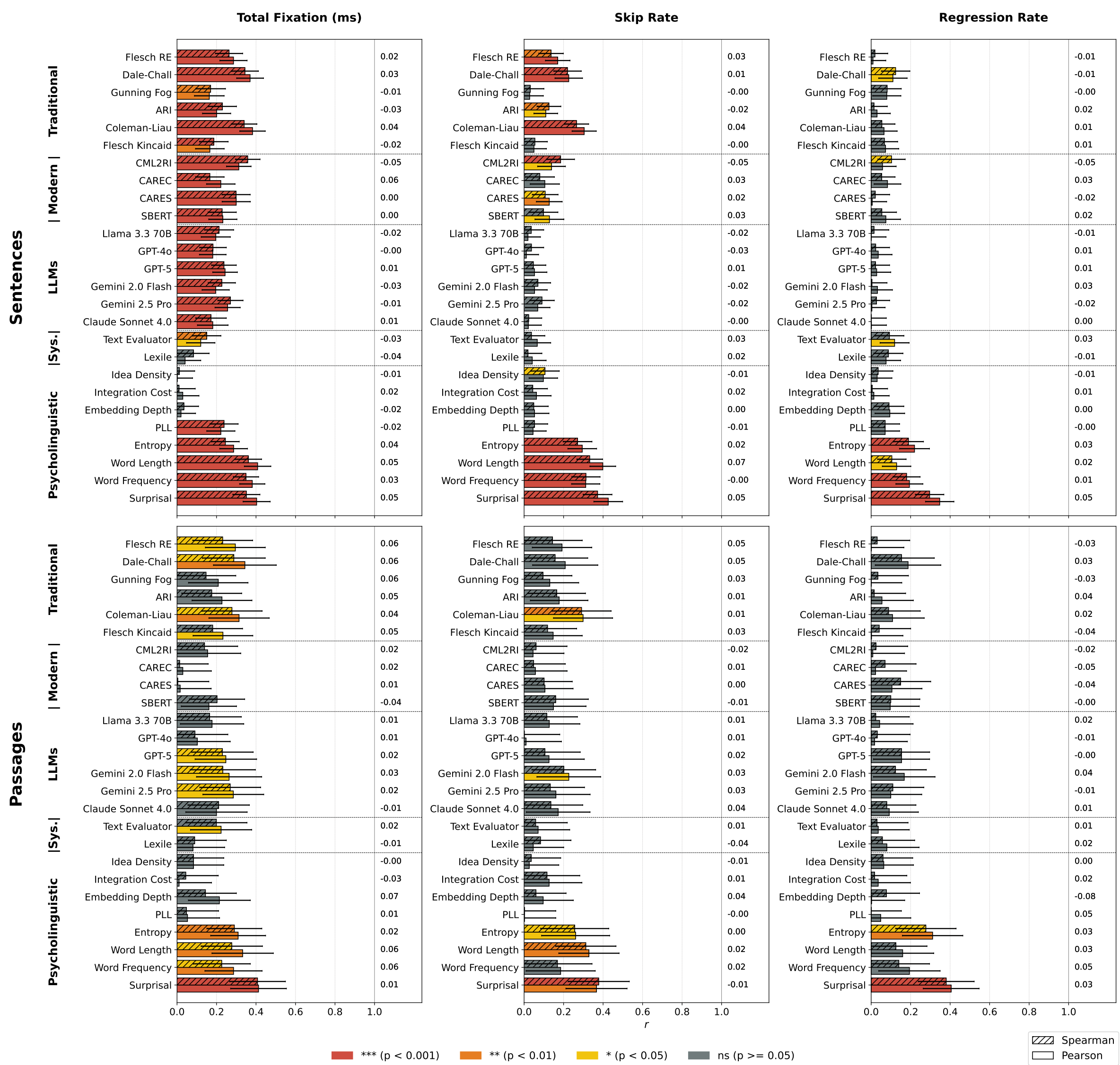

Figure A9: **Pearson and Spearman correlations.** Each pair of bars shows the Spearman ρ correlation (top, striped) and Pearson r correlation (bottom). To the right of each pair, we display the difference between the correlation coefficients (Pearson - Spearman). Bar colors indicate the statistical significance level of the correlation.

