Supercalifragilisticexpialidocious: Why Using the "Right" Readability Formula in Children's Web Search Matters

Garrett Allen[1 [0000−0003−4449−1510]], Ashlee Milton[3 [0000−0002−0320−6122]], Katherine Landau Wright[2 [0000−0002−6782−3453]], Jerry Alan Fails[1 [0000−0001−6139−1162]], Casey Kennington[1 [0000−0001−6654−8966]], and Maria Soledad Pera[1 [0000−0002−2008−9204]]

[1] Dept. of Computer Science - Boise State University - Boise, ID
[2] Dept. of Literacy, Language and Culture - Boise State University - Boise, ID
cast-group@boisestate.edu
[3] University of Minnesota, Minneapolis MN 55455, USA
milto064@umn.edu

Abstract. Readability is a core component of information retrieval (IR) tools as the complexity of a resource directly affects its relevance: a resource is only of use if the user can comprehend it. Even so, the link between readability and IR is often overlooked. As a step towards advancing knowledge on the influence of readability on IR, we focus on *Web search* for *children*. We explore how traditional formulas–which are simple, efficient, and portable–fare when applied to estimating the readability of Web resources for children written in English. We then present a formula well-suited for readability estimation of child-friendly Web resources. Lastly, we empirically show that readability can sway children's information access. Outcomes from this work reveal that: (i) for Web resources targeting children, a simple formula suffices as long as it considers contemporary terminology and audience requirements, and (ii) instead of turning to Flesch-Kincaid–a popular formula–the use of the "right" formula can shape Web search tools to best serve children. The work we present herein builds on three pillars: Audience, Application, and Expertise. It serves as a blueprint to place readability estimation methods that best apply to and inform IR applications serving varied audiences.

Keywords: readability, web search, readability, relevance

1 Introduction

Readability, or "the overall effect of language usage and composition on readers' ability to easily and quickly comprehend the document" [58], has a rich history of research surrounding its methods of estimation. These methods range from traditional formulas to advanced lexical and semantic models [16, 31, 58]. Traditional formulas, based on shallow features and developed using highly-curated printed materials like novels and journal articles [31], are routinely applied in

real-world environments [15, 26]. They target varied audience groups [43, 71], languages [39, 35, 72], and content domains [75]. State-of-the-art counterparts leverage complex models [16, 37, 62] based on feature engineering and/or neural-network architectures. They can also adopt a featureless design approach [55, 62]. Yet, how non-traditional models estimate readability is not intuitively understood, nor are these models as easy to deploy as the traditional formulas.

Readability plays a prominent role in *Information Retrieval* (**IR**) for children. In the literature focused on studying and facilitating information access for children, readability is strongly intertwined with the concept of *relevance*. Children must be able to read and understand resource content for it to be deemed relevant, i.e., children must comprehend the text presented to them to extract information that satisfies their needs [10, 56, 68]. The relationship between relevance and readability is discernible in the design of search and recommendation tools that explicitly target children, such as EmSe [32], Read-X [61], and Rabbit [66]. This association is not limited to informing algorithm design but also serves as a perspective for exploratory studies. For instance, a recent study uses readability as a performance measure when inspecting how Web search engines respond to children's queries in the classroom [14]. Bilal et al. [17, 18] rely on readability to examine search result snippets generated by commercial search engines, i.e., Google or Bing, for children's queries. These are meaningful explorations in view of works showing that materials retrieved in response to Web search tasks are inaccessible to many users [28, 83]. In general, top-ranked Web pages retrieved by Google are easier to read than those ranked lower [13]. Still, the average readability of top pages is around the 12^{th} grade [13, 14], which exceeds children's reading skills. This is a concern, as children often browse Search Engine Result Pages (**SERP**) from top to bottom [41]. Despite how interconnected readability and IR for children are, there is no consensus as to what formula to use for readability estimation, nor is there careful consideration about the link between IR applications and the formulas they use.

In this paper, we examine the connection between readability and IR to deepen understanding among researchers and practitioners. We anchor our exploration on three pillars that enable us to study the natural interactions of users with differing skill-sets and the IR applications they use to access information: (i) `Target Audience`, (ii) `Application`, and (iii) `Expertise`. Among other traits, resource relevance depends on the requirements of a user. The diversity in reading ability among *children* in Kindergarten–12^{th} grade allows them to serve as an opportune demographic for our `Target Audience`.[4] Due to the ubiquitous presence of search engines like Google and the fact that children commonly turn to these tools to access online information [14], we designate *Web search tools* as our `Application`. For `Expertise` we use *readability*. We favor traditional formulas for estimation of *English texts*, as opposed to neural methods, due to their simplicity of calculation, portability, prevalence among IR tools [32, 36, 50, 70], and use in real-world general settings [15, 26, 74]. With the analysis presented in this paper, we seek to answer two research questions.

[4] Grade levels according to the United States' educational system.

RQ1: Do traditional formulas effectively estimate the readability of resources targeting children? To answer this question, we undertake an empirical exploration to gauge the applicability of ten traditional formulas on resources written in English targeting children. We first compare and contrast the performance of these formulas across grade levels when applied to books, the medium they were intended to assess. Given our `Application` we further analyze the performance of these formulas when applied to digital resources, not print. We find that the effectiveness of these formulas greatly varies across grades and that lexicon-based formulas fare better than the most popular ones, e.g., Flesch-Kincaid [43], when predicting the readability of Web resources for children. This leads us to another question.

RQ2: Does the choice of readability formula impact the performance of Web search? We investigate if and how readability influences different scenarios related to Web search. We quantify the differences in performance observed by solely exchanging formulas when (i) estimating the readability of children's queries and snippets generated by search engines in response to children's inquiries, (ii) providing query suggestions for children, and (iii) re-ranking resources retrieved in response to children's queries to prioritize those suitable to them. Results from this analysis showcase that the choice of readability formula has the potential to affect children's online information discovery.

