The influence of text characteristics on perceived and actual difficulty of health information

Gondy Leroya,∗, Stephen Helmreichb, James R. Cowiec

a School of Information Systems and Technology, Claremont Graduate University, 130 E. Ninth Street, Claremont, CA 91711, United States
b Department of Information Sciences and Security Systems, Physical Science Laboratory, New Mexico State University, United States
c Department of Psychology, New Mexico State University, United States

ARTICLE INFO

Article history:
Received 11 November 2009
Received in revised form 20 January 2010
Accepted 5 February 2010
Available online xxx

Keywords:
Consumer health information
Communication barrier
Linguistics
Comprehension
Readability
Perceived difficulty
Understanding
Retention

ABSTRACT

Purpose: Willingness and ability to learn from health information in text are crucial for people to be informed and make better medical decisions. These two user characteristics are influenced by the perceived and actual difficulty of text. Our goal is to find text features that are indicative of perceived and actual difficulty so that barriers to reading can be lowered and understanding of information increased.

Methods: We systematically manipulated three text characteristics, – overall sentence structure (active, passive, extraposed-subject, or sentential-subject), noun phrases complexity (simple or complex), and function word density (high or low), – which are more fine-grained metrics to evaluate text than the commonly used readability formulas. We measured perceived difficulty with individual sentences by asking consumers to choose the easiest and most difficult version of a sentence. We measured actual difficulty with entire paragraphs by posing multiple-choice questions to measure understanding and retention of information in easy and difficult versions of the paragraphs.

Results: Based on a study with 86 participants, we found that low noun phrase complexity and high function words density lead to sentences being perceived as simpler. In the sentences with passive, sentential-subject, or extraposed-subject sentences, both main and interaction effects were significant (all \(p < .05\)). In active sentences, only noun phrase complexity mattered (\(p < .001\)). For the same group of participants, simplification of entire paragraphs based on these three linguistic features had only a small effect on understanding (\(p = .99\)) and no effect on retention of information.

Conclusions: Using grammatical text features, we could measure and improve the perceived difficulty of text. In contrast to expectations based on readability formulas, these grammatical manipulations had limited effects on actual difficulty and so were insufficient to simplify the text and improve understanding. Future work will include semantic measures and overall text composition and their effects on perceived and actual difficulty.

Limitations: These results are limited to grammatical features of text. The studies also used only one task, a question-answering task, to measure understanding of information.

© 2010 Published by Elsevier Ireland Ltd.

∗ Corresponding author. Tel.: +1 909 607 3270.
E-mail address: gondy.leroy@cgu.edu (G. Leroy).
1386-5056/$ – see front matter © 2010 Published by Elsevier Ireland Ltd.

1. Introduction

Every day, the number of available medical tests and treatments increases, more patients are diagnosed with chronic diseases that require long-term management, and to make matters worse, healthcare professionals face increasing pressure to see and treat more patients in a limited amount of time. In the US, an estimated 89 million people have insufficient health literacy to understand treatments or preventive care [1] and the costs associated with such limited health literacy are estimated at between $50 and $73 billion per year [2]. Patients with low health literacy are usually less knowledgeable about their condition, e.g., when suffering from diabetes [3], less inclined to search for information, e.g., about colorectal screening [4], and less capable to learn from text [5–7]. Although English and Spanish are the languages most commonly evaluated, comparable conclusions are made for other languages and in other countries, e.g., in Japan, low health literacy was associated with poorer health [8].

Inadequate health literacy is defined by the Institute of Medicine [9] as “limited ability to obtain, process, and understand basic health information and services needed to make appropriate health decisions and follow instructions for treatment.” Although personal interaction with physicians is preferred, the most common method for educating consumers is still providing them with texts. Willingness and ability are two important aspects of becoming health literate and learning from such text. When the available information is too difficult to understand, two problems can be expected to develop. The first, most obvious problem is that consumers will not understand the information [5–7] and will be less informed. The second, less obvious problem is that consumers may resort to reading texts that seem easier to read. This easy-to-read information may be overly represented by less reputable sources, such as blogs or targeted advertising. This is worrying since in general, online health information affects decisions about health, healthcare, and visits to a healthcare provider for at least a third of the readers [10].

Reading and learning from text can be encouraged by providing texts that pose low barriers and that are at an appropriate difficulty level for consumers’ expertise and ability. A low barrier to reading is important to encourage consumers to educate themselves and read the provided materials. An appropriate difficulty level is important to facilitate actual understanding and learning.

2. Related work

2.1. Perceived and actual text difficulty

The consumers’ willingness to read a text is the first, necessary step, toward their education. However, anecdotal evidence suggests that consumers who consider a text too difficult are not inclined to read or study it [11], which is confirmed by many physicians who can list excuses, e.g., “I broke my glasses,” made by patients who did not read provided educational pamphlets. The distinction between perceived and actual difficulty is important since they may independently influence willingness to read and understanding of materials. Evidence for this distinction comes from two research streams.

