Published in final edited form as:
Int J Med Inform. 2013 August ; 82(8): 717–730. doi:10.1016/j.ijmedinf.2013.03.001.

A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty

Gondy Leroy, Ph.D.,
School of Information Systems and Technology, Claremont Graduate University

David Kauchak, Ph.D., and
Computer Science Department, Middlebury College

Obay Mouradi
School of Information Systems and Technology, Claremont Graduate University

Abstract

Purpose—Low patient health literacy has been associated with cost increases in medicine because it contributes to inadequate care. Providing explanatory text is a convenient approach to distribute medical information and increase health literacy. Unfortunately, writing text that is easily understood is challenging. This work tests two text features for their impact on understanding: lexical simplification and coherence enhancement.

Methods—A user study was conducted to test the features' effect on perceived and actual text difficulty. Individual sentences were used to test perceived difficulty. Using a 5-point Likert scale, participants compared eight pairs of original and simplified sentences. Abstracts were used to test actual difficulty. For each abstract, four versions were created: original, lexically simplified, coherence enhanced, and lexically simplified and coherence enhanced. Using a mixed design, one group of participants worked with the original and lexically simplified documents (no coherence enhancement) while a second group worked with the coherence enhanced versions. Actual difficulty was measured using a Cloze measure and multiple-choice questions.

Results—Using Amazon's Mechanical Turk, 200 people participated of which 187 qualified based on our data qualification tests. A paired-samples *t*-test for the sentence ratings showed a significant reduction in difficulty after lexical simplification ($p < .001$). Results for actual difficulty are based on the abstracts and associated tasks. A two-way ANOVA for the Cloze test showed no effect of coherence enhancement but a main effect for lexical simplification, with the simplification leading to worse scores ($p = .004$). A follow-up ANOVA showed this effect exists only for function words when coherence was not enhanced ($p = .008$). In contrast, a two-way ANOVA for answering multiple-choice questions showed a significant beneficial effect of coherence enhancement ($p = .003$) but no effect of lexical simplification.

Corresponding Author: Gondy Leroy, School of Information Systems and Technology, Building ACB, Claremont Graduate University, 130 E. Ninth Street, Claremont, CA 91711, Phone: 909-6007-3270, gondy.leroy@cgu.edu.

Authors Contributions
All three authors contributed equally to the design of algorithms and executing and analysis of the study.

Conflict of Interest Statement
None of the authors has a conflict of interest.

Conclusions—Lexical simplification reduced the perceived difficulty of texts. Coherence enhancement reduced the actual difficulty of text when measured using multiple-choice questions. However, the Cloze measure results showed that lexical simplification can negatively impact the flow of the text.

Keywords

Comprehension; Readability; Health Literacy; Consumer Health Information; Patient Education; Information Systems; Medical Informatics Computing

Introduction

With increasing availability of medical web sites on the Internet, millions of patients and caregivers now look online for information. This trend is growing with prolonged lifespans, a rise in the number of diseases and problems that require long-term care, and increasingly sophisticated procedures and therapies. Unfortunately, the medical information presented to readers is not always clear. Weis [1] summarizes several studies demonstrating the increased burden and cost associated with limited comprehension. In general, there are millions of people in the US without sufficient health literacy to understand their treatments or preventive care [2]. Many examples of how low health literacy has had an impact on specific groups have also been noted including changes in participation in cervical cancer screening [3] and glycemic control in type 2 diabetes patients [4]. Moreover, readability levels are sometimes too high for healthcare providers, e.g., Singh [5] reports that behavioral treatment plans in psychiatric hospitals were not optimally carried out because they were not sufficiently understood by the staff.

Earlier work focused on evaluating the quality of the websites and the information they contained. Bernstam et al. reviewed twenty-two different such website criteria [6] and 273 different evaluation instruments [7]. They found that the inter-rater agreement between experts was low for several of the criteria but the agreement could be improved with very precise definitions. Of the instruments evaluated, they concluded that few were practical or useful for judging quality. Others have focused on automating the analysis of the sites, for example, Wang and Liu [8] developed instruments to automatically evaluate web pages using eighteen criteria.

Being able to rely on understandable information is a crucial aspect to increase health literacy. While there are many possible approaches to improving the provided information, for example, Choi and Bakken [9] add visuals, such as drawings, few have been demonstrated to have a clear beneficial impact. Among all available tools, ranging from interactive and personalized education to the use of demonstrative video, we believe that text can be an effective and cost efficient method to bring information to patients. Informative, preventive, persuasive or explanatory text is accepted in society and easily accessible for readers in either paper or electronic format.

Perceived and Actual Text Difficulty

Learning from text can be difficult. For a text to be effective at conveying information, people need to be both willing to read it (perceived difficulty) and capable of understanding its contents (actual difficulty). We argue that these two requirements impact the usefulness of a document and are important in studying health literacy and must both be accounted for when simplifying text. The actual difficulty of a text will influence how well readers understand the information. However, perceived difficulty cannot be ignored since it may prevent a person from even reading a document.

We base this distinction on previous work in the context of the Health Belief Model (HBM) and the Theory of Planned Behavior (TPB). In a review of twenty-four studies, Janz and Becker [10] looked at the four dimensions of the HBM model: perceived susceptibility, perceived severity, perceived benefits, and perceived barriers. They found the last dimension, *perceived barriers,* to be the most significant in explaining health behavior. Others have also used the HBM model as a guideline. For example, Davis et al. [11] used it as a model to design their study evaluating factors influencing colorectal cancer screening. The TPB, an extension of the Theory of Reasoned Action, also provides supporting evidence for the distinction between perceived and actual difficulty. In particular, the model includes a factor that encompasses beliefs: *perceived behavioral control*. This factorhas been shown to contain two distinct components, *perceived difficulty* and *perceived control,* which can bemanipulated independently, with perceived difficulty being the stronger predictor of intentions and behavior [33].

In current work, perceived and actual difficulty are seldom distinguished. The lack of strong evidence supporting an increase in understanding using simplification methods that rely on readability formulas may indicate that such methods affect perceived difficulty more than actual difficulty. These methods may generate text that *looks* easier but may not necessarily *be* easier to understand. In previous work, we also found indirect evidence for this distinction, since it is often much easier to improve the perceived difficulty of a document than the actual difficulty [12]. Simplifying text so that readers increase their understanding and recall of the information is more challenging.

Factors Influencing Learning and Understanding

How much is understood and learned from a text depends on many different factors. This learning process is commonly seen as an interaction between two types of factors: text characteristics and reader characteristics [13]. However, a third factor that can influence results is how understanding and learning are measured. Each of these three factors is composed of distinct, but not necessarily independent characteristics, see Figure 1. For example, the length, topic and writing style are text characteristics that affect text difficulty; age, education and language skills are example reader characteristics that affect comprehension; and open-ended questions or teach-back methods are different measures which have been shown to differentiate between aspects of understanding and learning.

Text Characteristics

Over the last few decades, many studies have focused on simplifying text by changing words (lexical) and shortening sentences (grammatical) following the assumption that this leads to increased understanding. The most popular approach has been the application of readability formulas and it is frequently suggested that following these formulas will lead to easier to understand text [14]. There exist several formulas, such as the Flesch-Kincaid readability grade level or the SMOG measure [15], but most operate on the same simple principles and use syllable and word counts in a sentence to assign a difficulty rating. The formulas have been used to test a variety of texts including patient education materials [16], injury prevention or hospice bereavement materials [17], HPV educational materials [18], informed consent forms [19] and even survey instruments [20]. Most writing guidelines, for example by the American Medical Association, advise writing at a 6th to 8th grade level. By applying these formulas, it has been concluded that most English-language Internet sites are too difficult because they require on average a 10^{th} grade level and sometimes even college level education [7–9] regardless of which of the five different readability formulas are used [14, 15].

