Comparing Google’s Readability of Search Results to the Flesch Readability Formulae: A Preliminary Analysis on Children’s Search Queries

Dania Bilal
School of Information Sciences, University of Tennessee
1345 Communications 451, Knoxville, TN 37996
dania@utk.edu

ABSTRACT
Results retrieved by Google on fifteen search queries formulated by middle school children were examined for readability based on Google Reading Level filter (Basic, Intermediate, Advanced). Three hundred retrieved results (links, snippets, and corresponding Web page texts) were analyzed for readability. Google’s assigned reading levels were aggregated at the query level and across the queries and averaged using percentages. The readability scores and grade levels calculated for these results using the Flesch Reading Ease and Flesch-Kincaid Grade Level were averaged at the query level and across the queries and compared to Google’s reading levels. On most queries, a mismatch between Google’s assigned reading levels and the Flesch scores/grade levels was found. The Flesch formulae predicted that a high number of results Google retrieved on the queries were Fairly Difficult, Difficult, or Very Confusing to children in middle grades. A fair number of results Google retrieved did not have assigned reading levels. Implications are made for Google’s reading algorithm, as well as for teachers, parents, and school librarians in guiding children’s use of Google.

INTRODUCTION
The exponential growth of information on the Web and the content of documents indexed by leading web search engines (Google, Yahoo!, and Bing) raise issues about the reading difficulty of text for web users of all ages. While traditional reading skills remain necessary, researchers argue that as Web users, children and adolescents must possess an additional set of reading strategies, dispositions, and digital literacy skills in an effort to locate, read, comprehend, derive meaning, and use information effectively (Coiro & Dobler, 2007). In the past few years, research on the readability of text retrieved from the Web has increased significantly (Benjamin, 2012). Generally, search engines designed specifically for children (i.e., Yahoo! Kids, Ask Kids) index web documents that are deemed suitable for children’s reading and age levels. This is not the case for Google, Yahoo! or Bing as these engines are designed for use by the general public rather than by children or adolescents (Bilal, 2012). Recent work on reading difficulty of websites (Schutten & McFarland, 2009), personalization of web search results by reading level (Collins-Thompson, Bennett, White, De la Chica, & Sontag, 2011), design of probabilistic models to predict web topic distribution based on reading level (Kim, Collins-Thompson, Bennett, & Dumais, 2012), and personalization of content selection based on text comprehensibility (Tan, Gabrilovich, & Pang, 2012) have unraveled how personalization of Web results by reading level can increase relevance and retrieval performance by search engines. Research on children’s interaction with Google continues to show that while these young users prefer Google to other engines, they experience difficulties, including ability to interpret search results retrieved by the engine (Druin, et al, 2009; Druin, et. al, 2010; Foss, et al., 2012). Could the reading level of these results be at the crux of the difficulty children experience in using Google?

In 2010, Google developed a filtering feature by three reading levels: Basic, Intermediate, and Advanced. Google involved a group of teachers in rating the readability level of sample Web pages and, subsequently, developed an algorithm that computes the readability level of retrieved results (Schwartz, 2010). To date, no work has examined the reading level of results (links, snippets, and corresponding Web page texts) Google retrieves on children’s search queries. Similarly, no research has evaluated and compared the reading level of these results against the reading difficulty and grade level calculated by the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (F-K) formulae. In recent years, these formulae have been widely researched in an effort to integrate a readability index into Web search results and build model systems and algorithms that predict Web text difficulty (see for example, Miltsakaki & Troutt, 2011;
Guo, Zhang, & Zhai, 2011; Collins-Thompson & Callan, 2004). As Web content continues to increase in complexity, and as children and other young users resort to Google to find information, it becomes necessary to evaluate Google’s reading filter so that we begin to build understanding of how much of what this engine retrieves supports children’s reading abilities. Previous studies (see for example, Kelly & Belkin, 2001) have revealed that adult users spend more time reading documents that they judge as relevant. Conversely, for younger users (i.e., middle school children), the ability to read and comprehend retrieved results may stand as a precursor for judging relevancy, making the examination of Google’s assigned reading levels of all forms of retrieved text on children’s search queries an essential area of investigation.

