Readability Measures as Predictors of Understandability and Engagement in Searching to Learn

Yasin Ghafourian[1,2][0000−0001−9683−9748], Allan Hanbury[2][0000−0002−7149−5843], and Petr Knoth[3][0000−0003−1161−7359]

[1] Research Studios Austria FG, Vienna 1090, Austria
[2] Technische Universität Wien, Vienna 1040, Austria
[3] Knowledge Media Institute, Open University, Milton Keynes, The United Kingdom

Abstract. Search engines have become essential tools for learning, providing access to vast amounts of educational resources. However, selecting the most suitable resources from numerous options can be challenging for learners. While search engines primarily rank resources based on topical relevance, factors like understandability and engagement are crucial for effective learning as well. Understandability, a key aspect of text, is often associated with readability. This study evaluates eight commonly used readability measures to determine their effectiveness in predicting understandability, engagement, topical relevance, and user-assigned ranks. The empirical evaluation employs a survey-based methodology, collecting explicit relevance feedback from participants regarding their preferences for learning from web pages. The relevance data was then analyzed concerning the readability measures. The findings highlight that readability measures are not only reliable predictors of understandability but also of engagement. Specifically, the FKGL and GFI measures demonstrate the highest and most consistent correlation with perceived understandability and engagement. This research provides valuable insights for selecting effective readability measures to tailor search results to the users' learning needs.

Keywords: Empirical Evaluation · Relevance· Understandability · Engagement · Readability Measures · User Study

1 Introduction

Search engines provide access to large quantities of learning resources contributing to the trend of using web search as a means for learning [8, 3]. However, it can be challenging for learners to choose which resources are most suitable from many of the available options [10, 13]. Search engines typically rank resources by their topical relevance, but other characteristics, such as understandability and engagement of the learning resources, are also important to learners.

Several readability measures have been established in the literature. They have been used to assess the complexity of written text and estimate its reading difficulty. In 1969, G. Harry McLaughlin, the creator of one of the widely used

readability measures, defined readability as: "the degree to which a given class of people find certain reading matter compelling and comprehensible [17]." As a result, readability is inherently linked to engagement and understandability.

Although there are several categories of approaches in the literature to measure the readability of text, More research is needed to assess their performance in different use cases. Vajjala's survey [23] summarizes two decades of literature on Automatic Readability Assessment (ARA) and concludes that a clear understanding of effective modeling techniques is still lacking in ARA.

Readability measures consider surface-level language features in web pages, such as sentence structure and word choice. Therefore, lengthy sentences, multisyllabic words, and uncommon vocabulary typically yield readability scores indicating a more complex text. Readability is one of the aspects of the text that contributes to its understandability [26, 5]. Thus, readability has been used in the literature as a proxy for understandability [18, 31, 30]. Table 1 provides a summary of eight of the most frequently used readability measures.

Table 1: Summary of the most common readability measures. S, W, Syl, and Ch show the number of sentences, words, syllables, and characters in the text respectively. W_Polly shows the number of words with 3 or more syllables, and W_long is the number of words with 6 or more characters. DC_DW is the number of difficult words after excluding Dale-Chall's list of 3,000 common words.

Readability Measure	Abbreviation	Formula	Description
Flesch-Kincaid Grade Level Index [12]	FKGL	$11.8 \times (\frac{Syl}{W}) + 0.39 \times (\frac{W}{S}) - 15.59$	Outputs a score as a U.S. grade level needed to understand the text. It can also mean the years of education needed to read the text.
Gunning's Fog Index [9]	GFI (or FOG)	$0.4 \times (\frac{W}{S} + (100 \times \frac{W\_Polly}{W}))$	The Gunning Fog Index formula promotes shorter, plain English sentences for better readability scores, while scores above 12 become difficult for most readers.
Flesh Reading Ease [7]	FRE	$206.835 - 1.015 \times \frac{W}{S} - 84.6 \times \frac{Syl}{W}$	Outputs a number between 0 and 100. The easier the text, the higher the score that it receives.
Coleman-Liau Index [4]	CLI	$0.0588 \times (\frac{Ch}{W} \times 100) + 0.296 \times (\frac{S}{W} \times 100) - 15.8$	Originally developed to help the U.S. Office of Education, CLI approximates a U.S. grade level to understand the text.
Dale-Chall Readability index [6]	DCI	$0.1579 \times (\frac{DC\_DW}{W} \times 100) + 0.0496 \times (\frac{W}{S})$	Outputs a score that corresponds to the U.S. grade system and is based on the use of familiar English words.
Automated Readability Index [22]	ARI	$0.5 \times (\frac{W}{S}) + 4.71 \times (\frac{Ch}{W}) - 21.43$	Outputs a number that approximates the grade needed to understand the text according to U.S. school grade system (from kindergarten to college)
The Lasbarhetsindex [2]	LIX	$\frac{W}{S} + (\frac{W\_long}{W} \times 100)$	Originally developed by a Swedish scholar, LIX is based on a word factor and a sentence factor. It favors the texts with shorter words (less than 6 characters) and sentences.
SMOG Grading [17]	SMOG	$3.1291 + 1.0430 \times \sqrt{\frac{W\_Polly}{S} \times 30}$	This formula estimates the educational years required to comprehend a text with values corresponding from the 4th grade to the college level in the U.S. grading system.