Unfortunately, few studies demonstrate an increase in understanding after making text modifications guided by such formulas. Friedman et al. [21], for example, did not find a correlation between Flesch-Kincaid readability scores and a Cloze measure of understanding. In addition, Wubben et al. [22] identify a number of issues for using the Flesh-Kincaid for directing and evaluation simplification algorithms. At a minimum, a re-validation of the formulas with modern text covering healthcare-relevant topics is needed. For example, many formulas equate *long* words with *difficult* words. However, in medicine, this relationship may not hold true, e.g., “apnea” would be considered more difficult than “diabetes” or “obesity” by most readers.

Over the years and across different disciplines, evidence has emerged that specific text characteristics affect text difficulty. This work focuses on two such characteristics: term difficulty and text coherence. Term difficulty can be addressed by lexical simplification [23, 24] which focuses on substituting difficult terms with simpler ones or by elaborating on a difficult term. Kandula et al. [25] found an increase in Cloze scores after manual lexical simplification of text from journals and electronic health records. O’Donnell [26] shows the beneficial effects of lexical simplification using elaborations for difficult Spanish terms for English speakers. Students who read elaborated documents could recall more information and the effect was stronger for more difficult text.

The second text characteristic we focus on is text coherence. McNamara et al. [27] define coherence as comprising structural and explanatory coherence. Coherence is improved by ensuring that no gaps exist in the flow of a document by use of anaphoric referents, connective ties, synonyms, etc. As part of the Readability Assessment Instrument (RAIN), both local and global coherence are defined by 12 characteristics, such as the inclusion of cause/effect relations, temporal relations, etc. Variants of these definitions exist [13], but all emphasize the need for good flow in a document combined with a structured, logical argument. Tools that improve the coherence of a text are not as popular as those based on readability formulas and very few exist. In early work, Adkins et al. [16, 28] and Kirkpatrick and Moheler [29] estimated the readability of fact sheets using a manual approach based on the RAIN characteristics. Norvig [30] developed an algorithm to estimate coherence using mappings to knowledge bases instead of literal text overlap.

The evidence for the importance of coherence is subtle and its effects often depend on other characteristics. Looking at overall coherence, Boscolo and Mason [13] found it did not affect their recall tasks but affected question answering for those readers who had high knowledge of but low interest ina topic, and vice versa. In contrast, Goldman and Murray [31] evaluated logical connectors as words that increase text coherence. Using an adjusted Cloze procedure they found that some connectors, additive and causal connectors were easier for readers to fill in than others such as adversative or sequential connectors.

Reader Characteristics

Many characteristics associated with the readers have been found to be associated with increased understanding. Van Servellen et al. [32] found that the level of education influences the ability to distinguish myths from facts about HIV and to understand HIV-related terms. Specific knowledge of a topic affects reading and learning, but the outcome is also influenced by how the information is presented. For example, Potelle and Rouet [33] found that readers with low prior knowledge of a topic scored higher on a survey and recalled more information after reading a hierarchical map as compared to a list or concept map. However, they found no differences for readers with high prior knowledge. Motivation and interest have also been shown to be important [13], with higher interest in a topic leading to better performance based on a number of measures of learning.

A reader's health literacy plays an important role and is typically measured with an instrument such as the Rapid Estimate of Adult Literacy in Medicine (REALM) [34], a word recognition test, or the Test of Functional Health Literacy in Adults (TOFHLA) and its shortened version (S-TOFHLA)[1]. Studies evaluating the effects of health literacy on understanding show that low health literacy often leads to a poor understanding of information. McWhirter [35] found that S-TOFHLA scores correlated with Cloze scores but not with a teach-back method for their texts on colon cancer. Gazmararian et al. [36] found that lower health literacy may play a role in prescription drug refill adherence.

Stress is another reader characteristics ordinarily measured with a survey, for example, with the Perceived Stress Scale-10 (PSS-10) [37]. Increased stress has been shown to be related to lower comprehension of medical terminology, such as HIV terms [32]. In earlier work [38], we also found an effect of both health literacy and stress; vulnerable subjects, those with high stress or with low health literacy, relied much more heavily on a visual table of contents that accompanied a text to answer questions.

Several other reader characteristics have been shown to play a role in understanding and learning. Goldman and Murray [31] found that English-as-a-Second Language (ESL) study participants performed worse on a Cloze test than native speakers, but the pattern of errors was the same for both groups. Other reader characteristics and their effects are less well understood. Van Servellen et al. [32] found a relationship between understanding HIV terms and the participants' education, income and the quality of communication received by healthcare providers. In addition, age and level of education were found to relate to understanding prescriptions.

Measurement Characteristics

The method of measuring understanding and learning, which is seldom discussed explicitly as a factor, will also affect the results. To measure perceived difficulty, it suffices to judge the difficulty of a document using a simple scale. However, to measure actual text difficulty and its relation to understanding and learning, a variety of tests have been used.

A common approach to evaluate understanding and learning is to pose questions during or after reading a text. This method has been adopted for centuries by educational institutes to conduct examinations. The questions can be factual questions or can require more detailed understanding including the ability to reason about the topic, or the ability integrate the information with one's personal knowledge and apply it. The type of question is important and can impact results. McNamara [27] found that high knowledge readers performed better on problem solving questions but not as well on inference questions with less coherent text. Skilled and unskilled readers also benefited differently depending on the coherence of the text. Skilled readers showed higher recall with poorly constructed text and suggest that it is due to the need for more active processing.

The Cloze measure is a fill-in-the-blanks approach introduced and validated by Taylor [39]. It was intended as a measure to distinguish between texts with different readability levels. With this method, every n^{th} word of a text is deleted and readers are asked to fill in the blanks. Taylor found that the measure differentiates text difficulty better than readability formulas. Although not originally intended as such, the Cloze measure has also been used for measuring understanding, though the results have been mixed. Kandula et al. [25] are one of the few groups to report an increase in understanding using the Cloze measure after

[1]Offered for sale at http://www.peppercornbooks.com

text simplification based on lexical term simplification and splitting of longer sentences into shorter ones.

We discuss the three categories of factors above as if they influence understanding independently of each other. This is an oversimplification and we expect to encounter many interactions between the three. Ideally, however, we can pinpoint some characteristics that are universal indicators of simplicity and lead to approaches that increase understanding after reading a simplified text.

Semi-automated Text Simplification

As pointed out by Seligman et al. [40], creating suitable and effective text manually requires both time and expertise. Unfortunately, few clinicians can spend the amount of time required to create text that is suitable for each different type of patient and hiring professional writers to simplify all texts is prohibitively expensive. Instead, our goal is to design and develop a writer support tool that can help writers simplify text.

We impose two constraints on our research and the algorithm development. The first is that any feature and its associated text changes used by our final algorithm must be backed by evidence showing a positive effect on reducing text difficulty. Such evidence is collected based on user studies that show improved understanding or retention. In practice, this means that we measure the understanding of text using a number of different metrics. In conducting these studies, we explicitly work with representative consumers and do not approximate difficulty levels via expert judgments or existing readability formulas.

The second constraint is that the final algorithm must be programmed and integrated into a writing support tool, e.g., as a plugin for Microsoft Word or as a standalone editor. This integration should allow for efficient and effective support for the writer without requiring specialized linguistics knowledge or writing skills. In practice, this means that the algorithm should be able to pinpoint overly difficult sections in a text without human intervention and suggest candidate alterations to the writer. The writer can then change the original text based on the suggestions.

In this study, we evaluated lexical simplification and coherence enhancement. Each approach is applied in a systematical manner to our text so that, if successful, algorithms can be programmed to automate the process.