The findings emerging from our study highlight the importance of choosing the "right" formula for readability estimation when dealing with children's Web resources, and how that decision exerts influence on Web search for children. The study also results in Spache-Allen, a new formula that extends Spache [71] by explicitly considering terminology familiar to children.[5] With our three pillars, we create a foundation for the investigation of the interaction between readability and IR; particularly the need to appraise the readability formulas used when designing information access tools and how to do so. These tools should be architected to provide user-friendly versions of resources, particularly for domains that use advanced technical jargon. This work has implications for the future development of fair and equitable resource access tools serving all users [38] and reinforces research on IR applications that leverage different readability approaches. Burgeoning research features (multi-modal) conversational applications that interact with users to clarify their information needs [6]. We envision readability playing a role in equipping these applications to formulate response utterances fitting disparate users' skills. In the spirit of accessibility [7, 59], these applications could support users beyond children who may have issues comprehending text, e.g., users with dyslexia or English language learners.

2 Background and Related Work

Readability has been a heavily-investigated area within the last century. Earlier works focused on traditional formulas that take a statistical approach considering

[5] The script used for analysis purposes, along with the Spache-Allen itself can be found at https://github.com/BSU-CAST/ecir22-readability

shallow features like the number of complex words, the number of syllables, or the length of sentences [31]. Among the many formulas in this group, the more well-known include the Flesch-Kincaid Reading Ease [43], the Coleman-Liau Index [24], the Dale-Chall Readability Formula [27, 22], the Gunning Fog Index [40], and the Spache Readability Formula [71]. With the advent of machine learning and neural networks, readability formulas transitioned to readability models, incorporating lexical, semantic, and even multilingual features alongside traditional shallow features to produce estimations [16, 37, 51, 62]. At the same time, we would be remiss not to mention existing commercial efforts, such as Wizenoze Readability Index [80, 81] and Lexile [1]. Unfortunately, there is a lack of standardization of reading levels used for estimations, with differing "scales" in readability prediction. For instance, some use grade levels, others binary labels (simple vs. complex), or varied categorical labels [43, 84, 55]. Consequently, it is increasingly difficult to explore which formula works best and why. Even with recent advancements, traditional formulas tend to be the ones most used in real-world scenarios [15, 26]. Still, traditional formulas are not without flaws. They can produce results that are inaccurate when assessing text that contains many simple, short terms that are highly technical in nature or build a complex, or subtle, story [21, 51, 72, 79]. Further, a critical evaluation of the predicted reading levels of passages used in academic readiness exams revealed that estimations yielded by traditional formulas were 1–3 grades higher than the intended grade levels [73, 74].

Works related to **readability and IR** that also align with our `Target Audience` and `Application` of interest include that of Bilal et al. [18] and Anuyah et al. [14], who study the complexity of resources retrieved by search engines in response to children's queries. Both agree that the reading levels of snippets and resources are too high for children to comprehend. Still, both explorations base their findings on traditional formulas, which can offer misleading estimations and might not be suitable for analyzing Web resources. The impact of readability is not constrained to IR for children. Literature shows that readability is far-reaching within IR. Lately, we see readability support a broad range of IR-related applications, from easing information access [36] and helping teachers locate news articles aligning with the readability levels of their students, to supporting classroom instruction [34] and fake news detection [64]. Through a Firefox plugin, Yu and Miller [85] provide readability support for Asian users who are not fluent in English by enhancing the readability of Web pages. Focusing on recommendation systems, researchers have considered readability as a trait for determining helpful reviews [70] as well as influencing algorithms that recommend books [66, 82, 5] and learning resources [50]. Readability also benefits question answering (QA). For example, researchers have used readability estimated via traditional formulas to identify high-quality developer chats [23] and educational answers in community QA [48], as well as aid detection of the "best" answers to questions in health QA communities [49], and the ranking of answers in community QA sites [29]. Concerning Web search, readability is a trait that has been considered to predict knowledge gain during Web search

[65]. It has also been used as a means to personalize retrieved resources [25, 61] and assess learning as a result of engaging with Web search tasks [69].

This brief overview exhibits the pervasive nature of readability within IR, making the pursuit of understanding its impact a must. With the analysis we discuss in this manuscript, we take initial steps towards that goal.

3 The Fit of Readability Formulas on Web Text

The lack of consensus around which readability formula to use on IR tools makes it uncertain which formula best suits complexity estimation of general Web texts, let alone those intended for young searchers (`Target Audience`). To address this concern, we examine the efficacy of readability formulas for their originally intended purpose: estimating the reading levels of published texts. We then probe their performance when applied to Web resources (`Application`). We study popular traditional formulas (`Expertise`): **(i) DC** - New Dale-Chall [22]; **(ii) SMOG** [57]; **(iii) GF** - Gunning-FOG Index [4]; **(iv) LIX** [20]; **(v) RIX** [12]; **(vi) CL** - Coleman-Liau Index [24], designed for digital texts; **(vii) FK** - Flesch-Kincaid [43], due to its widespread adoption; **(viii) Spache** - Spache Readability Formula [71], meant for texts targeting grades 1^{st}–3^{rd}; and **(ix) SS** - Spache-Sven [52], an enhanced version of Spache that augments its vocabulary with terms that frequently occur on children's websites. For formula details, see [16, 31]. It is apparent in traditional formulas which and how shallow features impact estimation. Instead, neural solutions often lack interpretability on how estimations are produced. Thus, traditional formulas, which are broadly adopted for research and mainstream applications alike, are the focus of this exploration.