In medicine, the Health Belief Model (HBM) has been used to explain and predict health-related behaviors. The model contains four dimensions: perceived susceptibility, perceived severity, perceived benefits, and perceived barriers. Although support for the model is mixed, in their review study in 1984, Janz and Becker [12] found that the perceived barriers dimension was the most significant of all four in explaining health behavior. In the context of consumer education, perceived difficulty of the text is a barrier encountered by many consumers who are expected to read text and educate themselves. Velayo’s results [13] point in this direction. He evaluated the effects of presentation media (e.g., text only versus text with graphics) and perceived difficulty on a comprehensive understanding test. He found that media type and perceived difficulty influenced understanding independently.

In psychology, the Theory of Reasoned Action (TRA) has been brought forward as a model to explain behaviors and what determines them. The Theory of Planned Behavior (TPB) can be seen as an extension [14]. The TPB adds perceived behavioral control as a factor to the original model. During recent years, several studies have shown support for the existence of two distinct components in perceived behavioral control: perceived difficulty and perceived control. Trafimow et al. [15] conducted 4 studies that demonstrated that perceived difficulty and perceived control can be manipulated independently while their 5th study showed perceived difficulty to be a stronger predictor of intentions and behaviors. In the medical field, Liu at al. [16] modeled drug information-seeking behavior by patients using questionnaire data for 1000 patients. They found that perceived difficulty was an important factor, sometimes together with self-efficacy, in explaining behaviors.

The HBM and TPB are both useful in guiding readability research. Health literacy relies on reading behaviors, which we believe are easily influenced by both perceived and actual difficulty. It is important to be aware of them so that reading materials can be optimized for both.

2.2. Readability metrics

The understanding consumers gain from reading texts is influenced by many characteristics, some of which can be influenced by writers and other that cannot. Background knowledge, education, language skills, and the reader’s environment are among those that cannot be changed by the writers of texts. However, several other factors can be influenced to increase understanding. For example, the vocabulary used, the complexity of sentences, the density of information in a text, and the composition of documents are among those factors that can be manipulated. Some factors will influence perceived difficulty of a text, others will influence actual difficulty, and some may influence both.

There are several writing guidelines, such as the Manual for Clinicians provided by the American Medical Association Foundation [9], which recommend the use of active voice, low readability grade levels, short paragraphs, and one-syllable or two-syllable words. In addition, these guidelines often use
formulas to assign a readability grade level to a text. The most commonly used formulas are the Flesch Readability Scores and the Flesch-Kincaid Grade Levels. They use word and sentence length as stand-ins for semantic and syntactic complexity or difficulty [17]. The formulas are readily available in Microsoft Word, which may partially explain their popularity. According to the guidelines, educational materials should be written at a 6th or 8th grade level. In response to these guidelines, many health literacy studies, including our own, have measured difficulty levels of text [5–7,18–20] and other online resources [21] using the formulas. The general conclusion of these studies is that most English-language Internet sites require at least a 10th grade (Flesch-Kincaid) reading grade level and many present information at college level [5,22–25]. Although readability formulas are among the most popular metrics, others have evaluated the content and style of written materials, such as completeness of information in the pamphlets [26,27] or cultural sensitivity [28]. An even broader, more inclusive approach is taken by the Evaluative Linguistic Framework, which takes both content and writing style into account [29].

Studies on readability that rely on the readability formulas usually conclude that the information is too difficult to read [5–7]. In the few cases where understanding is measured explicitly, the Cloze procedure [30] is often adopted. This test uses texts from which the nth word is removed. People are asked to fill in the blanks. Using a synonym or related word is considered incorrect. Trifilietti et al. [31] evaluated text of different grade levels using Cloze tests. They found that more people performed at an acceptable level with lower grade level texts. However few studies have been published that use different evaluation methods and show a relationship between grade levels and actual understanding. Weaver and Renken [32] did not find evidence of a relationship between understanding and readability in their review study. Other features, such as data presentation and report format, did not influence understanding; however, the education level of readers did [33]. This is not surprising since understanding requires complex cognitive process, not easily tested with simple tests or manipulated with simple text changes.

It is unclear if readability metrics are sufficient to measure difficulty of text and there is no common, accepted standard to measure understanding. Different metrics, such as Cloze tests, information retrieval tasks, or question-answering tasks, and expert evaluation, are used to evaluate online information and its difficulty on behalf of consumers. Each has its advocates and critics. For example, Birru et al. [34] point out that it is not clear if being able to answer questions by finding texts online indicates understanding. Some have dropped readability from text evaluation because of lack of evidence about its relation to actual understanding [35]. Yan et al. [36] measured both perceived readability and readability grade levels and demonstrated that they do not match very well. Others rely on experts. Provost et al. [37] developed a comprehensive scale for health information evaluation with 8 categories containing 98 items. The items were chosen by experts, who showed strong agreement, and the scale itself remains to be validated (as indicated by the authors). However, Bernstam et al. [38] found that experts do not always agree on seemingly objective text features, such as disclosure of advertising or sponsorship. Kundula and Zeng-Treitler [39] also worked with an expert panel to create a readability gold standard and found the relation with their judgments and the outcome of readability formulas to be unsatisfactory. In our own work, we found that experts and consumers do not always agree on difficulty levels [25], but the expert evaluation was more closely related to readability formula outcomes than the consumer evaluation.