Lexical Simplification

Manual lexical simplification has been shown to be helpful in increasing Cloze scores [25] and automatic lexical simplification techniques have been shown to improve readability formulas [22]. Our goal is to automate the lexical simplification process as much as possible. For our approach lexical simplification is accomplished in three phases: 1) identify difficult words, 2) generate candidate substitutions, and 3) choose the best substitution. The first two stages are completely automated and only the third phase of the algorithm requires any manual intervention.

The first phase consists of identifying difficult words. Text is split into individual sentences which are then parsed to get the parts-of-speech (POS) for each word. We use the Stanford Parser[2] [41] for this task. Then, the frequency of each word is retrieved from the Google Web Corpus[3], which contains *n*-gram counts from a corpus of a trillion words from public

[2] http://nlp.stanford.edu/software/lex-parser.shtml
[3] Available from the Linguistic Data Consortium, www.ldc.upenn.edu

web pages. We use this web frequency as an estimate of how familiar words are to readers. The rationale is that words that occur more on the web are more commonly written and read and can therefore be expected to be better known. Evidence for this approach comes from our corpus analysis studies where we found that there was a significantly higher average frequency of words in easy documents than in difficult documents [42, 43]. Words with lower frequency, i.e. less familiar, are flagged as difficult and become candidates to be changed.

Given the set of candidate difficult words in a text, the second phase of the algorithm identifies candidate improvements. Specifically, we identify synonyms, hypernyms and word definitions for the difficult words from four different resources: WordNet, Simple and Normal Wiktionary and the Unified Medical Language Systems (UMLS). Only synonyms and hypernyms that have the same POS as the original word are retrieved. If any new word has a Google *n*-gram frequency that is higher than the original word, i.e., it is more familiar, it is retained. Definitions and partial definitions can be used if all words with frequencies lower than the difficult word are replaced by easier words. These options are then ordered based on frequency with higher frequency words being preferred over lower.

Finally, in the last phase of the algorithm, the writer chooses the best replacement for each difficult word from the set of options generated in the second phase. When synonyms are the best alternative, a writer can simply replace the original word with the synonym. Hypernyms can be inserted in the sentence either between two commas or in parentheses. Longer phrases such as definitions can be added as stand-alone sentences before or after the sentence containing the difficult word. For example, a sentence such as:

> "Patients receiving HIV-related home care services provide an opportunity for assessment of oral health and smoking cessation needs; however, the majority of home care providers lack formal training to provide these services"

is rephrased to be:

> "Patients *getting* services *for* HIV-related home care provide an opportunity for assessment of oral health and the *need for quitting* smoking; however, *most* home care providers *do not have* formal training to provide these services.

Coherence Enhancement

Text coherence is an important complementary component since lexical simplification can reduce the flow of a document. This can occur when new information is added and may result in the text becoming choppy' to read, a characteristic resulting in more difficult rather than simpler text [44]. Coherence is therefore a logical feature to evaluate in combination with lexical simplification. We use the term coherence to represent both local (from one sentence to the next) and global (covering the entire text) coherence. Text coherence is improved by the application of a set of rules as outlined below. Currently the detection of potential incoherent sections and the application of these rules is done by the writer, however, in the future we plan to automate one or both steps.

First, local coherence is improved by adding pronouns to connect sentences more explicitly. For example:

> "The most common causes of morbidity and mortality in the western world can be accounted for by unhealthy patterns of behavior … . Interventions to improve health behavior are sorely needed."

can be changed to:

> "The most common causes of morbidity and mortality in the western world can be accounted for by unhealthy patterns of …. Interventions to improve ***such*** health behavior are sorely needed."

When there is repetition of a noun or noun phrase, often the result of lexical elaboration, it is replaced by an appropriate pronoun. This adjustment is based on work by Ledoux et al. [45], who show how the repetition of a prominent antecedent in a coreferential relation hinders reading and forms a disjoint reference. For example, the repetition of aspirin in the following sentence demonstrates this problem:

> "Taking aspirin will reduce the symptoms. The aspirin will help reduce the headache."

which after improving coherence becomes:

> "Taking aspirin will reduce the symptoms. ***It*** will help reduce the headache."

Second, global coherence is improved by adjusting the spacing of the text based on Gestalt principles [46]. Applied to text, these Gestalt laws dictate one idea per chunk of text. Spacing is then used to demarcate different chunks of text and to show how ideas relate to each other. When the ideas are disjoint, this involves creating new paragraphs. When the ideas are related, enumerations are created with each idea indented and on its own line. This type of organization, referred to as perceptual organization, has been shown to strongly correlate with scores on the SATA reading comprehension test [47]. For example, following these principles, the text:

> "…. behavior is required. In this commentary, we explore three relatively new possible roads ….: (1) genetics may influence …., (2) genetics may tone down …., and (3) genetics may be ….."

becomes the following after improving coherence:

> "…. behavior is required.
>
> In this commentary, we explore three relatively new possible roads ….:
>
> 1. genetics may influence ….,
> 2. genetics may tone down …., and
> 3. genetics may be ….."

User Study Design

We evaluated the impact of lexical simplification and coherence enhancement on text difficulty. To measure perceived difficulty, we selected individual sentences so that we could include several examples and several types of sentences. To measure actual difficulty, we selected stand-alone abstracts about which we could ask questions. To verify our setup and avoid software difficulty during the study, we first conducted a small pilot study with 17 participants using only lexical simplification [12]. The study presented here does not include this pilot study data, is based on different subjects and adds coherence enhancement as a variable.

Participants

We invited 200 participants to complete the study. They were recruited using Amazon's Mechanical Turk (MTurk) and were paid $1 for their participation. MTurk is an online service that facilitates human workers performing online tasks ranging from annotations to translation to user studies. It attracts a large number of participants from demographically diverse backgrounds [48] and data collected through MTurk has been shown to be as

reliable as traditional methods for both user studies [48, 49] and for a range of natural language tasks [50].

Although results from studies with MTurk have been reported as reliable, we included a number of validation tests to ensure the quality of our results and to avoid participants using programs to answer automatically or participants who do not take the study seriously. Data from participants who do not pass all validation tests are excluded. To test for quality, two multiple-choice validation questions were added in with the other multiple choice questions, one with each document. Each of these questions was extremely easy and could be answered correctly if the question was read. As an additional quality control, we measured the time taken to complete each section. If a section was completed in less than a minute, the data was discarded. Finally, to avoid multiple submissions from the same participant, IP addresses were temporarily recorded and results from submissions with the same IP address were discarded. Although there are situations where IP addresses can be shared, e.g., in public settings such as libraries, this is rare.

Stimuli

In April of 2012, we conducted an online search using PubMed for “smoking cessation”. We selected eight sentences from the results of this search. To obtain a representative sample, sentences were selected with varied length and varied Flesch-Kincaid scores. Flesch-Kincaid was measured for comparability with other research. We used Microsoft Word’s grammar feature to calculate the readability score, comparable to other researchers [7, 51–53]. Each sentence was simplified using our lexical simplification approach. Given the short length of individual sentences, we did not evaluate the effects of coherence enhancement for the sentences.

In addition to the sentences, we selected the first two abstracts of our PubMed search that were not descriptions of experiments. With this approach, we mimic online consumers who are looking for in-depth information but do not wish to read the results of individual experiments. One document discussed the influence of genetics on health behaviors (PMID = 22488456), which we refer to as the “genetics document”. The other document discussed the need for action in underdeveloped countries to improve smoking cessation (PMID = 22487605), which we refer to as the “countries document”. For each document, we created four versions to evaluate our independent variables:

1. no lexical simplification or coherence enhancement (original text)
2. lexical simplified but no coherence enhancement
3. coherence enhancement but no lexical simplification
4. lexical simplified and coherence enhancement

Table I provides an overview of the sentence and document characteristics. In the sentences, 54 words were considered too difficult because they had a Google frequency below the threshold. During simplification, 35 were replaced with an easier alternative by the writer (65%). Similarly, in the countries document 26 of 47 difficult words were replaced (55%) and in the genetics document 28 of 42 words (67%). The sentence and documents used in the study are available in the appendix [INSERT LINK TO APPENDIX].