For this empirical exploration we use two datasets built using existing corpora. We explicitly examine printed and digital mediums. DSBook is comprised of 235 book excerpts extracted from the appendices of the Common Core State Standards[6] [42], each associated with a range of grade levels. We use the minimum grade level from these ranges as the label, as children reading below their level experience less difficulty with comprehension versus when reading above their level [11]. DSBook also includes 2,084 books from Reading A-Z (RAZ) labeled with their corresponding reading level[7]. DSWeb is made up of 22,689 resources. It includes resources from the WeeBit corpus [77], which consists of samples extracted from WeeklyReader (an educational newspaper), each labeled with their corresponding grade level, and the NewsELA corpus [63], a set of curated news articles with their corresponding grade labels. Given the few resources targeting Kindergarten and 1^{st} graders, DSWeb also incorporates Web resources expertly curated from sites offering content for younger children.

In our experiment, we use Python's Textstat library [2] to estimate the readability of resources in DSBook and DSWeb. We quantify performance via Mean Error Rate (**MER**) and Root Mean Squared Error (**RMSE**). RMSE and MER

[6] A set of learning outcomes to inform curriculum for schools in the United States.

[7] RAZ uses a 26-letter scale assigned by experts for readability [47]. To enable fair comparison, we map letter labels to grade labels, using RAZ's conversion table [46].

exhibited similar trends, thus we omit detailed discussions on the former for brevity. To enable fair assessment for those formulas that provide a score rather than a grade, i.e., LIX, RIX, and DC, we map their outputs to a grade according to conversion tables from their original publications [12, 20, 27]. Through comparison of the results in each medium, we can discern disparities in performance and identify the formulas that better suit estimation of text difficulty of online resources for children. Significance of results are determined using the Kruskal-Wallis H-test [44] with a $p<0.05$. Unless otherwise stated, results reported in this section are significant.

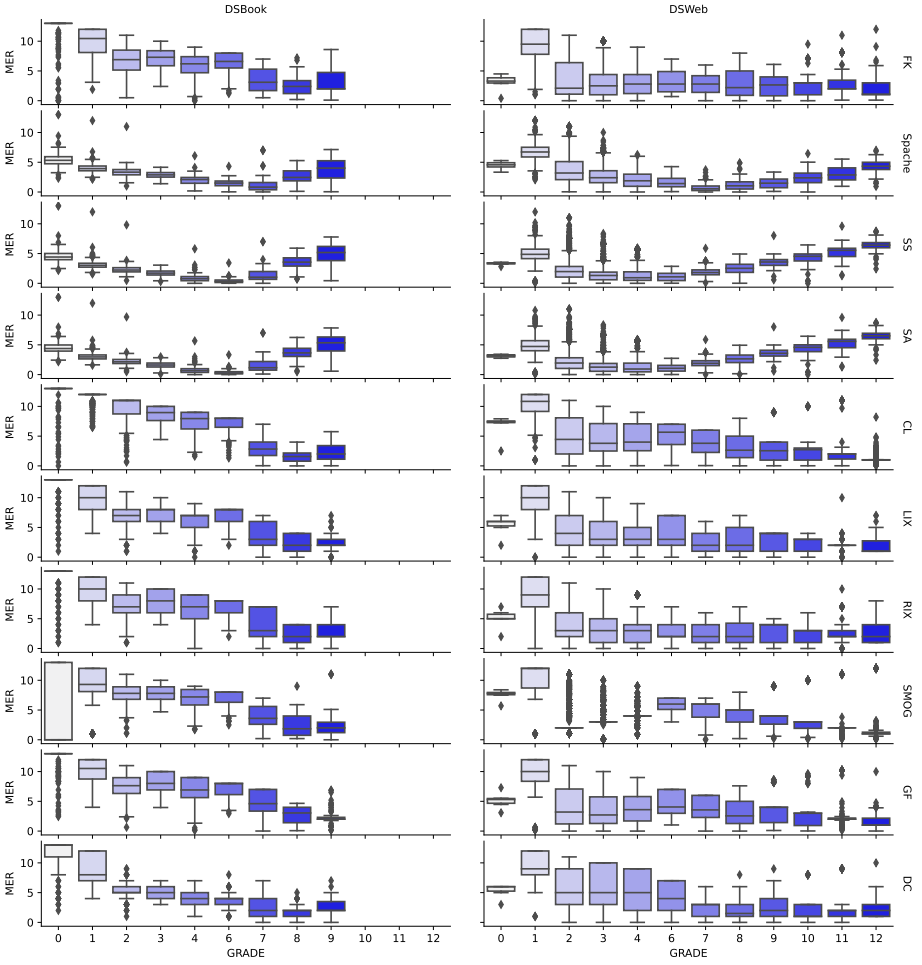


Fig. 1. MER across grades for readability formulas applied to DSBook and DSWeb. Resources in DSBook are labeled with a grade range indicating the corresponding target audience, so we take the lowest grade as ground truth.

We first investigate the capabilities of readability formulas using DSBook. As shown in Figure 1, Spache and SS tend to produce lower MER towards the middle grades, whereas the other formulas yield lower MER for the later grades. CL exhibits the lowest MER at the 9^{th} grade and above. Interestingly, even though it is commonly used [17], FK is one of the formulas that produces the highest MER, as compared to Spache and DC. Overall, Spache and SS are the least error-prone for resources intended for grades K–6^{th}.

To see if this performance translates to Web resources, we repeat the experiment using DSWeb. As illustrated in Figure 1, the results for Web resources are similar to those obtained for books in that Spache and SS are the least error-prone formula for resources targeting younger readers. With the exception of Spache and SS, traditional formulas are inconsistent when estimating the complexity of texts for earlier grades (K–6^{th}). Outcomes from the presented analysis serve as an indication of Spache and SS being formulas particularly well-suited for estimating the readability of Web resources for young readers.

Regardless of its effectiveness for our audience and resource type, Spache's static vocabulary–consisting of 1,064 words that are considered "easy" for children to comprehend [3]–is limited and includes terminology from the 1970s. As language changes over time [67], an outdated vocabulary may not capture easy terms for children in today's world, potentially pushing the formula to misleading text complexity estimations. The benefit of changing the 1940s vocabulary used in the original Dale-Chall formula [27] to the one used by the DC formula, more aligned to the 1990s, is apparent [22]. Similar boosts are seen with SS [52], which augments Spache's original vocabulary list through the inclusion of a dictionary of 48,000 non-stop lemmatized terms the authors extracted from children-related websites. Nevertheless, this enhancement relies on word frequency analysis and assumes that terms added to the vocabulary are understood by children, which may not always be the case.