### 2.3. Research interests

Our aim is to evaluate perceived and actual difficulty of text with metrics based on linguistic features. We limit our work to features that can be automated so that, if successful, tools can be developed to assist writers of health information texts. We focus here on grammatical features, but semantic and text composition features will be evaluated later. Ultimately, our work will lead to metrics with the potential to be automated, so that when successful, scalable solutions can be developed and provided to the community.

### 3. Linguistic structures

We aim to evaluate the effect of several linguistic features on difficulty. The first feature we evaluate is the overall sentence structure. Writing guidelines advise the use of active voice instead of passive voice. For example, “Fortunately, lifestyle changes can prevent further damage to coronary arteries” is active while “Fortunately, further damage to coronary arteries can be prevented by lifestyle changes” is passive voice. To our knowledge, no other sentence structures are discussed in the guidelines, or have been measured. However, when browsing consumer documents, such as clinical trials information, we noticed that most sentences use more complicated structures.

Some of the more complicated structures were sentences with a sentential-subject or with an extraposed-subject structure. A sentential-subject forms a sentence on its own, for example: “Identifying molecules involved in the immunologic response will aid researchers design better islet loss prevention methods”. Although the construction seems difficult to understand, it is used regularly in online clinical trials documents. This subject can be moved, extraposed, in the sentence and replaced with “it”, for example “It will aid researchers design better islet loss prevention methods to identify molecules involved in the immunologic response”.

The second feature we evaluate is noun phrase complexity. Noun phrase complexity increases the difficulty of sentences. In many texts, a complex single noun is the base of the noun phrases, e.g., “apoptosis.” Additionally, complex noun phrases are often compound nouns, i.e., a sequence of nouns that forms a larger referring expression, such as “common prostate cancer treatment.” These structures are repeatedly found in medical texts, as opposed to their more explicit equivalents using prepositions: “common treatment of prostate cancer.” Such phrases pose understanding difficulties in determining how the nouns are related (does “common” modify “prostate” or “cancer” or “treatment”?) which can be clarified with prepositional phrases.
The third feature we evaluate is rate of function words in a sentence. We use “function words” in a broad sense to refer to prepositions, wh-words, modals, auxiliaries, and determiners, for example, in, why, could, be, the. Our earlier comparisons between patient blogs and formal documents showed large differences in frequency of occurrence. The percentage of function words was often twice as high in blogs as in formal documents leading to a different cadence closer to spoken language. Those texts with many function words seem intuitively easier to read. A formal explanation is that these words may help space out individual concepts in text to facilitate memorization. In a small pilot study, which found that sentences with a higher percentage of function words were perceived to be easier by consumers [24]. Note that one of the effects of replacing compound nouns with prepositional phrases is to increase the percentage of function words in a sentence.

4. Methods

We designed our study so that we could measure both perceived and actual understanding of text with the same group of participants. The strength of this design is that the subjects are identical, reducing variance in the subject pool and allowing better comparisons of the effects of linguistics features on perceived and actual difficulty. The disadvantage is that both outcomes needed to be measured in a reasonable time, preferably during the same session, and that the measurement of one could not inform the measurement of the other.

We recruited students at undergraduate and graduate institutions in New Mexico and in Southern California. Most were working adults of different ethnicities. Some students were bilingual, from Mexico, and started to learn English in grade school. Participation was voluntary.

We used paper handouts which were assembled in advance. Each packet consisted of 16 pages which contained an informed consent form and four study sections. The sections had to be completed in order and participants could not go back to previous sections. We ordered the sections so that a time delay was introduced between answering questions with the text present (understanding) and without (retention). The order was as follows: the first section measured actual understanding. It contained texts with associated multiple-choice questions. Then, we provided individual sentences to measure perceived difficulty, followed by a short demographic questionnaire. The final section contained more multiple-choice questions about the text read in the beginning to measure retention of information. Each section is described in detail below.

4.1. Demographic questionnaire

The demographic questionnaire contained questions about the participants’ age, gender, native language, and highest level of schooling completed. We also included two questions about the level of education in medicine and linguistics to allow us to exclude participants whose knowledge would influence their results.

4.2. Perceived difficulty

To measure perceived difficulty using realistic text, we selected sentences from clinical trials documents and consumer health websites. Each sentence was adjusted to reflect the three linguistic characteristics discussed above. We evaluated active, passive, sentential-subject, and extraposed-subject sentence structures. Per sentences structure, we looked at noun phrase complexity (simple versus complex) and function word density (high versus low). Complex base nouns were replaced with easier nouns, e.g., “apoptosis” was replaced by “programmed cell death.” Compound nouns were replaced with their prepositional equivalents. Appendix A shows an example sentence for each feature combination.