Independent Variables

We evaluate the effect of two independent variables: lexical simplification (IV1) and coherence enhancement (IV2).

Because we use individual sentences to measure perceived difficulty, in which coherence cannot be enhanced, only the effect of lexical simplification (IV1) is measured for perceived difficulty. Although this prevents the analysis of lexical coherence, it allows subjects to finish the study in a reasonable time while still reading all the text. The sentences were shown to participants as pairs, with each pair containing the original and simplified sentence. Each participant viewed each pair in a randomized order; furthermore, the eight pairs were shown in a random order for each participant. We presented sentences in pairs since previous work has shown that a random set of 16 sentences offers less comparative value for the participants and would result in less useful judgments [54, 55].

For the documents, we measure actual difficulty and both independent variables (IV1 and IV2) are applied resulting in four versions of each document. Ideally, a completely within-subjects design would be used where each participant reads all four versions. However, this would result in an overly long study and can be problematic for testing actual difficulty (i.e. understanding and knowledge acquisition) because of carryover effects. Instead, we used a mixed design. One group of participants (Group 1) received documents that were not enhanced for coherence; the other group (Group 2) worked with the versions that were enhanced for coherence. Each group received a document that was lexically simplified and one that was not.

Since each participant worked with two texts it was necessary to use two different topics to avoid learning effects. The order of the conditions and the topic of the document were counter-balanced to avoid order bias. For example, the first group of participants was assigned to one the following orderings:

- Lexically Simplified Countries document – followed by – Original Genetics document
- Original Genetics document – followed – by Lexically Simplified Countries document
- Lexically Simplified Genetics document – followed by – Original Countries document
- Original Countries document – followed by – Lexically simplified Genetics document

The same orderings were also used with the Enhanced Coherence versions of each document and assigned to the second group of participants.

Dependent Variables

Perceived difficulty was measured with a subjective measurement using a 5-point Likert scale with labels ranging from “Very Easy” (score 1) to “Very Hard” (score 5). Participants were asked to judge each sentence on this scale. The instructions given in this section were “*Please rate each sentence for difficulty. Imagine that you are a patient. How easy is each sentence in helping you understand what is going on?*”

Actual difficulty was measured with the Cloze measure and with multiple-choice questions. The Cloze procedure was applied to both texts the participants received. Every seventh word was removed and left blank to be filled in by the participant. Table I provides an overview of the number of words deleted, including the distinction between content words, e.g. nouns, versus function words, e.g. conjunctions. Participants were asked to write a single word for each blank. In a later section of the study, participants were presented with the same texts (this time without removing words) and were asked to answer multiple-choice questions about the content of the text. For each text, we created seven multiple-choice questions, of

which six were content related and one was a validation question. The questions and answers were the same for all versions of a text. To prevent bias towards the original or simplified versions we avoided phrasings found in either text when generating the questions and answers.

Other Variables

We also collected demographic information for each participant. We included a section on general demographic information (race, ethnicity, native language, and age), a section on reading habits (frequency of reading books, magazines, online reading) and a section with the Short Test of Functional Health Literacy in Adults (S-TOFHLA).

Procedure

There were six sections in the study. Each participant worked through all sections in the following order:

1. Welcome page with instructions to finish each section sequentially (back button was disabled).
2. Actual difficulty measurement of the two documents using the Cloze test
3. Perceived difficulty measure of the eight sentences using a Likert-scale
4. Reader characteristics measurement:
    a. Demographic questions about age, gender, race, ethnicity, languages spoken, education level, linguistics or medical knowledge
    b. Questions about reading habits, i.e., how often they read books, printed news and magazine, and text online
    c. S-TOFHLA
5. Actual difficulty measurement of the same two documents used in section 2 using multiple-choice questions
6. Thank you page with contact information

Results

Participants

The data of 187 participants was accepted for this study. Other data (8 participants) was rejected because the participants failed one or more of the validation questions or because they used the same IP address (5 participants). On average, participants needed 33 minutes and 14 seconds to complete the study.

Table II shows an overview of the demographic information. A chi-square analysis was conducted for each demographic variable. The test showed no significant difference between the two groups of participants for any of the demographic variables and we therefore discuss the data as a single group. Adults of different ages were well represented in the group. The largest group was from participants who were between 21 and 30 years old (44%), and the smallest groups were those younger than 20 years (7%) and between 61 and 70 years old (5%). There were more female (64%) than male (36%) participants. Race and ethnicity also varied. The largest racial group was white (76%) and largest ethnic group Not Hispanic or Latino (92%). Because Internet access is global, we did not limit our participants to North America. Most users were North-American though, (87%) with a smaller group from Asia (11%) and very few from other continents such as Africa (0.5%) and Europe (2%).

Beyond demographics, we also collected participant characteristics that might impact text understanding and to explore how the amount of simplification interacted with perceived and actual text difficulty. We asked the participants about their education and language skills (Table II) and their reading habits for a variety of sources (Table III). Slightly more than half of the participants did not possess a university degree. There were 2% without a high school diploma, 36% with a high school diploma and 16% with an associate's degree. Of those with a university degree, most had obtained a bachelor's degree (31%) with fewer having obtained a master's degree (12%) or doctorate (3%). The majority of our participants spoke English at home exclusively (79%) or most of the time (7%). Very few spoke no English at home (1%), rarely English (5%) or half of the time English at home (8%).

We asked each participant about reading habits related to books, printed media and online media. For each category, we included examples with our description, such as "Time, The Economist and Newsweek" for printed news magazines, "email and newsletters" as examples of Internet communication, and "coupons, tickers and RSS feeds" as examples of Internet notifications. More than half of the participants read two or more fiction books per year with more than a third reading non-fiction books and textbooks. Most participants did not read printed magazines, with more than three quarters reporting they did not even read one printed magazine per week. In contrast, about half of our participants reported they read news online almost daily and almost everyone read email and other communications daily.

Perceived Difficulty

All participants evaluated the eight sentence pairs (original and simplified versions) on the 5-point Likert scale with 1 representing the easiest and 5 the most difficult text. Figure 2 (left side) shows that simplified sentences were consistently seen as easier. A paired samples *t*-test indicates that the overall average score for the lexically simplified sentences (2.16) is significantly lower/easier ($p < .001$) than for the original sentences (2.79). This simplification effect is significant for most sentences; taking a Bonferroni correction into account (for conducting 8 tests), the simplified versions of sentences 1, 2, 3, 5, 6 and 8 are significantly easier than the original versions ($p < .001$). However, the difference was not significant for sentence 4 ($p = .052$) or sentence 7 ($p = .140$).

For comparability with other research, we measured the Flesch-Kincaid grade level for each sentence (Figure 2, right side). The average Flesch-Kincaid grade level was 17.3 for the original sentences and 17.1 for the simplified sentences, a difference that is not significant (paired-samples *t*-test, $p = .839$). In some cases, the simplified sentence received a higher Flesch-Kincaid grade level than the original version, which was a result of longer sentences being created by adding information (e.g. definitions or hypernyms). For example, sentence 3 demonstrates this:

> Original: *"In practice, this risk of congenital heart defects is yet another reason to avoid using bupropion and to monitor the cardiac status of exposed fetuses."*
>
> Simplified: *"In practice, this risk of congenital heart defects (which are defects that happen while a baby is developing in a mother's body) is yet another reason to avoid using bupropion (a drug to treat depression) and to monitor the heart status of exposed unborn children."*

These results show the challenges with Flesch-Kincaid Grade Level for directing text simplification. Many of the sentences do not show major differences based on the level, however the responses by our 187 participants show a clear distinction between the original and simplified sentences, with simplified sentenced perceived as easier.

