Creating a Gold Standard for the Readability Measurement of Health Texts

Sasikiran Kandula, MS[1], Qing Zeng-Treitler PhD[1, 2]
[1]Decision Systems Group, Brigham and Women's Hospital, Boston, MA; [2]Harvard Medical School, Boston, MA

Abstract

Developing easy-to-read health texts for consumers continues to be a challenge in health communication. Though readability formulae such as Flesch-Kincaid Grade Level have been used in many studies, they were found to be inadequate to estimate the difficulty of some types of health texts. One impediment to the development of new readability assessment techniques is the absence of a gold standard that can be used to validate them. To overcome this deficiency, we have compiled a corpus of 324 health documents consisting of six different types of texts. These documents were manually reviewed and assigned a readability level (1-7 Likert scale) by a panel of five health literacy experts. The expert assigned ratings were found to be highly correlated with a patient representative's readability ratings (r = 0.81, p<0.0001).

Introduction

According to an Institute of Medicine report [1], "Nearly half of all American adults—90 million people—have difficulty understanding and acting upon health information." It stated that "Even people with strong literacy skills may have trouble obtaining, understanding, and using health information," and "Although causal relationships between limited health literacy and (poor) health outcomes are not yet established, cumulative and consistent findings suggest such a causal connection."

These statements suggest that communicating health information continues to be a challenge and good literacy skills, though necessary, are not sufficient to understand and use health information. Researchers have tried to address this information need by developing more 'readable' health material.

To assess the readability of health material, researchers have largely relied on readability formulae such as the Simplified Measure of Gobbledygook (SMOG), the Fry Readability Scale (FRY), and the Flesch-Kincaid Grade Level (FKGL) [2]. Most of these formulae were developed decades ago and estimate the complexity of the text mainly by word and sentence length. For instance SMOG is calculated as:

$$SMOG = 1.043 * \left(\sqrt{pw * \left(\frac{30}{sentences} \right)} \right) + 3.1291$$

where *pw* is the number of words having 3 or more syllables. Similarly, FKGL is defined as:

$$FKGL = \left(0.39 * \frac{words}{sentence} \right) + \left(11.8 * \frac{syllables}{word} \right) - 15.59$$

More recent research on readability has pointed out some of the limitations of these formulae. In particular, researchers have noted that cohesion or coherence between sentences is an important factor in readability [3, 4]. A piece of text with short sentences yet low cohesion between sentences can be less understandable than one that has longer sentences and high cohesion.

Most research on readability has focused on free-text prose, while content organization, layout and design also matter. To measure the overall readability of materials, Doak and Doak developed the Suitability Assessment of Materials (SAM) [5]. To measure the readability of charts and graphs, PMOSE/IKIRSCH has been developed by Mosenthal and Kirsch [6]. Both SAM and PMOSE, however, are complex instruments and not computerized.

While health-specific literacy tests are commonly used in medical research, there have been few health-specific measurements for text readability. When assessing the readability of health material, the difficulty of health vocabulary and concepts needs to be taken into consideration [7]. The use of word length or an easy-word list though simple is insufficient, since most health-related terms are deemed difficult enough that they don't appear on easy-word lists – in spite of the fact that some (e.g. diabetes) are familiar to the general public. Also, to understand some concepts (for instance, concepts related to metabolic pathways or pathophysiologic processes) we would require higher level knowledge. Furthermore, in some cases, the words that are being used may be easier than the concepts they are trying to describe.

Recently, we and other researchers have begun developing health-specific readability measures ([8, 9, 10]). The new readability measure we are developing attempts to take medical vocabulary, cohesion, and style into consideration while evaluating health texts.

One major challenge we encountered in the development of our measure, is the validation of its efficacy. To test general readability formulae that assign grade levels to documents, one may use the standard school textbooks as a gold standard. But in the health domain, there is no equivalent gold standard. This makes it difficult to demonstrate quantitatively that the new health-specific readability measures are indeed better than the existing formulae. In this paper, we describe our effort to develop a gold standard dataset for health text readability.

Background

In the health domain, researchers have used several approaches to estimate the "real" or gold standard text readability when developing readability assessment tools. Gemoets et al [8] applied Cloze test to 20 documents using 40 subjects. Cloze test [11] is a widely accepted method to measure text readability and comprehension. In this test, typically, every fifth word of the first 250 words of a document is replaced with a blank and subjects are asked to make a context-based guess of the missing word. The Cloze score of a document is calculated as the percentage of correct answers from the subjects. One limitation of the Cloze test is that it is time-consuming and challenging for the subjects, which limits the number of documents that can be tested on each subject. In Gemoets' study each subject received 2 documents and each document was tested on 4 subjects. At the same time, health texts and consumers are very diverse and a reasonable gold standard should include enough documents to accommodate these differences. Therefore, evaluating a large number of documents on a large number of subjects using Cloze test is daunting.