Ideally, we would construct all possible combinations for each example (4 sentence structures × 2 levels of noun phrase complexity × 2 rates of function word density). However, our pilot testing showed that this was practically impossible. Showing too many version of one sentence made the evaluation nonsensical and so we limited the number of combinations. Each participant received 8 sentences. Half contained a foursome of active sentences and a foursome of passive sentences. The other half contained a foursome of sentential-subject sentences and a foursome of extraposed-subject sentences. Each foursome had simple versus complex noun phrases and high versus low function word density. For each foursome, we asked participants to select the easiest and the most difficult of the four. This provided us with two measures of perceived difficulty: easy and difficult. We also asked participants to compare the foursome with active sentences to the foursome with passive sentences, and similarly the foursome with sentential-subject sentences to the foursome with extraposed-subject sentences.

To reduce other possible sources of variance, we controlled, as much as possible, the Flesh-Kincaid Readability Grade Levels and the word length of the sentences (Table 1). We conducted an ANOVA with sentence structure, noun phrase complexity, and function word density as independent variables for both Flesh-Kincaid scores and sentence length to verify our control. There are four example sentences in each condition (N = 4). As expected and unavoidable, the Flesch-Kincaid scores differed for noun phrase complexity, with more complex noun phrases leading to higher readability scores (F(1,48) = 8.33, p < .01). Similarly, sentence length differed for function word density (F(1,48) = 7.23, p = .01) with sentences with higher function word density being longer.

4.3. Actual difficulty—understanding and retention

To measure actual difficulty, we chose two paragraphs that were available online as part of the study description of clinical trials: one discussed heart disease and one discussed depression. The use of paragraphs instead of individual sentences has higher external validity; it is much closer to actual reading behaviors of consumers. However, some control is lost since the paragraphs cannot be as easily manipulated as individual sentences. The final paragraph still has to flow, has a consistent style, and different versions need to provide the same information. Appendix B shows the different paragraphs.

Table 1 – Average Flesh-Kincaid Grade Levels and word counts (N = 4).

<table>
<thead>
<tr>
<th>Sentence structure</th>
<th>NP complexity</th>
<th>Function word density</th>
<th>Flesch-Kincaid GL</th>
<th>Word ct.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>Simple</td>
<td>High</td>
<td>13.0</td>
<td>21.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>12.8</td>
<td>16.8</td>
</tr>
<tr>
<td>Complex</td>
<td>High</td>
<td>16.5</td>
<td></td>
<td>20.0</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>15.8</td>
<td></td>
<td>15.5</td>
</tr>
<tr>
<td>Passive</td>
<td>Simple</td>
<td>High</td>
<td>12.8</td>
<td>23.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>12.9</td>
<td>19.0</td>
</tr>
<tr>
<td>Complex</td>
<td>High</td>
<td>15.1</td>
<td></td>
<td>21.3</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>14.8</td>
<td></td>
<td>17.0</td>
</tr>
<tr>
<td>Sentential</td>
<td>Simple</td>
<td>High</td>
<td>12.3</td>
<td>20.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>11.7</td>
<td>16.3</td>
</tr>
<tr>
<td>Complex</td>
<td>High</td>
<td>14.6</td>
<td></td>
<td>18.3</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>12.8</td>
<td></td>
<td>14.8</td>
</tr>
<tr>
<td>Extraposed</td>
<td>Simple</td>
<td>High</td>
<td>11.3</td>
<td>21.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>11.0</td>
<td>17.5</td>
</tr>
<tr>
<td>Complex</td>
<td>High</td>
<td>13.1</td>
<td></td>
<td>20.0</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>12.1</td>
<td></td>
<td>16.5</td>
</tr>
</tbody>
</table>

Table 2 – Descriptive statistics for the four texts.

<table>
<thead>
<tr>
<th></th>
<th>Heart disease</th>
<th>Depression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Difficult (original)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Words</td>
<td>329</td>
<td>239</td>
</tr>
<tr>
<td>Noun phrases</td>
<td>81</td>
<td>56</td>
</tr>
<tr>
<td>Flesch-Kincaid Grade Level</td>
<td>15.8</td>
<td>16.6</td>
</tr>
</tbody>
</table>

The original version of each paragraph was considered the difficult version. We constructed an easy version for each by simplifying noun phrases, adding function words, and simplifying sentence structures by changing them as much as possible into active sentences. Table 2 shows summary statistics for the 4 paragraphs. The easy versions each had 10 sentences, while the difficult versions had 8 sentences. The easier versions used many more shorter noun phrases because compound noun phrases were split up. Our count represents this, e.g., a prepositional phrase such as “a complete resolution of symptoms with antidepressant treatment” is composed of three noun phrases.

For each paragraphs, we constructed 3 questions to measure understanding. The first question was about the overall goal of the paragraph, the following two questioned about possible conclusions that could be drawn from the text. The questions were the same for the easy and difficult versions, i.e., questions were not rephrased. These questions were presented together with the text to measure understanding. Seven additional questions were asked toward the end of the study, without the text present, to measure retention.

To avoid a carry-over and learning effect, we did not present the easy and difficult version of the same document to each participant. However, to maximize our data points, we provided each participant with a difficult and easy text: half of the group received the easy heart disease document and the difficult depression document; the other half received the difficult heart disease document and easy depression document. The order of the documents was also reversed within each group. We calculated the number of correctly answered questions for both understanding and retention. We also asked each participant to choose from which texts they felt they remembered most.