I Analysis of Generality: Additional Reading Ease Measures

The main analysis uses three reading ease measures, Total Fixation, Skip Rate and Regresssion Rate. Figure A10, Figure A11, and Figure A12 present the same analysis with eight additional eye-tracking measures from the psycholinguistic literature, as well as for reading speed.

Single fixation measures and number of fixations:

- **First Fixation (FF)** The mean duration of the first fixation on a word, considering only words that were not skipped.
- **Fixation Duration (FD)** The mean duration of a fixation on a word.
- **Number of Fixations (NF)** The mean number of fixations per word.

First pass measures:

- **First Pass Gaze Duration (fpGD)** The mean time from first entering a word to first leaving it, during first pass reading.
- **First Pass Skip Rate (fpSR)** The fraction of words that were not fixated, during first pass reading.
- **First Pass Regression Rate (fpRR)** The number of saccades per word that go backward, during first pass reading.

Later measures and reading speed:

- **Gaze Duration (GD)** The mean time from first entering a word to first leaving it.
- **Higher Pass Fixation Duration (hpFD)** The sum of all fixations on a word during second and higher pass readings, averaged across words.
- **Reading Speed (RS)** The number of words read per second. Note that differently from the measures above, RS is an offline measure that can be obtained without eye tracking. Assuming the text is known, it requires only the total reading time of the text.