Our Study in this paper is the first that empirically evaluates the predictive capacity of these eight readability measures under the same experimental conditions in a searching to learn context. We aim to assess the degree to which they can be used as predictors of the understandability, engagement, and topical relevance of web pages as perceived by users as well as the rank that the users would assign to these web pages for learning. Our results contribute to a better understanding of the differences in performance and consistency of readability measures, thus helping to select the most effective readability measures in tailoring search results towards the needs of learners.

Our methodology is based on a survey design and proceeds by first collecting explicit relevance feedback focused on participants' preferences for learning about a topic from a set of Web pages. We then analysed how the relevance data provided is associated with the readability measures listed above in Table 1. More specifically, we aim to answer the following two questions:

Research Question 1. To what extent are existing readability measures associated with the perceived understandability, engagement, topical relevance, and user-assigned ranks?
Research Question 2. To what extent are existing readability measures consistent in estimating the perceived understandability, engagement, topical relevance, and user-assigned ranks?

The key contributions of this work are: 1) We show that readability measures are not only good predictors of understandability (as they have been used as a proxy for understandability), but also of engagement of web pages 2) We show that FKGL and GFI are the readability measures with the highest and the most consistent correlation with perceived understandability and engagement.

2 Methodology

Our methodology employs a survey design to gather explicit relevance feedback from online participants' preferences for learning about a specific topic. We selected four topics and created a knowledge test consisting of 10 questions for each topic in survey format, using available online quizzes. This test, administered only once at the start of the survey, aimed to assess participants' existing knowledge of the topics and provide us with insights on the topic knowledge distribution among them. Next, we sampled 10 web pages for each topic using Google as search engine and SerpAPI[4] as a tool to retrieve the results returned by Google. For each topic, we submitted a query that covered the most important concepts in its knowledge test 10 times from different locations and in 10-minute intervals. We then merged the 10 retrieved search engine result pages (SERPs) in a paginated manner and sampled a link from each page of the merged SERP.

For each topic, more than 50 participants were hired from Prolific[5] resulting in a total of 207. Participants were instructed to re-rank the given web pages in descending order based on their opinion of how suitable they found the web pages for learning about the topic. Simultaneously, they were asked to provide three labels for each web page on a 5/7 point Likert scale: 1) the topical relevance, 2) the understandability, and 3) the level of engagement offered by the web page meaning its motivational value for learning about the topic.

Having conducted the survey, we proceeded to calculate the readability value for each web page in our collection using eight different readability measures from Table 1. To extract readability features, we pre-processed the documents using trafilatura[6]. It is worth noting that Palotti et al. [19] have done an investigation on the impact of web page pre-processing on readability measure values.

3 Results

Table 2 provides an overview of the participants who took part in each of the four topics, including their demographic distribution, average declared knowledge of the topic, and average obtained knowledge score after taking the knowledge test.

[4] https://serpapi.com
[5] https://www.prolific.co/
[6] https://trafilatura.readthedocs.io

Table 2: An overview of participants' demographics and characteristics. The Average Declared Knowledge is reported using a 5-point Likert scale and the attained Average Knowledge Test Scores are mapped to the same scale to allow comparison.

Topic Name	Number of Participants	Time Spent on the Survey (Minutes)		Gender Distribution			Age Distribution					Average Declared Knowledge (1-5)	Average Knowledge Test Score	Difference between Declared Knowledge Score and Test Score
		Mean	Standard Deviation	Female	Male	Other	18-24	25-34	35-44	45-54	55+			
World War 2	56	21.25	10.37	44%	56%	0%	5%	51%	22%	15%	7%	3.4	3.48 (62%)	0.08
Financial Literacy	51	22.10	10.12	45%	55%	0%	12%	27%	37%	20%	4%	3.1	3.20 (55%)	0.10
Covid-19	50	18.8	7.41	58%	40%	2%	12%	52%	18%	14%	4%	3.82	2.84 (46%)	0.98
Theory of General Relativity	50	26.62	14.12	44%	54%	2%	4%	38%	28%	24%	6%	2.26	2.32 (33%)	0.06

In order to investigate the association between the readability of web pages and the perceived relevance of those web pages, we utilized the Pearson correlation. Pearson correlation explores the strength and direction of the relationship between user-assigned values and readability measures for research question 1. To study the consistency of the readability measures for research question 2, the standard deviation of correlations across topics is used as a measure of variation.