The purposes of this study are twofold: 1. To examine to what extent Google retrieves results that are appropriate for middle school children’s reading levels, and 2. To compare the reading levels Google assigns to these results against the reading scores generated by the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade level (F-K).

Findings from this study should begin to build understanding of the difficulty middle school children may experience in comprehending the Web text that Google retrieves on search queries they formulate to solve different types of information problems represented in specific real-world tasks. The findings have implications for the design of Google’s Reading Level filter and reading algorithm; as well as for teachers, school librarians, and parents in guiding these young users’ information retrieval in Google.

RESEARCH QUESTIONS
Two research questions have guided this study:

1. To what extent do Google’s retrieved results (links, snippets, and corresponding Web page texts) on children’s search queries mesh with their reading abilities and grade levels?
2. To what extent do Google’s assigned reading levels (Basic, Intermediate, Advanced) to retrieved results (links, snippets, and corresponding Web page texts) on children’s search queries match with the reading and grade level scores generated by the Flesch Reading Ease and Flesch-Kincaid Grade level formulae?

RELATED STUDIES
A handful of studies have investigated the reading complexities of online text and estimated Web text difficulty. Coiro & Dobler (2007) examined the nature of the reading process and the choices made by eleven sixth-grade skilled students while searching within a multilayered Website called 5 Tigers to answer seven questions for an assigned task. The students were selected from six classrooms in three schools based on skills in traditional reading and experience in finding information on the Internet. Data were collected through individual interviews, think-aloud during reading, observations, and field notes. Findings showed that the students were confronted with complexities in reading online texts. That is, self-regulated text construction, assessing relevancy, drawing inferences, and cognitive flexibility.

Miltsakaki & Troutt (2011) describe Read-X, a Web application they designed for locating and evaluating potential reading material using Yahoo! Web services. Read-X searches the Web for text based on keywords input by a user, extracts from matched Web pages, analyzes the readability of each text by using Lix’s, Rix’s, and Coleman-Liau’s formulae, and calculate the grade level for the text or a readability classification (very easy, easy, standard, difficult or very difficult). Read-X also classifies the results by content themes and returns results according to readability levels. Read-X is targeted for adolescents and adults with reading challenges who seek information on the Web that is appropriate for their reading levels. The authors are developing a new reading formula that considers the user’s content knowledge in assigning readability level. Read-X should be tested with younger users to assess its potential in helping children (e.g., in middle school levels) identify results that Web search engines retrieve for their search queries and that are appropriate for their reading and grade levels.

Guo, Zhang, & Zhai (2011) have designed an application for integrating a readability index into Twitter search engine and embedded the two Flesch measures into search results. For each retrieved Twitter result, the user sees the (FRE) readability score and the F-K grade level. The authors are building a trial search engine to test the application on a large data corpus in Twitter.

Collins-Thompson, Bennett, White, de la Chica, & Sontag (2011), Microsoft Research, have investigated whether classifying retrieved results from Web search engines by reading levels could improve relevancy. They extracted a large log data corpus from the Open Directory Project, Kids & Teens Category (~30,000 total queries and ~20,000 unique queries) and characterized user behavior in certain subject areas of the queries and employed statistical models to estimate both user reading proficiency and the difficulty of search results. They found that re-ranking retrieved results by reading proficiency yielded statistically significant gains in the relevancy of these results. Other researchers have also examined the connection between reading levels and relevancy ranking of search.
results. For example, Tan, Gabrilovich, and Pang (2012), Yahoo! Research, developed a comprehensibility classifier to predict Web text difficulty and improve relevance ranking of search results. The classifier assigns a comprehensibility score to any document using a set of readability formulae (Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning FOG, ARI, SMOG, and Coleman-Liau). The authors modeled users’ reading preferences based on search click logs in a search engine and a community question answering forum (CQA). They constructed topic-specific models of users’ reading proficiency based on the logs. The authors found that modeling text comprehensibility could significantly improve relevance ranking of retrieved results and answers in the CQA forum.

The reviewed studies show that Google’s Reading Level of information retrieval on children’s search queries and its effectiveness against the Flesch measures is an under-studied area of research.