Leroy et al [10] used a corpus of 250 documents which were assumed to be easy, intermediate or difficult to read depending on their source and target audience. The easy documents were 'medically themed' blog entries written by lay users. The documents at the intermediate level were written by 'professionals to educate lay users'. The difficult set of documents consisted of journal articles. As such, the actual readability level of these texts was presumed but never tested.

Similarly, in one of our previous studies, we used examples of easy, moderate and difficult texts to test our readability assessment tool [9]. In this case, we used self-labeled easy-to-read consumer education materials as examples of easy documents, while news stories and scientific journal articles were used as examples of moderate and difficult texts respectively. Here we have assumed that the self-labeled consumer education materials are generally easier than health news stories and the news stories are easier than medical journal articles. While the authors of the paper and most researchers consider this to be a reasonable assumption, a few researchers have expressed reservations.

Methods

We assembled a panel of 5 health literacy and clinical experts and a patient representative to assess the readability of a set of 324 documents. To obtain a diverse sample, we selected 6 different types of documents: consumer education materials, news stories, clinical trial records, clinical reports, scientific journal articles and consumer education material targeted at kids (Table 1). All documents in the set are related to diabetes mellitus. As can be seen in the table, we included a much larger number of consumer education material in the corpus since these are more likely to be provided to and read by consumers.

Document Type	Count
Consumer education material (CEM)[1]	142
News report (NWS)[2]	34
Clinical trial record (CLT)[3]	39
Scientific journal article (JNL)[4]	38
Clinical report(REP)[5]	38
Consumer Education Material targeted at kids (KID)[6]	33

Table 1. Sample size for each document type

The experts included: a health literacy consultant with extensive experience in content development and health communication education, a health communication researcher with background in education psychology and nursing, a patient education content developer and nurse, a practicing nurse and certified diabetes educator, and a librarian from a hospital-based consumer health library. The patient representative is the family member of a diabetes patient.

The expert panel went through the following three steps to assess the readability of the corpus:

1. *Establish initial rating criteria*. The initial rating criteria required the experts to rank the readability of documents on a 1-7 Likert scale.

[1] Source: MedlinePlus, National Institute of Diabetes and Digestive and Kidney Diseases, Mayo Clinic
[2] Source: The New York Times, CNN. BBC, TIME
[3] Source: ClinicalTrials.gov
[4] Source: DiabetesCare, Annals of Internal Medicine, Circulation - Journal of the American Heart Association, Journal of Clinical Endocrinology and Metabolism and the British Medical Journal
[5] Source: Brigham and Women's Hospital internal records
[6] Source: American Diabetes Association

The rating of 1 represents the easiest level, i.e. readers with elementary school level English proficiency and minimum health literacy should be able to understand the material; 7 represents the most difficult level, i.e. readers need to have college-level English proficiency and professional-level knowledge in a particular health domain (e.g. biochemistry or epidemiology) to understand the material. Using these criteria, the panel reviewed a balanced set of 12 documents, i.e. two documents of each type.

2. *Group review to refine the rating criteria*. In this step, the ratings for the above set of documents were compared and analyzed. It was found that, for a majority of the documents (n=8), the ratings by the panel members were fairly consistent: the maximum difference between an individual's rating and the average rating was lower than 1.16. For the other four documents there was a higher variation in ratings. These disagreements between the experts were resolved through discussion. Based on these results, the panel refined the rating criteria as below:

Rating	Description
1	Can be understood by anyone with basic literacy
3	Can be understood by an average high school graduate
4	Can be understood by an average reader with some college education
7	Can be understood only by someone with professional education/training in a health domain.

Table 2. Rating guidelines used by the experts

The rating levels 2, 5, and 6 were used to represent the 'in-between' cases.

While the experts recognized that the document's style (such as font, sub-headings, bulleted lists) affects the document's readability, they decided to place more emphasis on the textual characteristics (such as vocabulary, sentence structure, voice).

Using the refined criteria, the panel reviewed a second set of 12 documents. All documents were presented as PDF files and made available to the reviewers through a web interface. Some documents that were longer were truncated to five pages.

Compared to the first set, the ratings for the second set of documents showed better inter-rater agreement. A majority of the expert's ratings, 62.5%, differed by 0.33 or less with the average rating and only 5.55% differed by level 1 or more.

3. *Individual review*. Each of the 5 experts was assigned 60 different documents. The patient representative was randomly assigned 60 of the 300 documents that were to be reviewed by the experts. At the end of the review process, we calculated the correlation between the patient representative's and the experts' ratings. The Spearman correlation coefficient was found to be 0.81 ($p < 0.0001$), which provided an extra validation to the experts' rating.

We analyzed the distribution of the expert assigned readability ratings by level and by document type. To obtain a sense how the experts' assessment differs from that of popular readability formulae, we also applied FKGL and SMOG formulae to the documents and calculated their correlation with the gold standard.