Table 3: Pearson correlations between user-provided labels and readability measures across topics. FRE • is the measure obtained from negating FRE. The values associated with $FKGL^{\rho}$ and GFI^{σ} are marked in bold as $FKGL^{\rho}$ is the readability measure with the highest average correlation with all labels across topics, and GFI^{σ} is the most consistent readability measure across topics in all dimensions of relevance, as it has the lowest mean standard deviation.

Label Name	Readability Measure	Mean of Correlations	SD of Correlations	Label Name	Readability Measure	Mean of Correlations	SD of Correlations
Understandability	$FKGL^{\rho}$	**0.645**	**0.077**	Engagement	$FKGL^{\rho}$	**0.566**	**0.121**
	GFI^{σ}	**0.642**	**0.074**		GFI^{σ}	**0.566**	**0.094**
	FRE •	0.640	0.070		SMOG	0.560	0.122
	SMOG	0.639	0.042		FRE •	0.559	0.157
	ARI	0.593	0.079		ARI	0.526	0.132
	LIX	0.573	0.074		LIX	0.507	0.151
	DCI	0.434	0.455		DCI	0.424	0.388
	CLI	0.399	0.231		CLI	0.360	0.385
Rank	$FKGL^{\rho}$	**0.526**	**0.217**	Topical Relevance	DCI	0.277	0.452
	GFI^{σ}	**0.511**	**0.184**		FRE •	0.275	0.350
	FRE •	0.504	0.280		$FKGL^{\rho}$	**0.270**	**0.261**
	ARI	0.501	0.246		ARI	0.265	0.325
	SMOG	0.489	0.231		GFI^{σ}	**0.239**	**0.197**
	LIX	0.454	0.261		SMOG	0.223	0.244
	DCI	0.413	0.494		CLI	0.194	0.586
	CLI	0.329	0.503		LIX	0.193	0.286

Most readability measures, with the exception of Flesch Reading Ease (FRE), are inherently designed so that a higher readability score indicates lower text understandability and, more difficult text. In FRE, a higher score signifies higher text understandability and lower difficulty. We have taken this inherent behaviour of readability functions into account during the conversion of the user-assigned labels of relevance from the Likert scale to values to allow for a straightforward comparison. As a result, in the converted Likert values, a lower value indicates a higher preference. For example, "Very Easy" in the understandability label was assigned the value of 1, while "Very Difficult" received the value of 7.

Table 3 shows the results of computing the Pearson correlation between each of the user-assigned values as one variable and each readability measure as the other variable. The correlations were calculated for each pair of label and measure, and then the mean and standard deviation of these correlations were calculated across all topics. The results from Table 3 confirm that the readability

measures, apart from DCI and CLI, show high consistency across topics for understandability and to a large degree also for engagement. For the topical relevance and rank, the standard deviations are substantially higher indicating that these readability measures are not necessarily good predictors for them.

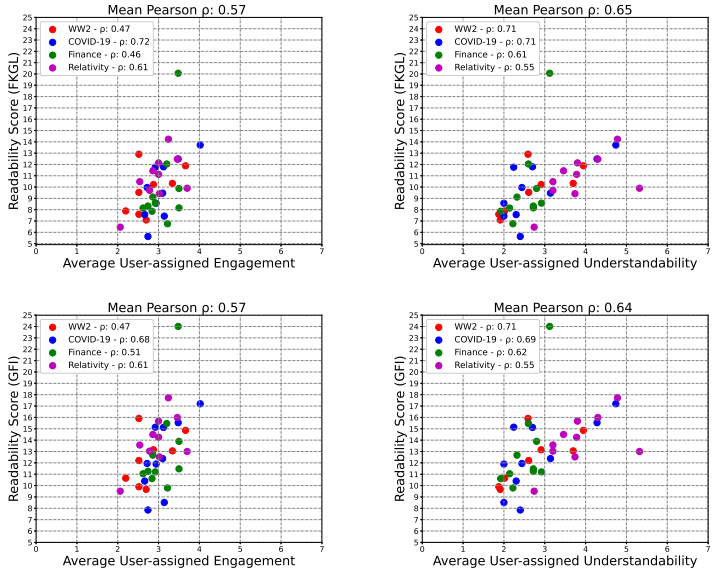

Fig. 1: Alignment of FKGL and GFI with values from user-provided labels, understandability, and engagement. Each dot is a web page.