METHOD

A quantitative research approach was employed to aggregate in percentage the reading levels (Basic, Intermediate, Advanced) Google assigned to each retrieved result (link, snippet, and corresponding Web page text) on each search query and across query types. The fifteen search queries submitted to Google consisted of five one-word, five two-words, and five natural language. The average text difficulty was calculated across query types.

The generated readability score and grade level for each retrieved result (link-snippet and corresponding Web page text) on each query was based on the Flesch Reading Ease (FRE) and the Flesch-Kincaid Grade Level (F-K) formulae. The score for each link and snippet for each given query was aggregated at the query level and across query types. Similarly, the score for each corresponding Web text page was aggregated at the query level and across query types. Both the average FRE and F-K scores were calculated at the query level and across query types.

Tasks and Search Queries

Two sets of search queries were employed in this study. The researcher compiled the first set through researching the published literature in information science from 1989 to late 2010 pertaining to children’s information behavior, use, and interaction with digital resources. This resulted in a corpus of (n=130) search queries (i.e., query formulations) that children in elementary and middle schools involved in the studies reported in this literature performed. These tasks represented different types of information problems and covered closed and open-ended tasks, simple and complex, one-word, two-words, and phrase/natural language in varied subject domains. The tasks and search queries used in this study include a variation of these queries. The rationale for this variation is to determine whether the readability level of text retrieved by Google varies accordingly. The queries the children constructed for these tasks and were extracted and submitted in Google verbatim as they were reported in this literature.

The second query set (n = 25) was compiled by the researcher’s graduate teaching assistant who was completing a practicum in a middle-school library in the spring of 2011. The student gathered tasks for which children formulated 25 queries to find information on topics assigned by their teachers or selected based on interest. The characteristics of these tasks and corresponding search queries are similar to those extracted from the published literature. The children’s intentions behind the formulations of these queries are unknown since they are not provided in studies from which they were extracted, or elicited from children in the middle school who shared their queries (For more detail, see Bilal, 2012). The fifteen tasks and search query formulations included in this study were selected from both query sets. Only unique (non-redundant) query formulations were included for each given task, resulting in 15 tasks and 15 unique query formulations.

Tasks and one-word query formulations

The tasks children were asked to perform are:
- Search for material on the Muslim pilgrimage Hajj;
- Find information about minority games in Britain;
- How do you find which direction is south using the sun and a stick?; Research high blood pressure; Choose a topic related to medieval times.

Children’s query formulations for these tasks include: Hajj; Rugby; compass; hypertension; and plague, respectively.

Tasks and two-word query formulations

The tasks children were asked to perform are: Pick a topic to research. It could be a trend, a sports figure, anything you want that I approve… If you prefer not to pick a topic, you can research one of the following topics: Tarantulas, Clint Dempsey, tarantula habitat, social networking, underground railroad, or find information about oil reserves.

Children’s query formulations for these tasks include: Clint Dempsey; Tarantula habitat; social networking; underground railroad; and oil reserves, respectively.

Tasks and natural language query formulations

The tasks children were asked to perform include: How does the heart work?; For a class, you need to
learn about medicines that can help people stop smoking; Using the Internet find three of these medicines; In what year was the speed skating event was introduced in the modern Olympics?; How do the types of dreams and daydreams affect how we sleep?; Environmentalists are concerned that the ozone layer is being depleted. Find how the lack of ozone in the earth’s atmosphere is affecting our forests.

Children’s query formulations for these tasks include: How does heart work; medicine stop smoking; dreams affecting sleep; what year speed skating Olympics; and ozone affecting forests, respectively.

Search Query Submission

Retrieved results by Google on five one-word, five two-words, and five natural language queries were analyzed for text readability. Each of the queries was submitted in Google’s search box verbatim as was formulated by the children. The engine’s Preferences feature at the time of the analysis (late 2011- early 2012) was set to ten results per page (Bilal, 2012), and the Reading Level option in Google was activated so that retrieved results indicate the readability level. Sponsored results were excluded for data analysis. Each result retrieved for a given query consisted of a link, snippet, and URL, which were copied and pasted into a notepad document. Each URL of each given result was activated and the corresponding Web page text was analyzed for text readability. Each of the queries was formulated by the children. The engine’s Preferences feature at the time of the analysis (late 2011- early 2012) was set to ten results per page (Bilal, 2012), and the Reading Level option in Google was activated so that retrieved results indicate the readability level. Sponsored results were excluded for data analysis. Each result retrieved for a given query consisted of a link, snippet, and URL, which were copied and pasted into a notepad document. Each URL of each given result was activated and the corresponding Web page text was analyzed for text readability.