To include vocabulary that children learn through instruction, we take advantage of the Age of Acquisition (**AoA**) dataset. This dataset contains acquisition ratings in the form of ages, ranging from 1–17 years, for ∼30,000 English words [45]. We posit that there is a benefit to simultaneously accounting for terminology that children have been exposed to through websites as well as terminology that has been taught. Thus, we merge the original Spache vocabulary with the terms from AoA and the dictionary from [52]; we call this updated formula Spache-Allen (**SA**), which is computed as in Equation 1.

$$\textit{Spache-Allen}(R) = (0.141 \times w_R/s_R) + (0.086 * dif(R)) + 0.839 \qquad (1)$$

where R is a resource, w_R and s_R are the number of words and sentences in R, respectively. The function $dif(R)$ determines the percentage of difficult words in R, where a word is deemed difficult if it does not appear in the "easy" vocabulary–in this case it includes 65,669 unique terms that children learn through instruction and/or are exposed to online, in addition to the original Spache's term list.

Regardless of the dataset considered, augmenting Spache's original vocabulary has a positive effect on readability estimation as it leads to decreases in

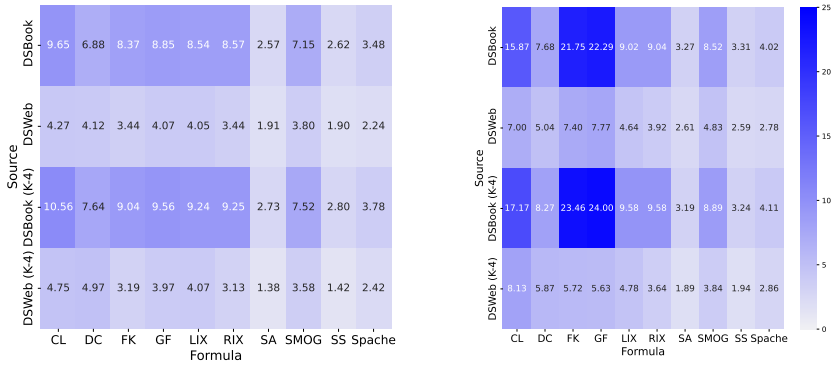
MER (Figure 1). SA consistently outperforms all other investigated traditional formulas through grade 5; it's performance is comparable to that of Spache and SS on higher grades. For grades 9 and above, formulas like CL yield the lowest MER, which is anticipated, given that Spache, SS, and SA have the express purpose of determining the difficulty of texts targeting younger readers.

With RQ1, we aimed to answer: *Do traditional formulas effectively estimate the readability of resources targeting children?* From trends in MER and RMSE, it is evident that the reliability of some formulas differs upon the source material they are applied to (e.g., DC averages a MER of 6.88 for books vs. 4.12 for Web resources). We see that, on average, book resources result in larger errors than Web resources; this is also prevalent among material targeting early readers, i.e., grades K–4^{th}. Interestingly, the MER and RMSE per formula varies depending on the grade of the text being assessed. This is particularly salient among early readers, both numerically and visibly in Figure 1. Even more so in Figure 2, when contrasting performance on the DSBook and DSWeb versus respective subsets of the datasets consisting of materials till the 4^{th} grade. The RMSE reported in Figure 2(b) is particularly telling as it doubles for CL and more than triples for GF, and FK, when contrasting overall performance for K–4^{th} grade resources. In the end, the Spache, SS, and SA formulas are the least error-prone when applied to Web resources targeting younger audiences. Though these three formulas perform similarly, the differences across them are significant (Kruskal-Wallis H-test, $p<0.05$). Therefore, the formula we see as most suitable to support tasks related to Web search for children is SA.

4 The Effect of Readability on Web Search for Children

It emerges from Section 3 that readability formulas do falter. With readability playing a prominent role in Web search for children, we investigate the cascading effect that the choice of readability formula can have on Web search. To do so, we consider four scenarios that spotlight different stages of the search process. In each scenario, we quantify the fluctuations in performance that result from using traditional readability formulas. As in Section 3, we use Python's Textstat library for readability estimation. A cursory search (on ACM Digital Library and Google Scholar) for recent literature focused on readability and IR applications reveal a plethora of recommender systems, QA, search, and text simplification strategies, to name a few, that depend upon readability as one of their components. Many of these applications default to FK as the readability formula of choice. For this reason, in each scenario, we treat performance based on FK as a baseline. For significance, we use a two-tailed student t-test with a Bonferroni correction (with α=0.05 and the number of tests N=10, which is the number of formulas) with $p<0.05$; all results are significant unless reported otherwise.

Scenario 1. In this scenario, we consider readability as a means to facilitate personalization, e.g., filtering and/or prioritizing retrieved resources. We posit that the readability of a query could serve as a proxy for the reading skills of the user initiating the search. In turn, this information can be used as a signal to



(a) MER. (b) RMSE.

Fig. 2. Error rate analysis by source and formula. We pay special attention to errors yielded by traditional formulas when applied to material targeting early readers.

Table 1. Fluctuations in performance observed in scenarios related to Web search for children when applying traditional readability formulas. Bold denotes best performing formula for the corresponding scenario, and '*' indicates significant w.r.t. Flesch-Kincaid (FK), a formula that is often used as a component of IR applications.