Among the readability measures, $FKGL$ stands out with the highest mean correlation across all user-assigned labels, while GFI demonstrates the highest consistency across all topics for all labels. The alignment of these two measures with the two labels that have shown the highest correlation with them (understandability and engagement) for all 4 topics is illustrated in Figure 1. A closer look at Figure 1 reveals that both the $FKGL$ and GFI measures exhibit a moderate to strong estimation of user-assigned understandability and engagement across all topics. Moreover, the figure highlights an interesting observation: the correlation values for each topic using both the $FKGL$ and GFI measures are nearly identical. These findings suggest that there is a consistent and robust relationship between these two measures and user-assigned understandability, and user-assigned engagement, regardless of the topic under consideration.

4 Discussion and Related Work

Readability measures have been used in combination with other frameworks to assess the understandability of text across different domains. Some works can be mentioned from the health informatics domain [1, 29, 25, 27]. It has also been previously shown that using readability measures to model text comprehensibility and personalize the search results to the user's understandability level can lead to significant improvements in content ranking [20, 30]. Readability measures have been also investigated in user studies as estimators and predictors of

user-provided data concerning understandability, comprehensibility, etc [11, 28, 24]. For instance, Leroy et al. [14] measured the association between values calculated by readability formulas and values assigned by users to a pair of difficult and easy sentences to measure the effectiveness of a text simplification tool.

Our work in this paper is the first study that looks into evaluating the predictive power of readability measures in a searching to learn context. We have asked our survey participants to provide scores for understandability, engagement, and topical relevance of web pages as they are re-ranking the pages for learning. This work falls in the same category as prior studies which empirically evaluate the performance of readability measures by conducting a user study and measuring the relationship between the user-provided data and the readability measures.

It comes as no surprise for us to see that user-assigned values for understandability exhibit the highest correlation with readability measures. However, it is intriguing to note that these measures serve also as predictors for engagement and, to a considerable extent, for topical relevance although with a weaker correlation with rank, indicating a less pronounced association. A slight surprise is that these measures demonstrate a weaker correlation with topical relevance compared to rank, despite topical relevance being commonly regarded as the primary component influencing the rank. This suggests that understandability and engagement might be equally, if not more, closely linked to rank. This observation is in line with the literature on information retrieval, stating that the overall relevance is not merely a function of topical relevance, but it is a multi-aspect concept including aspects like understandability, novelty, reliability, and other aspects. [16, 21, 5]. Similarly, the results of a user study by Li et. al [15] exploring a multidimensional user relevance model, concluded that "Topicality" does not show a significant contribution to users' relevance judgment.

5 Conclusion

In this paper, we studied the performance of eight of the most frequently used readability measures in predicting understandability, topical relevance, and engagement of online web pages as perceived by users in a learning context. We measured this performance in terms of correlation and consistency. We showed how each of these measures is correlated which each of the user-provided labels and how consistent is each measure across topics. We found out that not only are these measures moderate-strong predictors of understandability, but also they are good predictors of engagement as well. We also found out that in particular, two reading measures of FKGL and GFI have shown the highest correlation and consistency on average with all the user-provided labels. In our future work, we aim to assess the accuracy of readability measures in personalizing search results based on users' understanding of a topic, as estimated by those measures. We will explore this using our assessment of users' knowledge on the topic that are obtained through online quizzes before directing them to relevant web pages.

Acknowledgements This work was supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval – DoSSIER (H2020- EU.1.3.1., ID: 860721).