However, most recently, the Search Tools feature was removed from the default interface and the Reading Level feature has been transferred to the Advanced Search interface, making this option invisible to users accustomed to using the default interface. Finding the Readability Level feature is not intuitive in Google.

Flesch Reading Ease (FRE)

Rudolph Flesch developed the Flesch Reading Ease formula (FRE) in 1948 to determine the comprehensibility level of written text. He introduced this formula in an article he published in Journal of Applied Psychology in 1948. The calculation of text readability is based on six factors: 1) the number of sentences, 2) the number of words, 3) the number of complex words, 4) the percent of complex words, 5) the average words per sentence, and 6) the average syllables per word, as applicable. Flesch Reading Ease formula considers “periods, exclamation marks, colons and semicolons as sentence delimiters.” In addition, it has these special rules: “each group of continuous non-blank characters with beginning and ending punctuation removed counts as a word; each vowel in a word is considered one syllable subject to: (a) -es, -ed and -e (except -le) endings are ignored; (b) words of three letters or shorter count as single syllables; and (c) consecutive vowels count as one syllable.” (http://www.readabilityformulas.com/flesch-reading-ease-readability-formula.php). The Flesch Reading Ease formula provides a reading score based on a scale ranging from 0-100. The higher the score, the easier the text is to read. For example, a score ranging from 0-29 indicates that the text is very difficult to read, whereas a score ranging from 90-100 means that the text is very easy to read. (http://www.readable.com/check.php).

Flesch-Kincaid Grade Level (F-K)

The Flesch-Kincaid Grade Level (F-K) formula expands on the Flesch Reading Ease formula. John P. Kincaid enhanced the work of Rudolph Flesch and computed a grade level of readability score for English language text. This score translates to a grade level, meaning the number of years in schooling a person needs to be able to read the text, which is based on the U.S. public school system (K-12). For example, an F-K score of 8 means that an eighth grader will be able to read the text. This formula is used in the U.S. Department of Defense and the field of education. (Lefkovics, 2010). Microsoft Word and Outlook also employ these formulae to display text readability.

Despite their limitations (e.g., developed for print text; ignores context, reader’s interest and prior knowledge; requires 100 words minimum), the Flesch formulae have persisted in the Web environment as evidenced in recent studies that have employed them to...
develop models for predicting the readability of Web page text. In addition, these formulae are considered as the most reliable, especially when used to measure text difficulty for upper elementary to secondary materials (Guo, Zhang, & Zhai, 2011; Bravo, 2010).

Data Filtering

The Flesch formulae have general and special rules that should be followed in order to obtain a readability score and grade level for a given text as accurately as possible. Therefore, the punctuations (e.g., three dots (...) in snippets; colons (:); dashes (--); slashes (/; //)) displayed in each retrieved result for a given query were removed. In addition, the http addresses were excluded from analysis due to their embedded punctuations (slashes, colons, and dots).

Text Analysis in Flesch Reading Ease and Flesch-Kincaid Grade Level

The total number of links and snippets analyzed was 150 (15 queries*10 results per page per query) and 150 corresponding Web page texts, totaling 300 results. Each link and snippet for a given query was copied from the saved notepad document and uploaded into the Flesch-Kincaid online test tool located at, (http://www.readabilityformulas.com/free-readability-formula-tests.php). This website provides a reading score and grade level based on different formulae including the Flesch Reading Ease, Flesch-Kincaid Reading Level, Gunning FOG, Coleman-Liau Index, SMOG index, and Automated Readability Index. In addition, it calculates a Readability Consensus for the text’s reading level, grade level, and the appropriate reader’s age. While there are several readability applications available on the Web, the readability application employed by this website proved to be the most reliable based on the pilot-testing the author performed. The reading scores of a sample of five snippets and their corresponding web page texts were calculated manually using the Flesch Reading Ease formula and these scores were compared to the scores generated by the application on this website and two other applications available on the Web. The scores generated manually conformed more closely to those calculated by the selected application than the other two applications.