Scenario	**Metric**	**Formulas**									
		DC	*SMOG*	*GF*	*LIX*	*RIX*	*CL*	*FK*	*Spache*	*SS*	*SA*
1	MER	5.94*	4.68*	4.37*	4.49*	**2.1***	3.84	3.72	2.74*	2.3*	2.34*
2	MRR	**0.39**	0.38	**0.39**	0.36	**0.39**	0.38	0.36	0.4	n/a	0.37
3	MER	4.08*	5.23*	3.49	3.09	3.02*	3.61	3.31	**2.85***	3.0*	3.02*
4	MRR	0.36*	0.2*	0.27*	0.3*	0.27*	0.27*	0.42	**0.43**	0.3*	0.29*
4 (K-4)	MRR	0.26*	0.14*	0.26*	0.22*	0.22*	0.14*	0.49	0.48	**0.50**	0.48

filter and/or re-rank retrieved resources to match the users' inferred skills [25, 53, 76]. We use the 168 queries made available by the authors in [52, 8], each labeled with the grade of the child formulating the corresponding query. We estimate the readability of each query using the formulas in Section 3 and then compare their estimations with respect to the ground truth. For evaluation, we use MER (excluding RMSE for brevity, given similar trends). We are aware that the nature of traditional formulas makes them less suitable for short texts such as queries. Nevertheless, this is a limitation that affects all formulas (possibly with the exception of SS, which has been proven successful in identifying if a query was child-like [52]), therefore reported observations are not affected.

As reported in Table 1, it is clear that the choice of the formula used to estimate the readability of queries has the potential to skew the inference of the users' reading skills [76]. For example, formulas like RIX, Spache, SS, and SA lead to a MER of approximately ±2 grades, whereas DC or SMOG can predict

up to 4 grades above or below the grade of the child who initiated the query. While we expected discrepancy w.r.t. ground truth given that queries are short, if readability is used to enable personalization, the latter would be a concern as the tool would not be adequately supporting the target user.

Scenario 2. Children struggle with formulating queries when searching online [54]. Therefore, in this scenario, we examine the impact that readability has on query suggestion as a means to alleviate children's query formulation issues. We study the performance of ReQuIK [52], a state-of-the-art strategy that offers suggestions targeted at children. As ReQuIK utilizes both FK (popularity) and SS (Web applicability), we first observe fluctuations in its performance when exchanging FK for each of the remaining formulas. Motivated by the outcomes reported in Section 3, we also retain FK but instead replace SS with SA. In this experiment, we rely on ReQuik's implementation provided by the authors, and also use the 95 queries used in the original experiments [52]. To generate candidate query suggestions for each of the aforementioned queries, we use an $N-1$ approach: in each case, we use the prefix of each query (consisting of $N-1$ terms) to trigger Google's query suggestions via its API. Treating the original query as the ground truth, i.e., what should be ranked first, we calculate the Mean Reciprocal Rank (**MRR**) of the top-10 query suggestions ranked by ReQuIK.

Based on the results reported for Scenario 1, we expected the wide range of estimation errors to impact query suggestion generation. However, from the analysis of results reported in row 2 of Table 1, as well as the experiment using SA and FK to power ReQuIK (MRR of 0.37), it emerges that variations on ReQuIK's performance caused by swapping traditional formulas are not significant. Upon in-depth inspection, we attribute this to ReQuIK's design that incorporates neural architectures. Even though readability is an important trait considered in the wide model component of ReQuIK, it is the deep neural model component that most contributes to ReQuIK's overall success (cf. [52]).

Scenario 3. Snippets are meant to offer children a glimpse into the resources retrieved as they navigate a SERP. For the snippets to facilitate relevant resource selection, they must offer content that children can comprehend. We conduct a new experiment following the procedure outlined in Scenario 1, but on snippets instead of queries. We consider the snippets generated using Google's Custom Search API for a sample of 395 NewsELA resource titles acting as queries. We estimate snippet readability using the formulas in Section 3. Treating the original grade label for the corresponding NewsELA resource as ground truth, we compute the respective MER for assessment purposes.

As reported in row 3 of Table 1, there are significant performance fluctuations. As anticipated, Spache, SS, and SA lead to the lowest errors in estimation. On the other hand, SMOG and DC lead to more erroneous estimations. If SERP were to be personalized to ensure children could comprehend presented snippets, then the misleading readability estimations caused by some formulas could result in a SERP that excludes relevant resources. Additionally, as snippets act as proxies for resource content, they could be used in lieu of Web page content

for the purpose of re-ranking SERP for children and thus the use of misleading formulas could cause unhelpful changes to the SERP.

Scenario 4. Readability is a key relevance trait informing ranking, particularly given that children have different expectations and needs when it comes to retrieved resources [19]. In this scenario, we examine the effect that readability has on the performance of KORSCE [60], a re-ranking strategy that prioritizes resources for children in the classroom setting. Following the experimental protocol of Scenario 2, we exchange FK, the formula originally used by KORSCE, with each of the formulas under study enabling us to gauge potential performance implications. In this experiment, we sample 193 NewsELA resources and use their titles as queries. Using Google's Custom Search API, we collect the top-10 corresponding resources per query. We re-rank the resources associated with each query using KORSCE, treating the original resource as ground truth. To quantify performance, we use MRR.[8]

The results reported in row 4 of Table 1 show that just by interchanging the readability formula embedded in KORSCE's architecture, relevant resources move from position 5 in the ranking (i.e., SMOG's MRR is 0.2) to position ∼2 (based on MRR for Spache and FK, which are 0.43 and 0.42, respectively).[9] This is even more evident among rankings of resources for early readers, who would need the most help from tools when pursuing online information discovery tasks (row K–4 in Table 1). In their case, the relevant resources could move from position 7 in the ranking to 2, simply by exchanging SMOG or CL with Spache, SS, SA, or FK. As children tend to linearly examine SERP [41], the choice of readability formula could prompt the ranking algorithm to inadvertently position higher on the SERP resources children are unable to read or understand, thus negatively affecting their search experience.