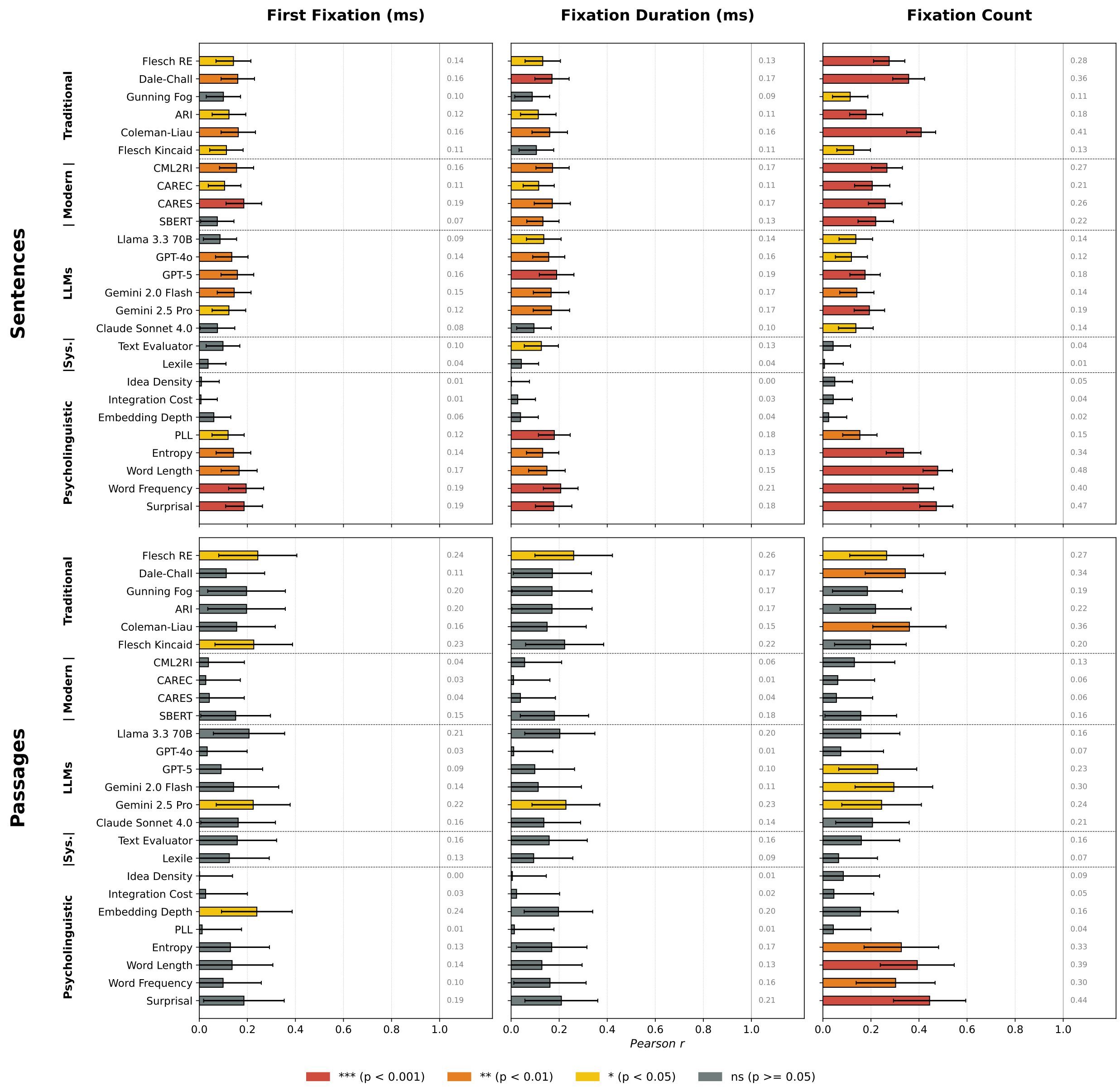

Figure A10: **Evaluation of readability scoring methods and psycholinguistic measures using their predictivity of reading ease.** Presented are Pearson correlation r coefficients for **First Fixation duration (FF), Fixation Duration (FD) and Number of Fixations (NF).** Error bars are 95% confidence intervals. Colors represent the statistical significance level of the correlation.