A table in Excel was created for each query, which contains each retrieved result (1-10), the Flesch Reading Ease of the link and snippet, Flesch-Kincaid Reading Level for the link and snippet, Flesch Reading Ease for the corresponding web page text of the activated URL, Flesch-Kincaid Reading Level of the same page text, and Google’s Reading Level for each retrieved result on a given query.

RESULTS

The results are provided in the context of the research questions posed in this study.

To what extent do Google’s retrieved results (links, snippets, and corresponding Web page texts) on children’s search queries mesh with their reading abilities and grade levels?

Table 1 shows the average percent of results Google retrieved by Reading Level. On one-word queries, 32% of the results were designated as Advanced, 34% as Intermediate, and 14% as Basic. Assuming that children at the middle school level who are able to read text at the Intermediate Level are also willing to read text at a lower level (i.e., Basic), they will have 48% in retrieved results appropriate for their reading levels. However, one should not assume that middle school children who possess adequate reading abilities are willing to read text at a level lower than their abilities. This assumption should be tested in future work.

On the five two-word queries, Google retrieved an average of 8% in results (including Web page text) at the Advanced Reading Level, 50% at the Intermediate Level, and 32% at the Basic Level. Here, Google yielded a higher average percent in results appropriate for the children’s reading levels. In combining the average percent of results at the Intermediate and Advanced levels, and assuming that children who possess adequate reading skills at the Intermediate level are also willing to read text at the Basic level, they will have 82% of the text suitable for their reading abilities. Note that the reading level of information retrieved by Google across all fifteen queries has not been analyzed vis-à-vis the relevancy of these results. This analysis is beyond the scope of this study.

As to the five natural language queries, Google retrieved a lower percent in average results (20%) at the Intermediate reading level, whereas it yielded a higher percent (38%) at the Basic Level. An equal percent (38%) of results was produced at the Advanced reading level, meaning that it is scholarly in nature and, therefore, inappropriate for the children’s reading level. As natural language dominates children’s search formulations (Bilal, 2000; Foss, et al., 2012) and they expect to find information relevant to their needs, Google’s retrieval on these types of queries may be disappointing to them.

A fair number of results Google retrieved on the queries did not have designated reading levels. As shown in Table 1, 12% of retrieved results on one-word queries, 10% on two-word queries, and 4% on natural language queries did not have assigned reading levels. Although Google does not assign a reading level to most media it retrieves (e.g., videos), media items retrieved on the queries were excluded from
analysis. This means that even in using the Readability Level filter to personalize search results for children’s reading abilities, this personalization will be incomplete.

Table 1. Average text readability in Google by query type.

<table>
<thead>
<tr>
<th>Queries</th>
<th>A</th>
<th>I</th>
<th>B</th>
<th>Not designated</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-Word (n=5)</td>
<td>32%</td>
<td>34%</td>
<td>14%</td>
<td>22%</td>
</tr>
<tr>
<td>Two-words (n=5)</td>
<td>8%</td>
<td>50%</td>
<td>32%</td>
<td>10%</td>
</tr>
<tr>
<td>Natural language (n=5)</td>
<td>38%</td>
<td>20%</td>
<td>38%</td>
<td>4%</td>
</tr>
</tbody>
</table>

Google’s Readability Level (A=Advanced; I=Intermediate; B=Basic).

To what extent do Google’s assigned reading levels (Basic, Intermediate, Advanced) to retrieved results (links, snippets, and corresponding Web page texts) on children’s search queries match with the reading and grade level scores generated by the Flesch Reading Ease and Flesch-Kincaid Grade level formulae?