References

1. Adkins, A.D., Singh, N.N.: Reading level and readability of patient education materials in mental health. Journal of Child and Family Studies **10**, 1–8 (2001)
2. Björnsson, C.H.: Readability of newspapers in 11 languages. Reading Research Quarterly pp. 480–497 (1983)
3. Câmara, A., Roy, N., Maxwell, D., Hauff, C.: Searching to learn with instructional scaffolding. In: Proceedings of the 2021 conference on human information interaction and retrieval. pp. 209–218 (2021)
4. Coleman, M., Liau, T.L.: A computer readability formula designed for machine scoring. Journal of Applied Psychology **60**(2), 283 (1975)
5. Cosijn, E., Ingwersen, P.: Dimensions of relevance. Information Processing & Management **36**(4), 533–550 (2000)
6. Dale, E., Chall, J.S.: A formula for predicting readability: Instructions. Educational research bulletin pp. 37–54 (1948)
7. Flesch, R.F., et al.: Art of readable writing (1949)
8. Gadiraju, U., Yu, R., Dietze, S., Holtz, P.: Analyzing knowledge gain of users in informational search sessions on the web. In: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval. pp. 2–11 (2018)
9. Gunning, R.: The technique of clear writing. mcgraw-hill. New York (1952)
10. Head, A.J., Eisenberg, M.B.: What today's college students say about conducting research in the digital age. Project information literacy progress report **4**(7) (2009)
11. Kauchak, D., Leroy, G., Hogue, A.: Measuring text difficulty using parse-tree frequency. Journal of the Association for Information Science and Technology **68**(9), 2088–2100 (2017)
12. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep., Naval Technical Training Command Millington TN Research Branch (1975)
13. Lee, S.S., Tay, S.M., Balakrishnan, A., Yeo, S.P., Samarasekera, D.D.: Mobile learning in clinical settings: unveiling the paradox. Korean Journal of Medical Education **33**(4), 349 (2021)
14. Leroy, G., Kauchak, D., Mouradi, O.: 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 (2013)
15. Li, J., Zhang, P., Song, D., Wu, Y.: Understanding an enriched multidimensional user relevance model by analyzing query logs. Journal of the Association for Information Science and Technology **68**(12), 2743–2754 (2017)
16. Mao, J., Liu, Y., Zhou, K., Nie, J.Y., Song, J., Zhang, M., Ma, S., Sun, J., Luo, H.: When does relevance mean usefulness and user satisfaction in web search? In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 463–472 (2016)
17. Mc Laughlin, G.H.: Smog grading-a new readability formula. Journal of reading **12**(8), 639–646 (1969)
18. Palotti, J., Goeuriot, L., Zuccon, G., Hanbury, A.: Ranking health web pages with relevance and understandability. In: Proceedings of the 39th international ACM SIGIR conference on Research and development in information retrieval. pp. 965–968 (2016)

19. Palotti, J.R.d.M., Zuccon, G., Hanbury, A.: The influence of pre-processing on the estimation of readability of web documents. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 1763–1766 (2015)
20. Palotti, J.R., Zuccon, G., Goeuriot, L., Kelly, L., Hanbury, A., Jones, G.J., Lupu, M., Pecina, P.: Clef ehealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In: CLEF (Working Notes). pp. 1–22 (2015)
21. Pasi, G.: Contextual search: issues and challenges. In: Symposium of the Austrian HCI and Usability Engineering Group. pp. 23–30. Springer (2011)
22. Smith, E.A., Senter, R.: Automated readability index, vol. 66. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air ... (1967)
23. Vajjala, S.: Trends, limitations and open challenges in automatic readability assessment research. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 5366–5377. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lrec-1.574
24. Verma, M., Yilmaz, E., Craswell, N.: On obtaining effort based judgements for information retrieval. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. pp. 277–286 (2016)
25. Wu, D.T., Hanauer, D.A., Mei, Q., Clark, P.M., An, L.C., Lei, J., Proulx, J., Zeng-Treitler, Q., Zheng, K.: Applying multiple methods to assess the readability of a large corpus of medical documents. Studies in health technology and informatics **192**, 647 (2013)
26. Xu, Y., Chen, Z.: Relevance judgment: What do information users consider beyond topicality? Journal of the American Society for Information Science and Technology **57**(7), 961–973 (2006)
27. Yan, X., Song, D., Li, X.: Concept-based document readability in domain specific information retrieval. In: Proceedings of the 15th ACM international conference on Information and knowledge management. pp. 540–549 (2006)
28. Yaneva, V., Evans, R.: Six good predictors of autistic text comprehension. In: Proceedings of the International Conference Recent Advances in Natural Language Processing. pp. 697–706 (2015)
29. Yılmaz, F.H., Tutar, M.S., Arslan, D., Çeri, A.: Readability, understandability, and quality of retinopathy of prematurity information on the web. Birth Defects Research **113**(12), 901–910 (2021)
30. Zuccon, G.: Understandability biased evaluation for information retrieval. In: Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38. pp. 280–292. Springer (2016)
31. Zuccon, G., Koopman, B.: Integrating understandability in the evaluation of consumer health search engines. In: MedIR@ SIGIR. pp. 32–35 (2014)