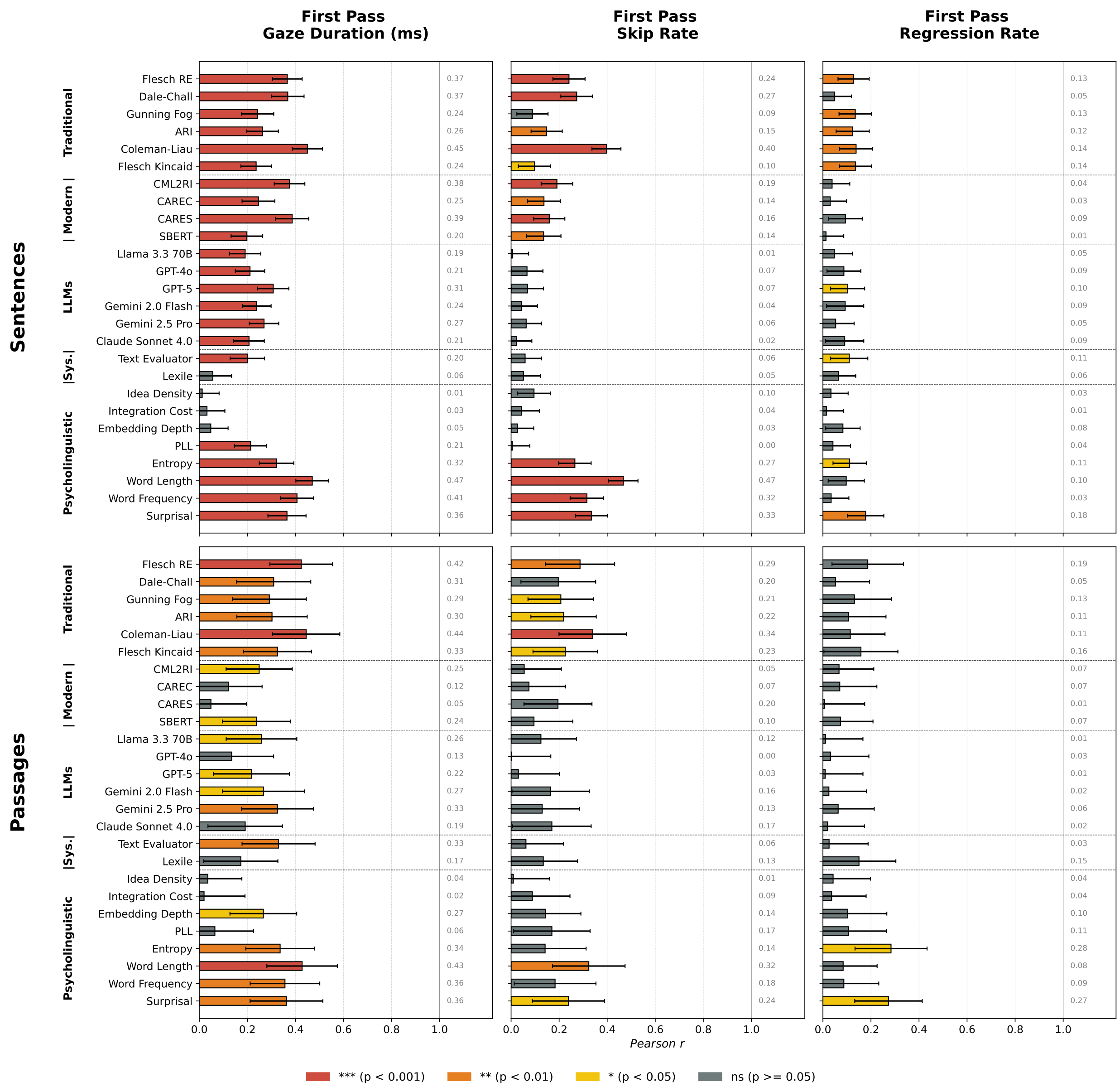

Figure A11: **Evaluation of readability scoring methods and psycholinguistic measures using their predictivity of reading ease.** Presented are Pearson correlation r coefficients for **first pass Gaze Duration (fpGD), first pass Skip Rate (fpSR) and first pass Regression Rate (fpRR).** Error bars are 95% confidence intervals. Colors represent the statistical significance level of the correlation.