Table 2 displays the average FRE and F-K reading scores and grade level on each one-word search query, as well as the percent of reading at the three reading levels assigned by Google. Based on the FRE formula, the lower the score, the more difficult is the text. Across the one-word queries, a vertical view of the FRE average scores of the link and snippet shows that the average scores range from 29.95 (Very Confusing to Difficult) to 55.71 (Fairly Difficult). For this text, the F-K grade level, as shown in Table 2, ranges from 9.07 (9th grade) to 11.59 (11th grade). As to the corresponding Web page text, the FRE scores (vertical view) range from 30.06 (Difficult) to 53.69 (Fairly Difficult), and the F-K Grade Level ranges from 11.03 (11th grade) to 17.18 (college level). It seems that Flesch scores and grade levels of results retrieved on the one-word queries could be at a higher reading difficulty than what Google assigned. The fact that 12% of the results did not have assigned reading levels by Google should be taken into consideration in comparing these results.

On the two-word queries (Table 3), a vertical view of the FRE average scores of links and snippet Google retrieved across the five queries ranged from 29.15 (Very Confusing) to 58.86 (Fairly Difficult), and the corresponding F-K grade level ranged from 15.28 (college level) to 12.64 (12th grade and up), respectively.

Table 2. FRE and F-K average readability scores vs. Google’s readability level on one-word queries.

<table>
<thead>
<tr>
<th>Query</th>
<th>FRE</th>
<th>F-K Grade level</th>
<th>FRE</th>
<th>F-K Grade level</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plague</td>
<td>29.95</td>
<td>11.59</td>
<td>43.73</td>
<td>11.03</td>
<td>A 50% ; I 40% ; n.d. 10%</td>
</tr>
<tr>
<td>Hajj</td>
<td>55.71</td>
<td>9.07</td>
<td>52.43</td>
<td>11.23</td>
<td>A 20% ; I 60% ; B 10% ; n.d. 10%</td>
</tr>
<tr>
<td>Compass</td>
<td>37.90</td>
<td>10.81</td>
<td>30.06</td>
<td>14.31</td>
<td>A 10% ; I 60% ; B 30%</td>
</tr>
<tr>
<td>Rugby</td>
<td>35.26</td>
<td>10.62</td>
<td>33.69</td>
<td>17.18</td>
<td>I 40% ; B 30% ; n.d. 30%</td>
</tr>
<tr>
<td>Hypertension</td>
<td>30.26</td>
<td>11.40</td>
<td>34.93</td>
<td>13.64</td>
<td>A 80% ; I 10% ; n.d. 10%</td>
</tr>
</tbody>
</table>

FRE=Flesch Reading Ease; F-K=Flesch-Kincaid Reading Level; Google (A=Advanced; I=Intermediate; B=Basic); n.d.=not designated; F-K Grade Level scale: 0-29 (Very Confusing); 30-49 (Difficult); 50-59 (Fairly Difficult): 60-69 (Standard); 70-79 (Fairly Easy); 80-89 (Easy); 90-100 (Very Easy).

This finding means that the Web page text retrieved on these queries could be more difficult to read by middle school children than the links and snippets. Conversely, Google’s average reading of these results was 50% at the Intermediate Level, 32% at the Basic level, and 8% at the Advanced Level. This finding indicates a mismatch between Google’s readability and the Flesch measures.

Table 3. FRE and F-K average readability scores vs. Google’s readability level on two-words queries.

<table>
<thead>
<tr>
<th>Query</th>
<th>FRE</th>
<th>F-K Grade level</th>
<th>FRE</th>
<th>F-K Grade level</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clint Dempsey</td>
<td>58.86</td>
<td>8.50</td>
<td>48.47</td>
<td>12.64</td>
<td>I 40% ; B 30% ; n.d. 30%</td>
</tr>
<tr>
<td>Oil reserves</td>
<td>43.35</td>
<td>10.34</td>
<td>32.28</td>
<td>14.79</td>
<td>A 30% ; I 60% ; n.d. 10%</td>
</tr>
<tr>
<td>Social networking</td>
<td>29.15</td>
<td>11.35</td>
<td>25.84</td>
<td>15.28</td>
<td>A 10% ; I 60% ; B 30%</td>
</tr>
<tr>
<td>Tarantula habitat</td>
<td>29.37</td>
<td>11.11</td>
<td>45.00</td>
<td>10.76</td>
<td>160% ; B 30% ; n.d. 10%</td>
</tr>
<tr>
<td>Underground railroad</td>
<td>38.87</td>
<td>10.11</td>
<td>36.01</td>
<td>14.64</td>
<td>I 30% ; B 70%</td>
</tr>
</tbody>
</table>