Actual Difficulty

To evaluate the actual difficulty of text, we use two different measures: the Cloze measure and multiple-choice questions. Figure 3 shows the results for the Cloze measure as the percentage of blanks that were filled in correctly. Only exact matches were accepted as correct, ignoring capitalization.

We conducted a two-way ANOVA with lexical simplification and coherence enhancement as the two independent variables and repeated it three times: for the percentage of correctly filled in blanks (All Words), for the blanks representing content words (Content Words), i.e., the meaning bearing words such as nouns and verbs, and for the blanks representing function words (Function Words), i.e., the prepositions, conjunctions etc.

Overall, for all types of words, the participants performed better with text that was not lexically simplified. We found a main effect for simplification ($F(1,370) = 8.575$, $p = .004$) with participants performing better with the original text. The interaction is not significant ($p = .07$). A closer look at the data reveals that this may be due to performance with function words. We found no significant effect of our independent variables on the percentage of content words filled in correctly. However, our ANOVA for function words showed a significant interaction effect ($F(1, 370) = 7.187$, $p = .008$), while the main effect for simplification is not significant ($p = .054$). The document that was lexically simplified but without coherence enhancement led to worse Cloze scores.

Figure 4 shows the percentage of correctly answered questions. There was no significant difference between the original and simplified text for our lexical simplification variable, however, there is a significant difference for coherence enhancement. We conducted a two-way ANOVA and found a main effect of coherence enhancement ($F(1,370)= 8.673$, $p = .003$) with more correct answers with enhanced coherence for documents with lexical simplification (60% versus 53% correct) and without (62% versus 56% correct).

The results show a clear difference in the results depending on the measure used to evaluate actual understanding. The Cloze measure shows no beneficial effects of our treatment of the text, while the multiple-choice questions shows more correct answers with coherence enhanced texts.

User Characteristics

To understand how user characteristics are related to understanding we conducted two-tailed Pearson correlations between user characteristics and the perceived and actual difficulty measures. Table IV shows an overview of the results. For perceived difficulty, we found that only the S-TOFHLA scores correlated with perceived difficulty. A higher S-TOFHLA score (higher health literacy) was associated with lower (easier) scores for perceived difficulty ($r = −.159$, $p = .030$). For the measurement of actual difficulty, we found two significant correlations. The S-TOFHLA scores were negatively correlated with the Cloze measure ($r = −.232$, $p = .001$). Higher S-TOFHLA scores (higher health literacy) were related to lower (worse) scores on the Cloze measure. The second significant correlation was for book reading: people who read more books scored higher (better) on the Cloze measure ($r = .172$, $p = .018$).

We believe these correlations scores are a testament to the importance of the type of measure used. The Cloze measure scores are likely reflecting the lack of coherence of the documents due to the elaboration occurring during lexical simplification. This may have been an important enough effect to show in these correlations as a negative relation between higher literacy and lower scores on the Cloze measure.

Discussion

Today, patients and consumers are increasingly responsible for educating themselves on diseases, treatments and therapies. With the overall popularity of the Internet, text has become the main tool for this self-education. Unfortunately, not all readers understand the information in the text they read. The texts and their topics are often difficult and the reader would benefit from text written in an easier form. Readability formulas have been popular in assisting this effort, however, they only provide a high-level evaluation of text, do not pinpoint difficult sections or suggest alternative writings, and there is little evidence that their outcome is associated with actual understanding.

We believe text difficulty and the associated reader understanding and learning is influenced by at least three factors: reader characteristics, text characteristics and measurement characteristics. In the study presented here, we influenced the lexical difficulty and the coherence of text. We measured both perceived and actual difficulty of the results using a Likert-scale for perceived difficulty, and Cloze measure and multiple-choice questions for actual difficulty. We found that perceived difficulty was improved by applying lexical simplification, but actual difficulty was not. We did not measure perceived difficulty for our coherence modifications (due to practical limitations) but found that increasing coherence positively influenced actual difficulty with increased understanding shown by the multiple-choice question task.

The perceived difficulty of our sentences was reduced after lexical simplification, our first tested text characteristic. Interestingly, we found significant differences between the original and simplified sentences when relying on reader judgments even though applying the Flesch-Kincaid readability formulas did not show a difference in the sentences. In addition, the reader characteristics were not strongly related to perceptions of difficulty. Only one characteristic, the health literacy as measured by the S-TOFHLA, was significantly related to perceived difficulty with people with higher health literacy scoring sentences as easier.

Reducing the actual difficulty of a text in an efficient manner is a difficult task. Our lexical simplification approach had a negative effect on one measurement, the Cloze measure, and no impact on the multiple-choice question scores. Although this was unexpected, the detailed analysis showed that this effect was mainly due to worse scores for function words but not for content words. It is probable that the simplification process introduced awkwardness in the texts that made the task of filling in function words more difficult. This may be the result of decreased coherence when elaboration was used to add information for difficult terms, which in turn led to poorer scores on the Cloze measure and a lack of impact on the multiple-choice questions.

Actual difficulty was improved with coherence enhancement with higher scores on the multiple-choice questions than without coherence enhancement. The size of this improvement, almost ten percent, is somewhat surprising since the coherence enhancement did not change the content. Considering the importance of online information, this is a very encouraging result. First, increasing coherence in this manner does not require a medical background, which makes the process much more accessible. Second, the text provided on the Internet lends itself well to optimal use of spacing and grouping of items.

Limitations and Future Work

The proposed algorithms and the study to evaluate them have some limitations which we hope to address in the future. With regards to our algorithms, they require more fine-tuning before they can be completely integrated in a writing support tool. For example, we plan to improve our lexical simplification algorithm by using a phrase-based instead of word-based

approach for flagging difficult terms. While it may complicate flagging difficult sections, i.e., phrases may be rare but the individual words common, it may help provide more appropriate alternative wordings. We also plan to investigate how to better integrate term elaborations in the original text so that they are less disruptive, such as hints or links that lead to explanations shown separate from the main text. We also plan to fine-tune and automate our coherence algorithm, measure the effects on perceived difficulty and replicate the effect on actual difficulty. There are also several additional limitations to our study setup that we hope to address in the future. Although there were a number of situations where we did not find statistical significant differences in this study, it may be possible that the effects are small and that more participants are required for their discovery. To get a broader view, however, we decided to first improve our algorithm and then conduct larger studies in the future, including different document types. We also aim to test different user characteristics, text characteristics and measurement characteristics to complement this study. Finally, studies based on participation by MTurk participants should also be complemented by work with patients in clinics. Given the importance of interest and motivation, it is possible that the results would be even stronger with patients than with MTurk users, especially when using text and topics relevant to the patients.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would to thank their study participants and Dr. Jennifer Unger, expert in smoking and public health at the University of Southern California (USC) for advise on the study and the questions posed. The study was reviewed by the Institutional Review Board (IRB) of Claremont Graduate University.

This work was supported by the U.S. National Library of Medicine, NIH/NLM 1R03LM010902-01.