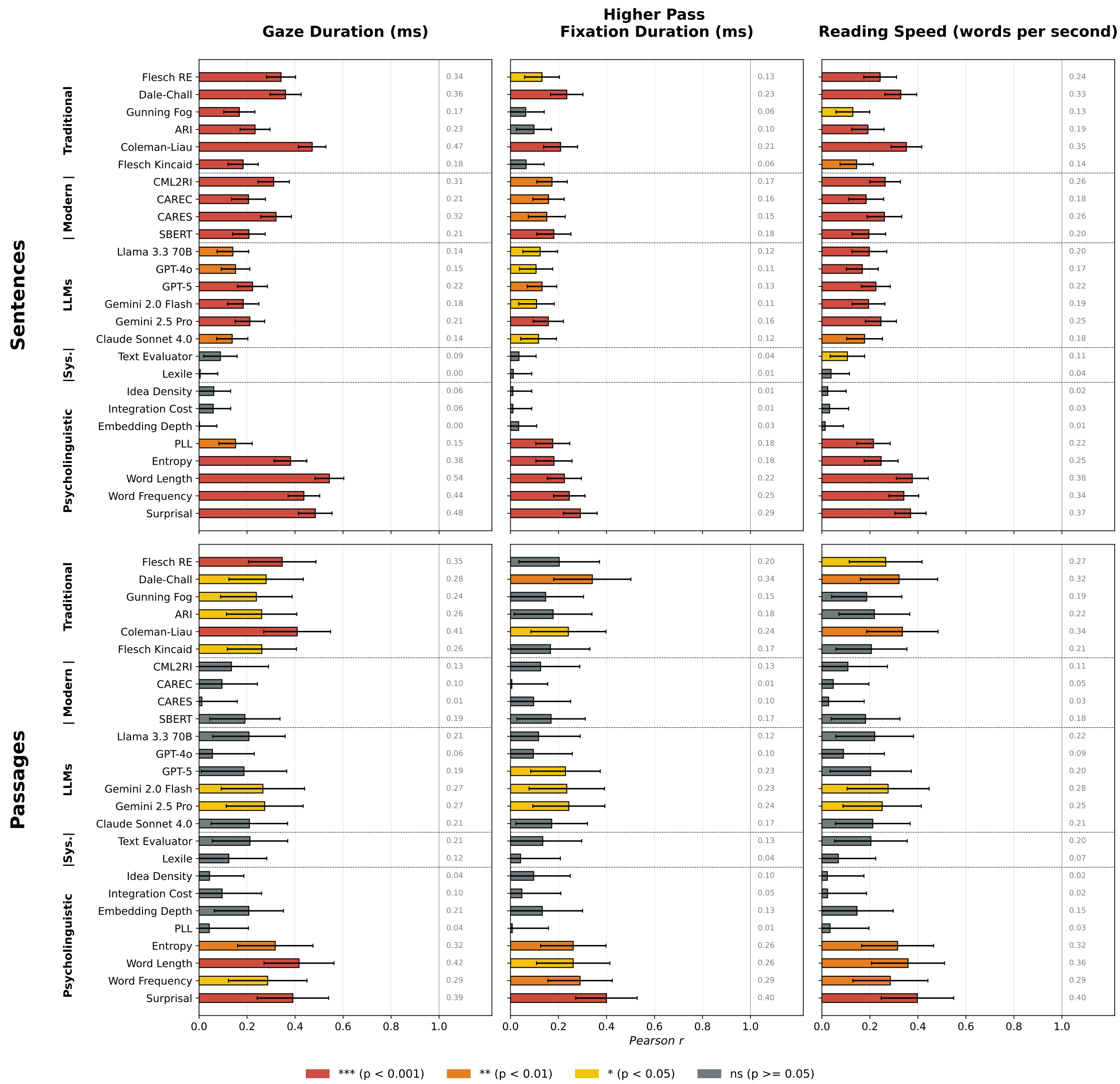

Figure A12: **Evaluation of readability scoring methods and psycholinguistic measures using their predictivity of reading ease.** Presented are Pearson correlation r coefficients for **Gaze Duration (GD), higher pass Fixation Duration (hpFD) and Reading Speed.** Error bars are 95% confidence intervals. Colors represent the statistical significance level of the correlation.

J Results with and without Control for Text Content

In the main analysis, we use an evaluation methodology which controls for text content by regressing the differences of readability scores between original and simplified versions of the same texts against the corresponding differences in reading ease measures. Here, we compare this approach to direct regression of readability scores on reading ease measures for different texts. The results are presented in Figure A13.