5. Results

5.1. Demographic information

A total of 97 volunteers participated. The data of eleven subjects were discarded because they did not follow the instructions completely, e.g., some did not evaluate each example of the sentences. No participants had more than introductory knowledge of medicine or linguistics. The following analysis is based on the remaining 86 complete evaluations.

The detailed demographic data can be found in Table 3. Both genders were almost equally represented with 56% of participants being female and 43% male. The average age was 26 years for the 83 participants who reported age. The youngest participant was 17 years old while the oldest was 72 years old. English was the most common native language spoken by 56% of the participants, followed by Spanish which was spoken by 25% of the participants. Five other languages were reported as native language, with Chinese being the...
most common (10%) and followed by Arabic, Kazak, Italian, and Ibo. Two participants did not report their native language.

Slightly more than half of the participants (55.8%) had not completed any degree beyond high school: 5.8% had a high school diploma, while 18.6% and 31.4% had some community college or some college education. The other half of the participants had completed some degree (44.2%) with most having a bachelor’s degree (17.4%), very few with a doctoral degree (4.7%) and about equal numbers with an associate’s (11.6%) or master’s (10.5%) degree.

5.2. Perceived difficulty

To evaluate perceived difficulty, we counted how often a particular version of a sentence in a foursome was selected by the participants as the easiest or the most difficult. Perceived difficulty is low when many participants indicate a sentence as the easiest. Perceived difficulty is high when many participants indicate a sentence as the most difficult. (Note: scores for ‘difficult’ have been reported as part of Ref. [40].) The scores are averaged over the four examples shown to each participant and over all participants.

Fig. 1 shows the results for all four sentence structures. Overall, we found very strong effects of both noun phrase complexity and function word density. The ‘easy’ and ‘difficult’ scores mirror each other as expected. Although not all effects are significant, the figures show how they convey the same message: sentences that are considered easy by most participants are considered difficult by few participants, and vice versa. Table 4 shows how the effects of noun phrase complexity and function word density are more pronounced in sentences with more difficult structures: there are few significant differences in the active sentences but many in the other sentence structures.

Since each participant chose between all the four options for each sentence structure, we conducted a repeated-measures analysis of variance per sentence structure (active, passive, sentential-subject, extraposed-subject) with two independent variables: noun phrase complexity (NP) and function word density (FW). We conducted these studies once for each dependent variable: “Easy” or the percentage of participants who chose a particular structure as the easiest and “Difficulty” or the percentage of participants who chose a particular structure as the most difficult. We report only the significant main and interaction effects (Table 4). We included the exact p-values for significant effects; effects that were not significant were not included. Since no post hoc comparisons were needed, i.e., only 2 levels per variable, no Bonferroni corrections are included. However, from the p-values, it is clear that even with an adjustment (α/4 for each sentence structure), most effects would still be significant and since all differences point in the same direction, the conclusions remain unchanged.

For active sentences, which are usually considered the easiest, we found only an effect for noun phrase complexity (p < .001). As can be seen in Fig. 1, sentences with more complex noun phrases are selected by more participants as the most difficult in the foursome, e.g., by 26% of the participants when there was high function word density and by 33% when there was low function word density.

For passive sentences, we found a main effect for noun phrase complexity (p < .05), a main effect for function word density (p < .001), and also an interaction effect (p < .05) for the difficulty ratings. Low function word density and high noun phrase complexity are more difficult. The effect of function words is stronger in sentences with complex noun phrases. Although the data look complementary for the easiness rating in Fig. 1, Table 4 shows that with easiness ratings only the effect for function word density was significant (p < .001).

For sentential-subject and extraposed-subject sentences, both noun phrase complexity and function word density had a strong effect. Again, the results show mirrored patterns, but the differences are not always significant. For sentential-subject sentence difficulty scores, only the noun phrase complexity was significant (p < .001). However, when choosing the easiest sentence, both noun phrase complexity

<table>
<thead>
<tr>
<th>Variable</th>
<th>Values</th>
<th>N</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Female</td>
<td>49</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>37</td>
<td>43</td>
</tr>
<tr>
<td>Native Language</td>
<td>Arabic</td>
<td>2</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>8</td>
<td>9.3</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>47</td>
<td>54.7</td>
</tr>
<tr>
<td></td>
<td>English/Spanish</td>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>Ibo</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>Italian</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>Kazak</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>21</td>
<td>24.4</td>
</tr>
<tr>
<td></td>
<td>Unreported</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Highest schooling completed</td>
<td>High school</td>
<td>5</td>
<td>5.8</td>
</tr>
<tr>
<td></td>
<td>Some community college</td>
<td>16</td>
<td>18.6</td>
</tr>
<tr>
<td></td>
<td>Some college</td>
<td>27</td>
<td>31.4</td>
</tr>
<tr>
<td></td>
<td>Community college associate’s degree</td>
<td>10</td>
<td>11.6</td>
</tr>
<tr>
<td></td>
<td>Bachelor’s degree</td>
<td>15</td>
<td>17.4</td>
</tr>
<tr>
<td></td>
<td>Master’s degree</td>
<td>9</td>
<td>10.5</td>
</tr>
<tr>
<td></td>
<td>Doctorate</td>
<td>4</td>
<td>4.7</td>
</tr>
</tbody>
</table>

and function word density affected the scores. Their interaction was also significant ($p < .05$): Fig. 1 shows how the effects for easiness are stronger when noun phrases are simple.