FRE=Flesch Reading Ease; F-K=Flesch-Kincaid Reading Level; Google (A=Advanced; I=Intermediate; B=Basic); n.d.=not designated; F-K Grade Level scale: 0-29 (Very Confusing); 30-49 (Difficult); 50-59 (Fairly Difficult); 60-69 (Standard); 70-79 (Fairly Easy); 80-89 (Easy); 90-100 (Very Easy).
While 4-5 sentences of text may satisfy the requirement for calculating the FRE scores and F-K grade level, it is unknown how much text (i.e., words, syllables, or other) is required in order to predict readability based on Google’s algorithm. As seen in Table 3, there is a fairly high average percent of results (10% to 30%) for which Google did not assign a reading level. Again, reasons for this finding are unknown. In addition, a high average percent of results (60%) Google retrieved on Oil Reserves, Social Networking, and Tarantula Habitat were assigned at the Intermediate reading level that, based on the F-K formula, they could be appropriate for grades 14th to 16th, a much higher reading level for middle school grades. Moreover, Google retrieved an average of 70% of results on the query Underground Railroad assigned as Basic reading level, contrary to the F-K prediction as suitable for the 14th grade level. On the positive side, however, an average of 30% of results Google labeled at the Basic level on the query, Tarantula Habitat, matched with the F-K level as suitable for the 10th to 11th grade levels.

Table 4 shows a vertical view of the average FRE scores for the links and snippets on natural language queries ranging from 27.35 (Very Confusing) to 54.48 (Fairly Difficult). The F-K Grade Level of these results shows a range from the 9th to 13th grade levels. As to the corresponding Web page text, FRE scores ranged from 28.63 (Very Confusing) to 67.39 (Standard reading level); and the F-K Grade Level ranged from the 7th to nearly 15th grade levels. As seen in Table 4, an average of 70% of the results Google retrieved on the query, Ozone Affecting Forest, are assigned as Advanced, matching with the F-K grade level (14.90 or nearly grade 15th). On the query, How Does Heart Work, an average of 50% of what Google retrieved was labeled as Advanced, in contrast with the F-K formula that predicted difficulty of the text at the 7th grade level.

<table>
<thead>
<tr>
<th>Query</th>
<th>FRE Score</th>
<th>F-K Grade</th>
<th>FRE Score</th>
<th>F-K Grade</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dreams affecting sleep</td>
<td>35.80</td>
<td>12.23</td>
<td>51.44</td>
<td>10.48</td>
<td>A 30%; I 40%; B 20%; n.d. 10%</td>
</tr>
<tr>
<td>What year speed skating Olympics</td>
<td>38.17</td>
<td>11.36</td>
<td>48.91</td>
<td>12.15</td>
<td>I 20%; B 80%</td>
</tr>
<tr>
<td>Ozone affecting forest</td>
<td>41.18</td>
<td>11.42</td>
<td>28.63</td>
<td>14.90</td>
<td>A 70%; B 30%</td>
</tr>
<tr>
<td>How does heart work</td>
<td>54.48</td>
<td>9.45</td>
<td>67.39</td>
<td>7.26</td>
<td>A 50%; I 20%; B 20%; n.d. 10%</td>
</tr>
<tr>
<td>Medicine stops smoking</td>
<td>27.35</td>
<td>13.16</td>
<td>54.13</td>
<td>9.46</td>
<td>A 40%; I 20%; B 40%</td>
</tr>
</tbody>
</table>

DISCUSSION

Reading and understanding the different forms of text (links, snippets, Web page text) retrieved by Google on children’s search queries could be, in many cases, too complex for middle school children to comprehend. The Flesch formulae predicted that the retrieved results across the fifteen search queries would be difficult for the middle school grade levels.