References

1. Weis, BD. Health Literacy and Patient Safety: Help Patients Understand. Manual for Clinicians. 2. Vol. 2007. AMA and AMA Foundation; 2007.
2. Committee on Health Literacy - Institute of Medicine of the National Academies. Book Health Literacy: A Prescription to End Confusion. The National Academies Press; 2004. Health Literacy: A Prescription to End Confusion.
3. Garbers S, Chiasson MA. Inadequate Functional Health Literacy in Spanish as a Barrier to Cervical Cancer Screening Among Immigrant Latinas in New York City. Preventing Chronic Disease. 2004; 1(4):1–10.
4. Powell CK, Hill EG, Clancy DE. The Relationship Between Health Literacy and Diabetes Knowledge and Readiness to Take Health Actions . The Diabetes Educator. 2007; 33(1):144–151. [PubMed: 17272800]
5. Singh AN. Readability of Behavioral Treatment Programs in Mental Health. Journal of Child and Family Studies. 1999; 8(4):369–375.
6. Bernstam E, Sagaram S, Walji M, Johnson C, Meric-Bernstam F. Usability of Quality Measures for Online Health Information: Can Commonly Used Technical Quality Criteria be Reliably Assessed? Int J Med Inform. 2005; 74(7–8):675–683. [PubMed: 16043090]
7. Bernstam EV, Shelton DM, Walji M, Meric-Bernstam F. Instruments to Assess the Quality of Health Information on the World Wide Web: What Can our Patients Actually Use? International Journal of Medical Informatics. 2005; 74(1):13–19. [PubMed: 15626632]
8. Wang Y, Liu Z. Automatic Detecting Indicators for Quality of Health Information on the Web. International Journal of Medical Informatics. 2007; 76:575–582. [PubMed: 16750417]

9. Choi J, Bakken S. Web-based education for low-literate parents in Neonatal Intensive Care Unit: Development of a website and heuristic evaluation and usability testing. International Journal of Medical Informatics. 2010; 79(8):565–575. [PubMed: 20617546]
10. Janz NK, Becker MH. The Health Belief Model: A Decade Later. Health Education Quarterly. 1984; 11:1–47. [PubMed: 6392204]
11. Davis TC, Dolan NC, Ferreira MR, Tomori C, Green KW, Sipler AM, Bennett CL. The Role of Inadequate Health Literacy Skills in Colorectal Cancer Screening. Cancer Investigation. 2001; 19(2):193–200. [PubMed: 11296623]
12. Mouradi, O.; Leroy, G.; Kauchak, D.; Endicott, JE. Influence of Text and Participant Characteristics on Perceived and Actual Text Difficulty. Proc. Hawaii International Conference on System Sciences; Maui. HI. 2013.
13. Boscolo P, Mason L. Topic Knowledge, Text Coherence, and Interest: How they Interact in Learning from Instructional Texts. The Journal of Experimental Education. 2003; 7(2):126–148.
14. Mullan J, Crookes PA, Yeatman H. University of Wollongong juea, University of Wollongong pcuea. Rain fog, smog and printed educational material. Journal of Pharmacy Practice and Research. 33(4):284–286.
15. McLaughlin GH. SMOG Grading: a New Readability Formula. Journal of Reading. 1969; 12:636–646.
16. Adkins AD, Singh NN. Reading Level and Readability of Patient Education Materials in Mental Health. Journal of Child and Family Studies. 2001; 10(1):1–8.
17. Rathbun A, Thornton LA, Fox JE. Are Our Investments Paying Off? A Study of Reading Level and Bereavement Materials. American Journal of Hospice and Palliative Medicine. 2008 [Epub ahead of print].
18. Brandt H, McCree D, Lindley L, Sharpe P, Hutto B. An evaluation of printed HPV educational materials. Cancer Control. 2005; (Suppl 2):103–106. [PubMed: 16327760]
19. Brainard J. Study Finds Research Consent Forms Difficult to Comprehend. The Chronicle of Higher Education. 2003; 49(19):A21–A22. [PubMed: 15287115]
20. Maples P, Franks A, Ray S, Stevens A, Wallace L. Development and validation of a low-literacy Chronic Obstructive Pulmonary Disease knowledge Questionnaire (COPD-Q). Patient Education and Counseling. 2009; 81(1):19–22. [PubMed: 20044232]
21. Friedman DB, Corwin SJ, Dominick GM, Rose ID. African American Men's Understanding and Perceptions about Prostate Cacner: Why Multiple Dimensions of Health Literacy are Important in Cancer Communication. Journal of Community Health. 2009; 34:449–460. [PubMed: 19517223]
22. Wubben, S.; van den Bosch, A.; Krahmer, E. Sentence Simplification by Monolingual Machine Translation. Proc. Proceedings of thr Annual Meeting of the Association for Computational Linguistics; Jeju Island, Korea. 2012.
23. Specia, L.; Jauhar, SK.; Mihalcea, R. SemEval-2012 Task 1: English Lexical Simplification. Proc. Joint Conference on Lexical and Computerational Semantics (*SEM); Montreal, Canada. 2012.
24. Biran, O.; Brody, S.; Elhadad, N. Putting it Simply: A Context-Aware Approach to Lexical Simplification. Proc. Proceedings of Association of Computational Linguistics; Portland, OR. 2011.
25. Kandula, S.; Curtis, D.; Zeng-Treitler, Q. A Semantic and Syntactic Text Simplification Tool for Health Content. Proc. AMIA Annu Symp Proc; 2010.
26. O'Donnell ME. Finding Middle Ground in Second Language Reading: Pedagogic Modifications that Increase Comprehensibility and Vocabulary Acquisition while Preserving Authenthic Text Features. The Modern Language Journal. 2009; 93:512–533.
27. McNamara DS, Kintsch E, Songer NB, Kintsch W. Are Good Texts Always Better? Interactions of Text Coherence. Background Knowledge, and Levels of Understanding in Learning from Text', Cognition and Instruction. 1996; 14(1):1–43.
28. Adkins AD, Elkins EN, Singh NN. Readability of NIMH Easy-to-Read Patient Education Materials. Journal of Child & Family Studies. 2001; 10(3):279.
29. Kirkpatrick MAF, Mohler CP. Using the Readability Assessment Instrument to Evaluate Patient Medication Leaflets. Drug Information Journal. 1999; 33:557–563.

30. Norvig, P. Inference in Text Understanding. Proc. National Conference on Artificial Intelligence AAAI-87; Seattle, Washington. 1987.

31. Goldman SR, Murray JD. Knoweldge of Connectors as Cohesion Devices in Text: A Comparative Stud of Native-English and English-as-a-Second-Language Speakers. Journal of Educational Psychology. 1992; 84(4):504–519.

32. Van Servellen G, Brown JS, Lombardi E, Herrera G. Health Literacy in Low-Income Latino Men and Women Receiving Antiretroviral Therapy in Community-Based Treatment Centers. Aids Patient Care and STDs. 2003; 17(6):283–298. [PubMed: 12880492]

33. Potelle H, Rouet J-F. Effects of Content Representation and Readers' Prior Knowledge on the Comprehension of Hypertext. International Journal of Human-Computer Studies. 2003; 58:327–345.

34. Davis T, Long S, Jackson R, Mayeaux E, George R, Murphy P, Crouch M. Rapid Estimate of Adult Literacy in Medicine: a Shortened Screening Instrument. Family Medicine. 1993; 25(6): 391–395. [PubMed: 8349060]

35. McWhirter J, Todd L, Hoffman-Goetz L. Comparing Witten and Oral Measuress of Comprehension of Cancer Information by English-as-a-Second-Language Chinese Immigrant Women. Journal of Cancer Education. 2011; 26:484–489. [PubMed: 21445682]

36. Gazmararian JA, Kripalani S, Miller MJ, Echt KV, Ren J, Rask K. Factors associated with medication refill adherence in cardiovascular-related diseases: a focus on health literacy. Journal of General Internal Medicine. 2006; 21(12):1215–1221. [PubMed: 17105519]

37. Cohen S, Kamarck T, Mermelstein R. A Global Measure of Perceived Stress. Journal of Health and Social Behavior. 1983; 24(4):385–396. [PubMed: 6668417]

38. Leroy G, Miller T. Perils of Providing Visual Health Information Overviews for Consumers with Low Health Literacy or High Stress. Journal of the American Medical Informatics Association. 2010; 17:220–223. [PubMed: 20190068]

39. Taylor WL. Cloze procedure: A new tool for measuring readability. Journalism Quarterly. 1953; 30:415–433.

40. Seligman HK, Wallace AS, DeWalt DA, Schillinger D, Arnold CL, Shilliday BB, Delgadillo A, Bengal N, Davis TC. Facilitating Behavior Change with Low-literacy Patient Education Materials. American Journal of Health Behavior. 2007; 31(Suppl 1):S69–S78. [PubMed: 17931139]

41. Klein, D.; Manning, CD. Accurate Unlexicalized Parsing. Proc. 41st Meeting of the Association for Computational Linguistics; 2003.