With RQ2, we sought to answer: *Does the choice of readability formula impact the performance of Web search?* From the findings discussed in this section, we can surmise that yes, the choice of readability formula affects in a meaningful manner Web search for children. Altering the formula used leads to variations in performance across most of the scenarios examined for Web search for children. Variations were not significant for Scenario 2. We attribute this to ReQuIK's deep model dominating its wide counterpart. Overall, the shift in performance caused by the choice of formula matters, as retrieving resources at appropriate reading levels positively impacts user satisfaction [25, 33]. A further concern related to this shift is that searchers could be deterred from engaging with query suggestions or resources that are assumed to be above searchers' skills when the queries and resources could very well be comprehensible and hence relevant. Formulas underestimating difficulty could mistakenly direct searchers to query suggestions or prioritize resources that are far beyond what searchers can comprehend, thus unintentionally setting them up for a failed search.

[8] We use KORSCE's implementation made available by the authors.
[9] In Scenario 4, FK's performance is not unexpected as KORSCE is optimized for FK.

5 Conclusion and Future Work

In this paper, we aimed to highlight the natural connection between readability and IR tools. In particular, we focused our analysis on the impact readability has on Web search for children. We gauged the performance of traditional formulas when applied to estimate the readability of printed and digital material targeting children. Moreover, through different scenarios intended to draw attention to different stages of the search process, we studied performance fluctuations that are a direct consequence of exchanging readability formulas.

Analysis of the experimental results suggests that even though Flesch-Kincaid is commonly used to determine the readability of Web resources, it is not the one that best captures their level of difficulty, especially when these Web resources target younger audiences. We have shown that variations of the well-known Spache formula, which explicitly considers terminology children are exposed to online and/or learn as they grow, are better suited to estimate the readability of Web resources for young searchers (RQ1). Of note, we introduced Spache-Allen, which emerged as a result of the explorations conducted in pursuit of RQ1. The effect of readability on algorithms empowering information discovery for young searchers also became apparent during our explorations; making it imperative for developers and researchers to consider using the "right" formula, one best serving the target audience and application, as it directly translates to performance improvements (RQ2). From reported findings we surmise that (i) the performance of IR applications can indeed change based on the readability formula used and (ii) by carefully considering which readability formula supports the target audience of interest, IR applications can be optimized for performance or personalization with respect to an audience (echoing the reports in [78] on general Web resources, not just those targeting children).

Lessons learned from this work inform ongoing efforts related to better enabling children's information discovery through Web search. These include algorithmic solutions that rely on readability as one of their components to suggest queries [52], determine search intent [30], identify resources that are relevant to children [60, 25], aid teachers seeking texts for their classrooms [34], or offer teachers insights on students' abilities via search [9]. As decisions related to readability impact all areas of IR, the applicability of this work is far reaching. Further, the pillars introduced can serve as a blueprint that researchers can turn to as a guide for their own explorations towards finding a well-suited readability estimation solution for their intended tasks and audiences.

We limited our examination to traditional formulas applied to Web resources written in English. In the future, we plan to extend our analysis to state-of-the-art counterparts to identify the benefits and constraints inherent to dealing with these more complex models. As a step towards making information accessible worldwide, and given the rise of multilingual strategies for readability estimation, we will extend our exploration to written languages beyond English [56].

Acknowledgments. Work partially funded by NSF Award #1763649. The authors would like to thank Dr. Ion Madrazo Azpiazu and Dr. Michael D. Ekstrand for their valuable feedback.

References

1. https://www.lexile.com/
2. https://github.com/shivam5992/textstat
3. https://github.com/cdimascio/py-readability-metrics/blob/master/readability/data/spache_easy.txt
4. Albright, J., de Guzman, C., Acebo, P., Paiva, D., Faulkner, M., Swanson, J.: Readability of patient education materials: implications for clinical practice. Applied Nursing Research **9**(3), 139–143 (1996)
5. Alharthi, H., Inkpen, D.: Study of linguistic features incorporated in a literary book recommender system. In: ACM/SIGAPP SAC. pp. 1027–1034 (2019)
6. Aliannejadi, M., Zamani, H., Crestani, F., Croft, W.B.: Asking clarifying questions in open-domain information-seeking conversations. In: ACM SIGIR. pp. 475–484 (2019)
7. Allan, J., Croft, B., Moffat, A., Sanderson, M.: Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012. In: ACM SIGIR Forum. vol. 46, pp. 2–32 (2012)
8. Allen, G., Peterson, B.L., Ratakonda, D.k., Sakib, M.N., Fails, J.A., Kennington, C., Wright, K.L., Pera, M.S.: Engage!: Co-designing search engine result pages to foster interactions. In: ACM IDC. pp. 583–587 (2021)
9. Allen, G., Wright, K.L., Fails, J.A., Kennington, C., Pera, M.S.: Casting a net: Supporting teachers with search technology. arXiv preprint arXiv:2105.03456 (2021)
10. Amendum, S.J., Conradi, K., Hiebert, E.: Does text complexity matter in the elementary grades? a research synthesis of text difficulty and elementary students' reading fluency and comprehension. Ed. Psy. Review **30**(1), 121–151 (2018)
11. Amendum, S.J., Conradi, K., Liebfreund, M.D.: The push for more challenging texts: An analysis of early readers' rate, accuracy, and comprehension. Reading Psychology **37**(4), 570–600 (2016)
12. Anderson, J.: Lix and rix: Variations on a little-known readability index. Journal of Reading **26**(6), 490–496 (1983)
13. Antunes, H., Lopes, C.T.: Readability of web content. In: CISTI. pp. 1–4 (2019)
14. Anuyah, O., Milton, A., Green, M., Pera, M.S.: An empirical analysis of search engines' response to web search queries associated with the classroom setting. Aslib (2019)
15. Begeny, J.C., Greene, D.J.: Can readability formulas be used to successfully gauge difficulty of reading materials? Psychology in the Schools **51**(2), 198–215 (2014)
16. Benjamin, R.G.: Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. Ed. Psy. Review **24**(1), 63–88 (2012)
17. Bilal, D.: Comparing google's readability of search results to the flesch readability formulae: a preliminary analysis on children's search queries. American Society for Information Science and Technology **50**(1), 1–9 (2013)
18. Bilal, D., Huang, L.M.: Readability and word complexity of serps snippets and web pages on children's search queries: Google vs bing. Aslib (2019)
19. Bilal, D., Kirby, J.: Differences and similarities in information seeking: children and adults as web users. IPM **38**(5), 649–670 (2002)
20. Björnsson, C.H.: Läsbarhet: hur skall man som författare nå fram till läsarna? Bokförlaget Liber (1968)
21. Bruce, B., Rubin, A., Starr, K.: Why readability formulas fail. IEEE Transactions on Professional Communication (1), 50–52 (1981)