While children at all levels should be exposed to textual materials that could stretch their thinking and facilitate “deeper levels of processing and meaningful construction of knowledge,” (Coiro & Dobler, 2007) middle school children who possess “standard” reading skills could struggle in their pursuit to comprehend and derive meaning from different forms of text retrieved by Google. Previous studies by Druin, et al. (2009; 2010) and Foss, et al. (2012) have revealed that while middle school children were able to search Google, they experienced difficulty interpreting the results retrieved for their search queries. Based on the findings of the present study, it is possible that this difficulty is associated with the complexity of most Web text that Google indexes. Future studies should investigate this association to determine whether a significant difference exists between the ability of children to interpret results and the complexity of the text retrieved by Google.

This study revealed that based on the Flesch formulae, most forms of text retrieved by Google across the fifteen queries would be advanced, difficult, and very confusing for middle school children. Of the 300 retrieved results (links/snippets and corresponding web pages) only one result was at the standard reading level of 7th graders. While text between 8th to 10th grade levels might be manageable by skilled readers, text at higher levels could be extremely difficult for these children’s cognitive abilities to process and comprehend. If this difficulty is associated with the content of the landing sites and Web pages that Google’s spiders collect from the Web and from which snippets are extracted, the algorithm for crawling the
Web should be modified to augment the indexing of sites and pages designed for children.

The mismatch between Google’s assigned readability to text and the F-K Grade Level found in this study suggest that either the engine’s Reading Level algorithm is inadequate, or that the Flesch formulae may not be valid for predicting the readability of Web text (links, snippets, and corresponding pages) since it was originally developed for measuring the readability of print instead of Web text.

Google’s Readability Level scale (Basic, Intermediate, Advanced) does not provide sufficient variation between the three levels. Specifically, the threshold for the Intermediate level is too narrow for judging the readability level for lower and upper intermediate grade levels. Improving Google’s Reading filter may be achieved by widening the reading scale to four levels (Basic, Lower Intermediate, Upper Intermediate, Advanced) to support a better personalization of results. In addition, making the Reading Level filter available and visible in the default search interface rather than only embedded in the Advanced interface could expand its use by mediators and integrate its use in information literacy programs.

The findings of this study have implications for teachers, school librarians, and parents. Teachers and school librarians should provide effective information literacy training to children in how select relevant results from Google that are appropriate for their reading levels. While training children in using the Readability Level filter could alleviate some of the difficulties children experience in using Google, these mediators should recognize the limitations of this feature. Although children favor Google, directing them to search tools and sites designed for their age levels could leverage a better search experience in finding information on the Web.

LIMITATIONS
The main limitation of this study resides in the small number of queries used. Future research should consider a larger corpus of queries to enable generalization of the results.

CONCLUSIONS
This study examined the reading difficulty of search results and web page texts Google retrieves on fifteen search queries formulated by middle school children. It compared the Reading Level Google assigns to results it retrieved on the queries against the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (F-K) measures. The Flesch reading scores and grade level did not match with Google’s assigned reading levels. In most cases, the Flesch formulae predicted that the results and corresponding Web pages Google retrieved on the queries would be difficult for middle school children. It is possible that the Flesch formulae are not appropriate for measuring the readability of Web text, since they were designed to measure print instead of Web text. It is also possible that Google’s reading algorithm needs improvements.

A fair percent of results retrieved by Google across the queries did not have reading levels attached to them. This finding means that even in using the Readability Level filter to personalize search results, children will find that this personalization non-inclusive.

Although the findings of this study are based on fifteen search queries resulting in 300 snippets and corresponding Web page text, they provide a basis for additional research with a larger corpus of queries so that we may extend this investigation to other search engines and find common ground for generalizing the findings.

Research that involves children in the evaluation of Web text readability is needed so that we can compare their reading abilities to Google’s Reading Level and the Flesch formulae based on their perspectives. Involving children in such investigations could result in developing more effective reading algorithms and models for improving not only Google’s readability algorithm, but also the readability of text retrieved by other search engines.

REFERENCES


Coiro, J, & Dobler, E. 2007. Exploring the online reading comprehension strategies used by sixth-
grade skilled readers to search for and locate information on the Internet. *Reading Research Quarterly*, 42(2), 214-257.