42. Leroy, G.; Endicott, JE. Term Familiarity to Indicate Perceived and Actual Difficulty of Text in Medical Digital Libraries. Proc. International Conference on Asia-Pacific Digital Libraries (ICADL 2011) - Digital Libraries -- for Culture Heritage, Knowledge Dissemination, and Future Creation; Beijing, China. October 24–27 2011;

43. Leroy, G.; Endicott, JE. Combining NLP with Evidence-based Methods to Find Text Metrics related to Perceived and Actual Text Difficulty. Proc. 2nd ACM SIGHIT International Health Informatics Symposium (ACM IHI 2012); Florida, Miami. January 28–30 2012;

44. Zarcadoolas C. The Simplicity Complex: Exploring Simplified Health Messages in a Complex World. Health Promotion International. 2010; 26(3):338–350. [PubMed: 21149317]

45. Ledoux K, Gordon PC, Camblin CC, Swaab TY. Coreference and Lexical Repetition: Mechanisms of Discourse Integration. Memory & Cognition. 2007; 35(4):801–815.

46. Ellis, WD. A source book of Gestalt psychology. Vol. 1938. Paul, Trench, Trubner & Co; 1938.

47. Stothers M, Klein PD. Perceptual Organization. Phnological Awareess, and Reading Comprehension in Adults with and without Learning Disabilities.', Annals of Dyslexia. 2010; 60(2):209–237.

48. Paolacci G, Changler J, Ipeirotis PG. Running Experiments on Amazon Mechanical Turk. Running Experiments on Amazon Mechanical Turk. 2010; 5(5):411–419.

49. Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk A New Source of Inexpensive, Yet High-Quality, Data? Perspectives on Psychological Science. 2011; 6(1)

50. Novotney, S.; Callison-Burch, C. Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-expert Transcription. Proc. Proceedings of North American Association for Computational Linguistics; 2010.

51. D'Alessandro D, Kingsley P, Johnson-West J. The Readability of Pediatric Patient Education Materials on the World Wide Web. Arch Pediatr Adolesc Med. 2001; 155:807–812. [PubMed: 11434848]

52. Kasabwal, aK; Agarwal, N.; Hansberry, D.; Baredes, S.; Eloy, J. Readability assessment of Patient Education Materials from the American Academy of Otolaryngology--Head and Neck Surgery Foundation. Otolaryngol Head Neck Surg. 2012; 147(3):466–471. [PubMed: 22473833]

53. Schmitt P, Prestigiacomo C. Readability of Neurosurgery-Related Patient Education Materials Provided by the American Association of Neurological Surgeons and the National Library of Medicine and National Institutes of Health. World Neurosurg. 2011 (in press).

54. Leroy G, Helmreich S, Cowie J. The Influence of Text Characteristics on Perceived and Actual Difficulty of Health Information. International Journal of Medical Informatics. 2010; 79(6):438–449. [PubMed: 20202895]

55. Leroy, G.; Helmreich, S.; Cowie, JR. The Effects of Linguistic Features and Evaluation Perspective on Perceived Difficulty of Medical Text. Proc. Hawaii International Conference on System Sciences (HICSS); Kauai. January 5–8 2010;

Summary Points

What is already known

- Online text is generally considered too difficult for laymen to read and understand
- Readability formulas are the most commonly used tool to simplify text
- There is little evidence that 'fixing' text using readability formulas results in better understanding of the text

New insights

- Actual and perceived difficulty are two different text characteristics
- Lexical simplification reduces the perceived difficulty of text
- Coherence enhancement reduces the actual difficulty of text

Research Highlights

- There exists a distinction between actual and perceived difficulty of text
- The study showed how lexical simplification reduced perceived difficulty of text.
- The study showed how coherence enhancement reduced actual difficulty of text.

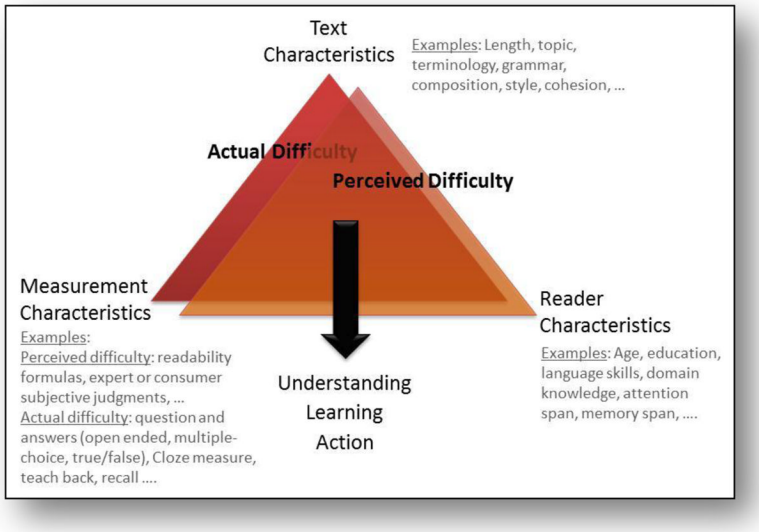

Figure 1.
Characteristics Important in Studies Measuring User Understanding

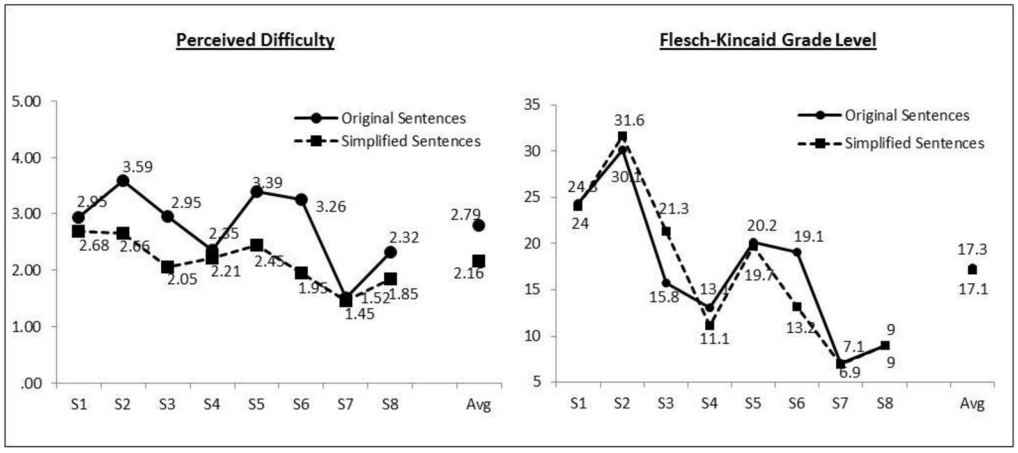

Figure 2.
Perceived Difficulty (1 = Easiest, 5 = Hardest) and Flesch-Kincaid Grade Levels for the eight sentence pairs.