22. Chall, J.S., Dale, E.: Readability revisited: The new Dale-Chall readability formula. Brookline Books (1995)
23. Chatterjee, P., Damevski, K., Kraft, N.A., Pollock, L.: Automatically identifying the quality of developer chats for post hoc use. ACM TOSEM **30**(4), 1–28 (2021)
24. Coleman, M., Liau, T.L.: A computer readability formula designed for machine scoring. Journal of Applied Psychology **60**(2), 283 (1975)
25. Collins-Thompson, K., Bennett, P.N., White, R.W., De La Chica, S., Sontag, D.: Personalizing web search results by reading level. In: ACM CIKM. pp. 403–412 (2011)
26. Crossley, S.A., Skalicky, S., Dascalu, M.: Moving beyond classic readability formulas: New methods and new models. Journal of Research in Reading **42**(3-4), 541–561 (2019)
27. Dale, E., Chall, J.S.: A formula for predicting readability: Instructions. Educational research bulletin pp. 37–54 (1948)
28. D'Alessandro, D.M., Kingsley, P., Johnson-West, J.: The readability of pediatric patient education materials on the world wide web. Archives of pediatrics & adolescent medicine **155**(7), 807–812 (2001)
29. Dalip, D.H., Gonçalves, M.A., Cristo, M., Calado, P.: Exploiting user feedback to learn to rank answers in q&a forums: a case study with stack overflow. In: ACM SIGIR. pp. 543–552 (2013)
30. Dragovic, N., Madrazo Azpiazu, I., Pera, M.S.: " is sven seven?" a search intent module for children. In: ACM SIGIR. pp. 885–888 (2016)
31. DuBay, W.H.: Smart Language: Readers, Readability, and the Grading of Text. (2007)
32. Eickhoff, C., Azzopardi, L., Hiemstra, D., De Jong, F., de Vries, A.P., Dowie, D., Torres, S.D., Glassey, R., Gyllstrom, K., Kruisinga, F., et al.: Emse: initial evaluation of a child-friendly medical search system. In: IIiX. pp. 282–285 (2012)
33. Eickhoff, C., de Vries, A.P., Collins-Thompson, K.: Copulas for information retrieval. In: ACM SIGIR. pp. 663–672 (2013)
34. Ekstrand, M.D., Wright, K.L., Pera, M.S.: Enhancing classroom instruction with online news. Aslib (2020)
35. El-Haj, M., Rayson, P.: Osman—a novel arabic readability metric. In: LREC. pp. 250–255 (2016)
36. Ermakova, L., Bellot, P., Braslavski, P., Kamps, J., Mothe, J., Nurbakova, D., Ovchinnikova, I., Sanjuan, E.: Text simplification for scientific information access. In: ECIR (2021)
37. François, T., Miltsakaki, E.: Do nlp and machine learning improve traditional readability formulas? In: First Workshop on Predicting and Improving Text Readability for target reader populations. pp. 49–57 (2012)
38. Garcia-Febo, L., Hustad, A., Rösch, H., Sturges, P., Vallotton, A.: Ifla code of ethics for librarians and other information workers. https://www.ifla.org/publications/ifla-code-of-ethics-for-librarians-and-other-information-workers--short-version-/
39. Gonzalez-Dios, I., Aranzabe, M.J., de Ilarraza, A.D., Salaberri, H.: Simple or complex? assessing the readability of basque texts. In: COLING. pp. 334–344 (2014)
40. Gunning, R.: The fog index after twenty years. Journal of Business Communication **6**(2), 3–13 (1969)
41. Gwizdka, J., Bilal, D.: Analysis of children's queries and click behavior on ranked results and their thought processes in google search. In: CHIIR. pp. 377–380 (2017)
42. Initiative, C.C.S.S.: Appendix b: Text exemplars and sample performance tasks (2020), http://www.corestandards.org/assets/Appendix_B.pdf