Finally, sentences with an extraposed-subject showed most effects. For the easiness scores, both noun phrase complexity ($p < .001$) and function word density ($p < .01$) affected the scores. For the difficulty scores, both main effects ($p < .001$ and .05) and the interaction effect ($p < .05$) were significant, again indicating a difference in strengths of effects depending on noun phrase complexity.

When asked to compare the active with passive foursome, 58% of the participants found the active sentences the easiest; the other 38% indicated the passive sentences as the easiest. A few participants indicated both or none of groups as the easiest (4%). When asked to compare the sentential-subject with the extraposed-subject sentences, 63% found the sentential-subject sentence the easiest and 34% found the extraposed-subject sentences. Again, a few comparisons were invalid because both or no group was chosen (3%).

### 5.3. Actual difficulty: understanding and retention

Actual difficulty was evaluated by calculating the number of questions that could be answered correctly (under-
standing) and without (retention) the text. Fig. 2 shows an overview of the scores for the different documents and difficulty levels. Overall, participants scored slightly more than 50% correct when the text was present (understanding). The scores for the depression document are slightly higher but this difference was not significant. With regard to the text difficulty level, we conducted an ANOVA and found only a trend, i.e., a nearly significant effect, that the difficulty level influenced understanding ($F(1,168) = 2.758, p = .099$). These differences have nearly disappeared for the retention scores and no significant effects were found. To compare the understanding and retention scores of individuals, we calculated Pearson’s $r$, which tests linear associations, and found them to be significantly correlated (Pearson $r = .388$, $p < .001$).

For the subjective evaluation, participants judged the depression document to be much easier than the heart disease document. This was the case regardless of the version (easy or difficult) they received. For those participants who received an easy depression document and a difficult heart disease document, 87% ($N = 38$, 5 did not answer the question) found the depression document to be the easiest. Surprisingly, for those participants who received the difficult depression document and the easy heart disease document, 60% ($N = 43$), still found the depression document to be easier. This contradicts what we would expect based on readability grade levels, since the difficult depression document was graded higher than the easy heart disease document. Even though participants indicated that perceived difficulty was different, the test scores do not indicate that actual difficulty was different, i.e., scores for the easy heart disease and difficult depression document are equal (57%).

5.4. Influence of demographic characteristics

To evaluate whether demographic characteristics influence the scores for perceived and actual difficulty, we calculated the effect of schooling and gender on perceived and actual difficulty. We recoded the schooling level by numbering each level in as follows: (1) high school, (2) some community college, (3) some college, (4) associate degree, (5) bachelor’s degree, (6) master’s degree, and (7) doctoral degree. There was no correlation between schooling and perceived difficulty scores within each foursome. An ANOVA with schooling as independent variable showed no effect on understanding or retention. However, since this is a post hoc variable, it had an unbalanced $N$. Therefore we completed this analysis by calculating Spearman’s rho, which is a nonlinear correlation coefficient. We chose this nonlinear coefficient since it can be argued whether ‘some college’ should be scored lower than ‘associate degree’. We did not find any significant correlation between schooling and the actual difficulty scores (understanding and retention) within each foursome. However, as indicated above, scores for understanding and retention were linearly correlated, indicating that people who did well on the understanding also did well on the retention task. Finally, a post hoc ANOVA with gender as the independent variable showed no differences between males and females in understanding or retention.

6. Conclusions

The goal of our work was to look at linguistic features of text that may influence perceived and actual difficulty of text. Since today’s computational resources are much more advanced than those available when the first readability formulas were conceived, we intentionally focused on features that can augment the formulas and that would provide very precise advice, i.e., pinpoint areas in text, on how to simplify text. Our intention was also to draw a clear line between perceived and actual difficulty. We tested perceived difficulty by asking participants’ feedback on sentences. We tested actual difficulty with question-answering tasks. We believe our work to be among the first to clearly distinguish between perceived and actual difficulty and measure both. Our results show that simple surface metrics are useful to help improve perceived difficulty, but reducing actual difficulty requires more than surface changes to text.

Perceived difficulty was heavily influenced by our sentence manipulations. Especially sentences with more complex overall structures were influenced by noun phrase complexity and function word density. Since perceived difficulty influences the first impression and so may form a barrier to reading of a document, more work leading to precise and supportive writing tools can help writers provide texts that will not deter consumers from starting to read. The effects of these linguistic variations diminish when more cognitive processing (understanding) and additional memory skills (retention) are needed.
In our studies, this may be due to the limitations of the text we worked with—longer texts may show stronger effects and provide more insights. However we believe other aspects, such as topic density, e.g., rate of different concepts per text unit, and text composition, will influence understanding more. Paragraphs structure and writing styles, e.g., an overview sentence to start a paragraph, may also help understanding.