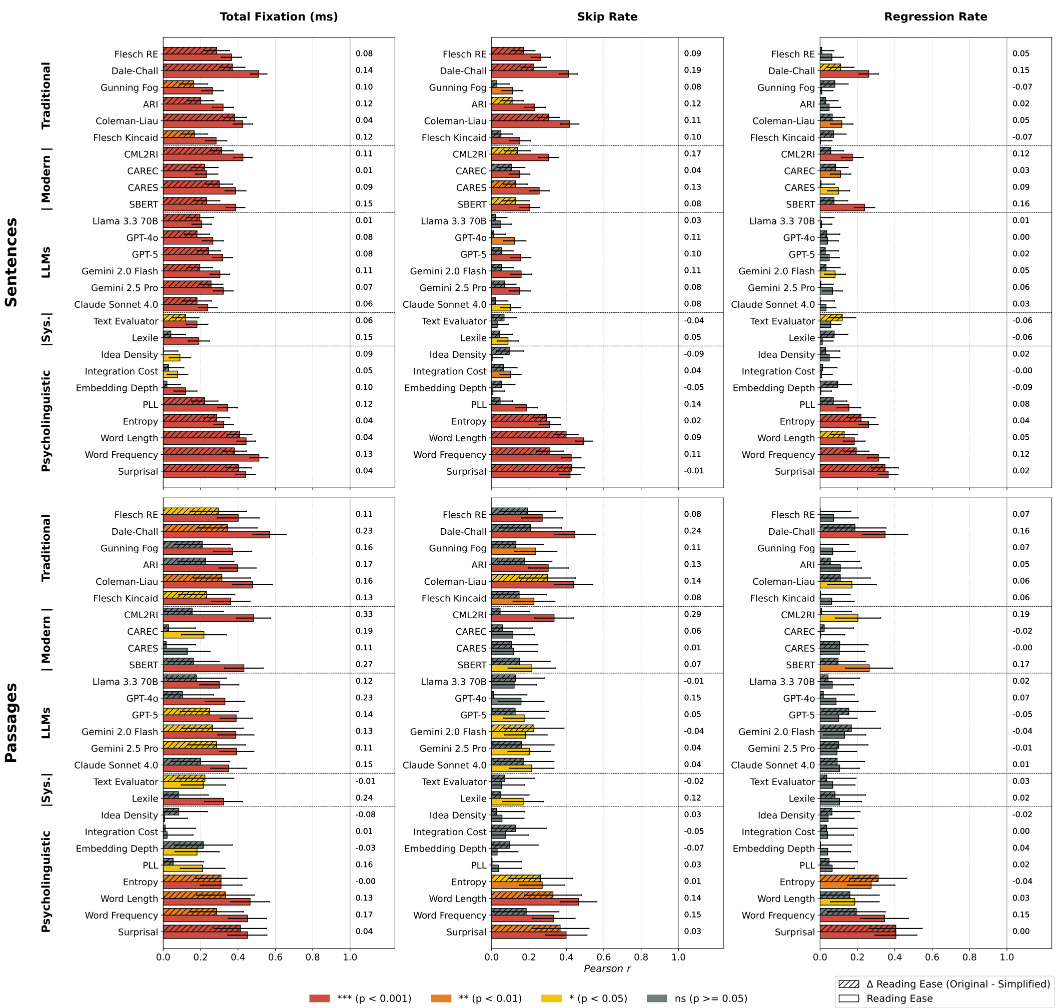

Figure A13: **Reading ease predictivity with and without control for text content.** Each pair of bars consists of (1) the main analysis Pearson correlation r between $\Delta\text{ReadingEase}_{E,T}$ and $\Delta\text{Score}_{M,T}$ (top, striped), and (2) the corresponding evaluation without control for the content of the texts, where we present the Pearson correlation r between $\text{ReadingEase}_{E,T}$ and $\text{Score}_{M,T}$, using all the texts in the OneStopŁ1&L2 corpus (bottom). To the right of each pair, we display the difference in correlations (ΔReading Ease - Reading Ease). Bar colors indicate the statistical significance level of the correlation.

K OneStop: Text Statistics

	Original	**Simplified**	**p value**
Number of passages	162	162	NA
Number of questions	486	486	NA
Words per passage	119.9 ± 4.3	97.1 ± 3.6	***
Sentences per passage	5.78 ± 0.31	5.75 ± 0.27	ns
Sentence length (words)	20.8 ± 0.6	16.9 ± 0.5	***
Mean word length (characters)	4.8 ± 0.04	4.6 ± 0.04	***
Mean word frequency (Wordfreq)	11.28 ± 0.11	10.99 ± 0.11	***
Mean word surprisal (Pythia-70m)	5.01 ± 0.06	4.77 ± 0.06	***

Table A2: Statistics of the original "Advanced" and simplified "Elementary" versions of OneStop texts. For mean values we include a 95% confidence interval, and p-value of a t-test comparing the means of the original and simplified text versions. ns ($p \geq 0.05$), *** ($p < 0.001$).

L Traditional Readability Formulas

Name	Formula	Meaning
Flesch Reading Ease [22]	$206.836 - 84.6 \times \frac{\text{total syllables}}{\text{total words}} - 1.015 \times \frac{\text{total words}}{\text{total sentences}}$	Inversely proportional to the grade level in which 50% of students achieved 75% on material from the McCall-Crabbs Standard Test Lessons in Reading.
Dale Chall Score [16]	$0.1579 \times \frac{\text{difficult words}}{\text{words}} \times 100 + 0.0496 \times \frac{\text{words}}{\text{sentences}} + 3.6365$	Grade level where 50% of the students score at least 50% on the McCall-Crabbs Standard Test Lessons in Reading.
Gunning Fog Index [27]	$0.4 \times \left(\frac{\text{words}}{\text{sentences}} + 100 \times \frac{\text{complex words}}{\text{words}}\right)$	Grade level, representing the estimated years of formal education required to understand the text on first reading.
ARI [52]	$4.71 \times \frac{\text{characters}}{\text{words}} + 0.5 \times \frac{\text{words}}{\text{sentences}} - 21.43$	Grade level where 50% of subjects scored at least 35% on a cloze test.
Coleman Liau Index [11]	$0.0588 \times L - 0.296 \times S - 15.8$ where: L = average letters per 100 words S = average sentences per 100 words	Grade level, scaled according to the expected performance of a college undergraduate in a cloze test.
Flesch Kincaid Grade Score [33]	$0.39 \times \frac{\text{total words}}{\text{total sentences}} + 11.8 \times \frac{\text{total syllables}}{\text{total words}} - 15.59$	Grade level where 50% of the subjects scored at least 35% on a cloze test.

Table A3: Traditional Readability Formulas

M Tools for Computing Readability Scores

The following resources were used to implement the readability measures:

- **textstat library**: version 0.7.4 https://github.com/textstat/textstat, used for extracting the traditional readability measures.
- **Automatic Readability tool for English (ARTE)**: https://nlp.gsu.edu/APIdoc [8], used for extracting the modern readability measures.
- **pycpidr library**: version 0.3.0 https://github.com/jrrobison1/pycpidr, used for calculating idea density measure [6].
- **icy-parses**: https://github.com/dmhowcroft/icy-parses, used for calculating integration cost measure.
- **text-metrics library**: version 1.1.7 https://github.com/lacclab/text-metrics, used for calculating per-word surprisal, length, and frequency.
- **spaCy model (en_core_web_sm-3.8.0)**: https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-3.8.0