43. 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)
44. Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. Journal of the American statistical Association **47**(260), 583–621 (1952)
45. Kuperman, V., Stadthagen-Gonzalez, H., Brysbaert, M.: Age-of-acquisition ratings for 30,000 english words. Behavior research methods **44**(4), 978–990 (2012)
46. LAZEL, I.: Level correlation chart. https://www.readinga-z.com/learninga-z-levels/level-correlation-chart/ (2021), (accessed Jan 18, 2021)
47. LAZEL, I.: Reading a-z: The online reading program with downloadable books to print and assemble. https://www.readinga-z.com/ (2021), (accessed Jan 18, 2021)
48. Le, L.T., Shah, C., Choi, E.: Evaluating the quality of educational answers in community question-answering. In: IEEE/ACM JCDL. pp. 129–138 (2016)
49. Lin, C.Y., Wu, Y.H., Chen, A.L.: Selecting the most helpful answers in online health question answering communities. Journal of Intelligent Information Systems pp. 1–23 (2021)
50. Liu, L., Koutrika, G., Wu, S.: Learningassistant: A novel learning resource recommendation system. In: IEEE ICDE. pp. 1424–1427 (2015)
51. Madrazo Azpiazu, I.: Towards Multipurpose Readability Assessment. Master's thesis, Boise State University (2016), https://scholarworks.boisestate.edu/td/1210/
52. Madrazo Azpiazu, I., Dragovic, N., Anuyah, O., Pera, M.S.: Looking for the movie seven or sven from the movie frozen? a multi-perspective strategy for recommending queries for children. In: ACM CHIIR. pp. 92–101 (2018)
53. Madrazo Azpiazu, I., Dragovic, N., Pera, M.S.: Finding, understanding and learning: Making information discovery tasks useful for children and teachers. SAL Workshop co-Located with ACM SIGIR (2016)
54. Madrazo Azpiazu, I., Dragovic, N., Pera, M.S., Fails, J.A.: Online searching and learning: Yum and other search tools for children and teachers. Information Retrieval Journal **20**(5), 524–545 (2017)
55. Madrazo Azpiazu, I., Pera, M.S.: Multiattentive recurrent neural network architecture for multilingual readability assessment. TACL **7**, 421–436 (2019)
56. Madrazo Azpiazu, I., Pera, M.S.: An analysis of transfer learning methods for multilingual readability assessment. In: Adjunct Publication of the 28th ACM UMAP. pp. 95–100 (2020)
57. Mc Laughlin, G.H.: Smog grading-a new readability formula. Journal of Reading **12**(8), 639–646 (1969)
58. Meng, C., Chen, M., Mao, J., Neville, J.: Readnet: A hierarchical transformer framework for web article readability analysis. Advances in Information Retrieval **12035**, 33 (2020)
59. Milton, A., Allen, G., Pera, M.S.: To infinity and beyond! accessibility is the future for kids' search engines. arXiv preprint arXiv:2106.07813 (2021)
60. Milton, A., Anuya, O., Spear, L., Wright, K.L., Pera, M.S.: A ranking strategy to promote resources supporting the classroom environment. In: IEEE/WIC/ACM WI-IAT. pp. 121–128 (2020)
61. Miltsakaki, E., Troutt, A.: Read-x: Automatic evaluation of reading difficulty of web text. In: E-Learn. pp. 7280–7286. AACE (2007)
62. Mohammadi, H., Khasteh, S.H.: Text as environment: A deep reinforcement learning text readability assessment model. arXiv preprint arXiv:1912.05957 (2019)
63. Newsela: Newsela article corpos (2016), https://newsela.com/data

64. Ngada, O., Haskins, B.: Fake news detection using content-based features and machine learning. In: IEEE CSDE. pp. 1–6 (2020)
65. Otto, C., Yu, R., Pardi, G., von Hoyer, J., Rokicki, M., Hoppe, A., Holtz, P., Kammerer, Y., Dietze, S., Ewerth, R.: Predicting knowledge gain during web search based on multimedia resource consumption. In: AIED. pp. 318–330 (2021)
66. Pera, M.S., Ng, Y.K.: Automating readers' advisory to make book recommendations for k-12 readers. In: ACM RecSys. pp. 9–16 (2014)
67. Ramiro, C., Srinivasan, M., Malt, B.C., Xu, Y.: Algorithms in the historical emergence of word senses. National Academy of Sciences **115**(10), 2323–2328 (2018)
68. Reed, D.K., Kershaw-Herrera, S.: An examination of text complexity as characterized by readability and cohesion. Journal of Experimental Ed. **84**(1), 75–97 (2016)
69. Roy, N., Torre, M.V., Gadiraju, U., Maxwell, D., Hauff, C.: Note the highlight: Incorporating active reading tools in a search as learning environment. In: ACM CHIIR. pp. 229–238 (2021)
70. Saptono, R., Mine, T.: Time-based sampling methods for detecting helpful reviews. In: IEEE/WIC/ACM WI-IAT. pp. 508–513 (2020)
71. Spache, G.D.: The spache readability formula. Good reading for poor readers pp. 195–207 (1974)
72. Spaulding, S.: A spanish readability formula. The Modern Language Journal **40**(8), 433–441 (1956)
73. Szabo, S., Sinclair, B.: Staar reading passages: The readability is too high. Schooling **3**(1), 1–14 (2012)
74. Szabo, S., Sinclair, B.B.: Readability of the staar test is still misaligned. Schooling **10**(1), 1–12 (2019)
75. Tahir, M., Usman, M., Muhammad, F., Khan, I., Idrees, M., Irfan, M., Glowacz, A., et al.: Evaluation of quality and readability of online health information on high blood pressure using discern and flesch-kincaid tools. Applied Sciences **10**(9), 3214 (2020)
76. Taranova, A., Braschler, M.: Textual complexity as an indicator of document relevance. In: ECIR. LNCS, vol. 12657, pp. 410–417 (2021)
77. Vajjala, S., Meurers, D.: On improving the accuracy of readability classification using insights from second language acquisition. In: seventh workshop on building educational applications using NLP. pp. 163–173 (2012)
78. Vajjala, S., Meurers, D.: On the applicability of readability models to web texts. In: Second Workshop on Predicting and Improving Text Readability for Target Reader Populations. pp. 59–68 (2013)
79. Wang, H.X.: Developing and testing readability measurements for second language learners. Ph.D. thesis, Queensland University of Technology (2016)
80. Westervelf, T.: Wizenoze search white paper. Available at https://cdn.theewf.org/uploads/pdf/Wizenoze-white-paper.pdf (2021)
81. Wizenoze: Wizenoze readability index©. http://www.wizenoze.com (2021)
82. Wojciechowski, A., Gorzynski, K.: A method for measuring similarity of books: a step towards an objective recommender system for readers. In: LTC. pp. 161–174 (2013)
83. Wong, K., Levi, J.R.: Readability of pediatric otolaryngology information by children's hospitals and academic institutions. The Laryngoscope **127**(4), E138–E144 (2017)
84. Xia, M., Kochmar, E., Briscoe, T.: Text readability assessment for second language learners. arXiv preprint arXiv:1906.07580 (2019)
85. Yu, C.H., Miller, R.C.: Enhancing web page readability for non-native readers. In: CHI. pp. 2523–2532 (2010)