We believe our work provides a stepping stone for writing tools that make good use of computational resources such as lexicons and parsers. Our work is not complete and listing its limitations is a good start for future studies. We hope that future work on perceived and actual difficulty will address these limitations. A limitation of this study is that it only focuses on surface features at the sentence level. It has been shown here that changing this presentation has significant effects on consumers' perceptions but not on their actual understanding. More surface features can be evaluated but future studies should also include the overall structure of the text, e.g., length of paragraph, and the relation between content and text presentation, e.g., how ideas divided over different paragraphs, storyline, and flow of an argument in a text. Another limitation of this study is its use of multiple-choice questions to measure understanding. Although these tests have been used in all school systems for centuries, other tasks may provide more and different insights. For example, Cloze based studies could facilitate comparison with existing studies, other tasks could focus on asking for a verbal explanation to questions, repeating of information, and even discussion and argument formation. Finally, only features that can be automatically detected should be of interested because only those will lead to scalable solutions. Function word density, noun phrases complexity, and overall sentence structure are within reach of today's parser. More advanced evaluations, such as flow of argument, will probably require combinations of rule-based and statistical approaches but are also possible. However, we suspect that with inclusion of more semantic information, metrics will become necessarily domain specific.

We intend to evaluate more text features and their effects on perceived difficulty and consumers' willingness to self-educate. Our future work will also continue to compare perceived difficulty and actual difficulty. We intend to include semantic measures, such as word familiarity, and information density in documents.

Summary points
Already known:

- Readability formulas are most often used to measure text difficulty. Writing guidelines rely on readability formulas.
- Most patient educational materials, online or not, are considered too difficult based on current readability metrics. Implicitly, that work assumed ‘actual difficulty’ to be related to readability formulae outcome.
- Active voice and short words are considered easier to understand.

Knowledge added by this study:

- Actual and perceived difficulty should not be considered identical, especially not since manipulating perceived difficulty does not necessarily lead to simplification (actual difficulty) of the text.
- Grammatical structures influence perceived difficulty.
- With today's resources, fine-grained and content-specific metrics can be constructed for health information.
- Influencing actual difficulty, and so understanding and retention of information, requires more than surface changes to the text. Those surface changes are currently most encouraged by use of readability formulas.

Acknowledgements

The authors would like to thank all participants of the study. This work was made possible by a grant from the National Science Foundation, SGER, 0742223, “U3—Understanding User Understanding”.

Contributions. All three authors contributed to (1) the conception and design of the study, acquisition of data, analysis and interpretation of data, (2) drafting the article and (3) final approval of the version submitted.

Appendix A. Overview of sentence structures

<table>
<thead>
<tr>
<th>Sentence structure</th>
<th>NP complexity</th>
<th>Function word density</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active (foursome)</td>
<td>Simple</td>
<td>High</td>
<td>We will ask those patients who are qualified to participate in this study to consider a research permission form which includes the following information.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>We will ask patients qualified to participate in this study to consider a research permission form which includes the following information.</td>
</tr>
<tr>
<td></td>
<td>Complex</td>
<td>High</td>
<td>We will ask those patients who are eligible to participate in this study to consider a research consent form which includes the following information.</td>
</tr>
</tbody>
</table>
Appendix A (Continued)

<table>
<thead>
<tr>
<th>Sentence structure</th>
<th>NP complexity</th>
<th>Function word density</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passive (foursome)</td>
<td>Simple</td>
<td>High</td>
<td>We will ask patients eligible to participate in this study to consider a research consent form which includes the following information.</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td></td>
<td>Those patients who are qualified to participate in this study will be asked to consider a research permission form which includes the following information.</td>
</tr>
<tr>
<td></td>
<td>Complex</td>
<td>High</td>
<td>Those patients who are eligible to participate in this study will be asked to consider a research consent form which includes the following information.</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td></td>
<td>Patients eligible to participate in this study will be asked to consider a research consent form which includes the following information.</td>
</tr>
<tr>
<td>Sentential (foursome)</td>
<td>Simple</td>
<td>High</td>
<td>Exercising under your doctor's supervision improves your heart functions.</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td></td>
<td>Exercising under doctor supervision improves heart functions.</td>
</tr>
<tr>
<td></td>
<td>Complex</td>
<td>High</td>
<td>Exercising under your physician's supervision improves your heart functions.</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td></td>
<td>Exercising under physician's supervision improves heart functions.</td>
</tr>
<tr>
<td>Extraposed (foursome)</td>
<td>Simple</td>
<td>High</td>
<td>It improves your heart functions to exercise under your doctor's supervision.</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td></td>
<td>It improves heart function to exercise under doctor supervision.</td>
</tr>
<tr>
<td></td>
<td>Complex</td>
<td>High</td>
<td>It improves your heart functions to exercise under your physician's supervision.</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td></td>
<td>It improves heart functions to exercise under physician supervision.</td>
</tr>
</tbody>
</table>

Appendix B. Overview of paragraphs

B.1. Heart Disease Document (difficult version)