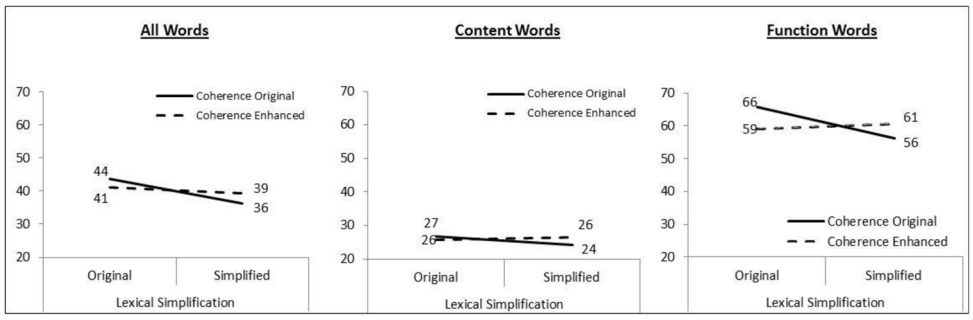

Figure 3.
Actual Difficulty Measured by Percentage Correct on the Cloze Test

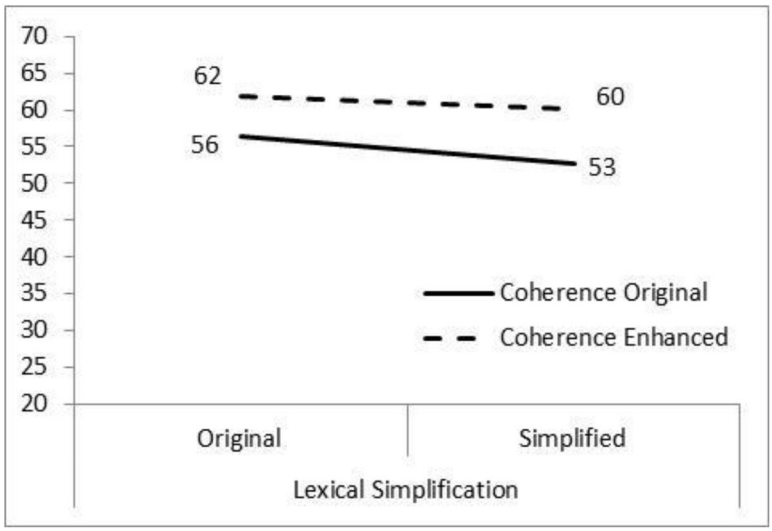

Figure 4.
Actual Difficulty Measured by Percentage Correct on the Multiple-Choice Questions

Table I

Sentence and Document Characteristics

SENTENCES				
Lexical Simplification:	**Original (N = 8)**		**Simplified (N = 8)**	
Word Count (Avg.)	24		32	
Flesch-Kincaid Grade Level (Avg.)	17.3		17.1	
DOCUMENTS				
Coherence:	Original		Enhanced	
Lexical Simplification:	Original (N=2)	Simplified (N=2)	Original (N=2)	Simplified (N=2)
Word Count (Avg.)	231	264	234	262
Flesch-Kincaid Grade Level (Avg.)	21	21	18	16
Cloze Measure:				
All Blanks Count (Avg.)	33	37	33	36
Content Word Blanks Count (Avg.)	20	23	20	22
Function Word Blanks Count (Avg.)	13	14	13	14

Int J Med Inform. Author manuscript; available in PMC 2014 August 01.

Table II

Participant Demographic Information (Group 1 = Original & Lexical Simplified Documents, Group 2: Coherence Enhanced Original & Coherence Enhanced Lexical Simplified Documents)

Characteristic	All Participants		Group 1		Group 2	
	N = 187	(%)	N = 93	(%)	N = 94	(%)
Age						
20 or younger	13	(7)	4	(4)	9	(10)
21–30	83	(44)	41	(44)	42	(45)
31–40	44	(23)	26	(28)	18	(19)
41–50	22	(12)	12	(13)	10	(11)
51–60	16	(9)	7	(8)	9	(10)
61–70	9	(5)	3	(3)	6	(6)
71 or older	-	-	-	-	-	-
Gender						
Female	119	(64)	55	(59)	64	(68)
Male	68	(36)	38	(41)	30	(32)
Race (Multiple choices allowed)						
American Indian/Native Alaskan	1	(.5)	-	-	1	(1)
Asian	31	(17)	15	(16)	16	(117)
Black or African American	12	(6)	3	(3)	9	(10)
Native Hawaiian or Other Pacific Islander	3	(1.5)	-	-	3	(3)
White	142	(76)	75	(81)	67	(71)
Ethnicity						
Hispanic or Latino	14	(8)	4	(4)	10	(11)
Not Hispanic or Latino	173	(92)	89	(96)	84	(89)
Location						
North America	162	(87)	82	(88)	80	(85)
South America	-		-	-	-	-
Africa	1	(.5)	-	-	1	(1)

Characteristic	All Participants		Group 1		Group 2	
	N = 187	(%)	N = 93	(%)	N = 94	(%)
Europe	4	(2)	-	-	4	(4)
Asia	20	(11)	11	(12)	9	(10)
Australia or Oceania	-		-	-	-	-
Education (Highest Completed)						
Less than High School	4	(2)	4	(4)	-	-
High School Diploma	68	(36)	29	(31)	39	(41)
Associate's Degree	30	(16)	17	(18)	13	(14)
Bachelor's Degree	57	(31)	26	(28)	31	(33)
Master's Degree	23	(12)	13	(14)	10	(11)
Doctorate	5	(3)	4	(4)	1	(1)
Language Skills (Frequency of Speaking English at Home)						
Never English	2	(1)	-	-	2	(2)
Rarely English	10	(5)	5	(5)	5	(5)
Half English	15	(8)	8	(9)	7	(7)
Mostly English	13	(7)	7	(8)	6	(6)
Only English	147	(79)	73	(78)	74	(79)

Table III

Participant Reading Habits (N=187)

	BOOKS		
	Fiction	**Non-fiction**	**Textbooks**
	N (%)	**N (%)**	**N (%)**
Not at all	21 (11)	29 (15)	66 (36)
Less than 1 per year	23 (12)	48 (26)	27 (14)
About one per year	20 (11)	44 (24)	25 (13)
2–6 per year	50 (27)	41 (22)	39 (21)
More than 6 per year	73 (39)	25 (13)	30 (16)

	PRINTED NEWS AND MAGAZINES		
	News magazines	Newspapers	Other magazines
Not at all	63 (34)	51 (27)	42 (23)
Less than 1 per week	78 (42)	69 (40)	83 (44)
About one per week	34 (18)	29 (16)	35 (19)
2–6 per week	7 (4)	27 (14)	23 (12)
More than 6 per week	5 (3)	11 (6)	4 (2)

	INTERNET			
	Online news	Social media	Comm.	Notifications
Not at all	14 (8)	21 (11)	1 (1)	60 (32)
Less than 1 per week	34 (18)	26 (14)	10 (5)	37 (20)
About one per week	28 (15)	13 (7)	12 (6)	28 (15)
2–6 per week	48 (26)	39 (21)	48 (26)	27 (14)
More than 6 per week	63 (34)	88 (47)	116 (62)	35 (19)

Table IV

Correlations between User Characteristics and Perceived and Actual Difficulty

	Perceived Difficulty	Actual Difficulty	
		Cloze Measure (all blanks)	Multiple-Choice Questions
Age	.104	−.056	−.057
General Education	.132	.006	−.024
Medical Education	−.094	.029	−.051
Language skills	−.014	−.076	.092
Health Literacy (S-TOFHLA)	−.159*	−.232**	.032
Reading Habits			
Books	−.031	.172*	.119
Printed News and Magazines	.035	−.018	−.038
Internet	−.030	.017	−.043

N=187, Pearson r two-tailed correlation coefficient,

* p < .05,

** p < .01