Primary care counseling for modifiable coronary artery disease risk factors, including obesity, during preventive health examinations is advocated by the American Heart Association and the American Diabetes Association (ADA, 2001; Grundy et al., 1997). In a recent study conducted in two primary care practices in Louisiana (Huang et al., 2004), primary care practitioner counseling on weight loss was well-received by patients and effective in increasing patients’ understanding of the negative health impact of obesity. However, also identified in this study was the lack of sufficient guidance on weight management strategies for primary care practitioners. Insufficient physician confidence, knowledge and counseling skills, as well as lack of time, resources and under use of dietitians that contribute to inadequate counseling on diet, physical activity, and weight loss were included as potential reasons for this deficiency (Yeager et al., 1996). The extent and content of physician counseling about diet, exercise, and weight loss are inadequate (Galuska et al., 1999; Nawaz et al., 2000). This is discouraging given the fact that physician–patient interactions regarding healthy diet habits have been shown to effect change resulting in improved eating habits (USPSTF, 2002) and weight loss (Nawaz, 2000). Given its potential, that little research has been conducted on primary care interventions for obesity in childhood is surprising. To our knowledge, a study by Saelens et al. (2000) is the only study to date evaluating a primary care-based behavioral therapy program for weight control management in adolescents.

B.2. Heart Disease Document (easy version)

The American Heart Association and the American Diabetes Association advocate that those who have risk factors for
heart disease that are under their control, including being overweight, should receive counseling from their primary caregivers about these factors, when they have their regular checkups (ADA, 2001; Grundy et al., 1997). In a recent study, which researchers conducted in two medical clinics in Louisiana (Huang et al., 2004), patients appreciated counseling by doctors and nurses that encouraged them to lose weight. And it was effective in helping them understand how their being overweight could be causing or result in their own poor health. However, this study also showed that there was not enough guidance given to doctors and nurses about ways to help people manage their weight. Some reasons for this lack of guidance might include not enough confidence on the part of doctors in their own knowledge and skill at counseling, as well as a lack of time and resources on their part. Not using the services of dietician also might have contributed to not enough counseling on food, exercise, and losing weight, according to Yeager et al., 1996. According to studies by Galuska et al. (1999) and Nawaz et al. (2000), the amount of counseling by doctors about diet, exercise, and losing weight, as well as the content of that counseling, was not adequate. It is discouraging that the counseling was inadequate, since studies have shown that conversations between doctors and patients about healthy eating habits can improve the eating habits of patients (USPSTF, 2002) and help them lose weight (Nawaz, 2000). Given the potential for improving the eating habits of patients, it is surprising how few researchers have looked into the effect of counseling by doctors and nurses on overweight children. We know of only one study up to now (by Saelens et al., 2000), that weighs the benefits of counseling by primary caregivers to change the behavior of adolescents in order to help them control their weight.

B.3. Depression Document (difficult version)

On average, patients with major depression are symptomatic 60% of the time, even when community-standard antidepressant treatment is being received. Unfortunately, a complete resolution of symptoms with antidepressant treatment is not experienced by most patients with depression and 10–20% of patients are refractory to all currently available modalities, including electroconvulsive shock (ECT) therapy. ECT is often effective in patients who have failed adequate trials of multiple antidepressants, but is associated with the risk of anesthesia and with significant short-term memory impairment. Responses to ECT are short-lived, and many patients who respond subsequently relapse, even when on maintenance antidepressants. In addition to efficacy issues, side effects associated with antidepressants or ECT cannot be tolerated by many patients. The risks of not responding to (or tolerating) treatment have been highlighted by recent studies documenting that partial (but incomplete) response is associated with an increased risk of full symptomatic relapse (even when on therapy) and a worse long term disease course, as well as with significantly impaired quality of life. Treatment resistance also results in a six times increase in direct health care costs. The tremendous need to identify novel treatment strategies is highlighted by these factors, especially for depressed patients who are unresponsive to conventional therapies.

B.4. Depression Document (easy version)

Overall, patients who have major depression experience symptoms 60% of the time, even if they are receiving treatment by antidepressant drugs that are standard in the medical community. Unfortunately, most patients with depression do not experience a complete recovery from their symptoms with treatment by antidepressant drugs. 10–20% of patients are not helped by any currently available treatment, including electric shock therapy (ECT). Shock therapy is often effective for patients who have not been helped by treatment with antidepressant drugs, but it runs the risk of complications from the anesthesia and they experience significant impairment of their short-term memory. Improvement from shock therapy does not last long, and many patients who do improve, later experience a relapse, even when they continue to take antidepressant drugs. In addition to the problem of treatments not working well, many patients are unable to tolerate the side effects that come with antidepressant drugs or shock therapy. Recent studies highlight the risks to patients if their treatment does not work or if they cannot tolerate it. These studies show that a partial (but incomplete) recovery is associated with a greater possibility of a full return of their symptoms (even when they are on therapy) and a worse long-term outcome of the disease, as well as with a quality of life that is significantly impaired. Resistance to treatment also results in direct health care costs that are six times higher. These factors highlight the tremendous need to find new methods of treatment, especially for depressed patients who are not responding to conventional treatment.

REFERENCES