Results

Through expert panel review, we were able to obtain the gold standard readability rating of a set of 324 documents. For the first set of documents, where each document was reviewed by all the experts, a consensus rating was used. For the second set of documents, the average rating was used.

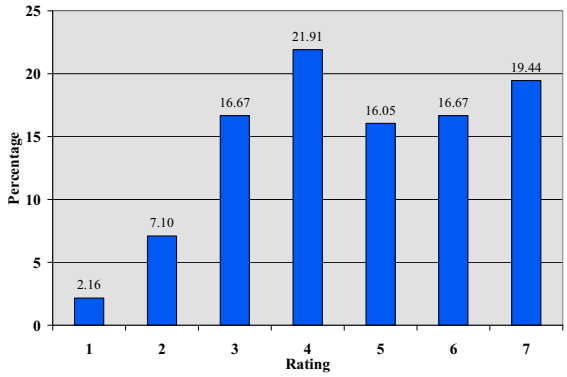

Figure 1. Rating Distribution

An analysis of the ratings showed that only 2.16% of the 324 documents have been rated 1. This indicates that the experts do not consider most of the documents, including those that have been designed for kids, to be understandable to people with basic literacy skills. In contrast, 74.06% were rated 4 or higher which indicates that these are appropriate for readers with at least some college education. Figure 1, gives the distribution of ratings among the 324 documents

We examined how the expert assigned ratings varied across the document types. Intuitively, we would

expect the education material aimed for kids to be the easiest, followed by the news articles and consumer health material, with the journal articles, personal health records and clinical trials occupying the difficult end of the spectrum. The expert ratings were found to concur with this general estimate of the difficulty of the documents (Table 3). However, each document type demonstrated a sizeable range of rating levels (Figure 2). For example, a few journal articles were rated 4 and 5 while some news articles were rated 6. This suggests that some journal articles could be easier than some news articles and we cannot assume a document's readability based solely on the document type.

Type	Mean	SD[7]	Q1[8]	Q3[9]	(min, max)
KID	2.39	0.86	2	3	(1, 5)
NWS	4.26	0.86	4	5	(3, 6)
CEM	4.02	1.18	3	5	(1, 7)
CLT	6.33	0.89	6	7	(4, 7)
JNL	6.55	0.72	6	7	(4, 7)
REP	6.13	0.84	6	7	(4, 7)

Table 3. Descriptive statistics for expert ratings

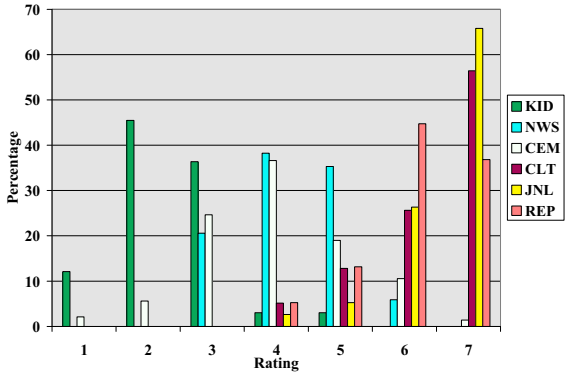

Figure 2. Readability level distribution over document types

The availability of these ratings, allowed us to examine how FKGL and SMOG perform in assessing the readability of health-related texts. The grades assigned by FKGL were found to have a Spearman correlation coefficient of 0.54 ($p < 0.0001$) with the expert ratings while SMOG grades had a similar correlation of 0.55($p < 0.0001$).

We calculated the mean FKGL and SMOG grades for each document type (Table 4). There is a significant difference between the difficulty of clinical reports as assessed by the experts and by FKGL/SMOG. While the experts rate these documents to be only slightly easier than the journal articles, both the readability formulae incorrectly rate these to be only slightly harder than the documents targeting kids.

We believe that the formulae's assessment of the clinical reports does not reflect the actual readability of the documents. For instance, a clinical report that was assigned a grade of 6.01 by FKGL (10.40 by SMOG) had the following excerpt: '*HEENT was normal. His lungs were clear to auscultation. Heart exam, S1, S2 normal. Regular rate and rhythm without murmurs, rubs or gallops. Abdomen soft and nontender. Extremities warm with 2+ pulses and there is no edema.*' The presence of short sentences and very few polysyllable words caused FKGL to assign this text a low grade. The human expert, however, assigned a rating 7 to this document.

Document Type	FKGL		SMOG	
	Mean	SD	Mean	SD
KID	6.56	1.59	9.83	1.40
NWS	11.29	2.22	13.68	1.80
CEM	10.19	2.05	12.62	1.69
CLT	15.71	2.37	17.10	1.93
JNL	15.82	1.43	17.24	1.07
REP	8.38	1.78	11.36	1.44

Table 4. Descriptive statistics for FKGL and SMOG grades (by document type)

Another observation is the formulae's under-estimation of document difficulty across all document types. For instance, majority of the news articles and consumer education materials were rated by experts at level 4 and above, which means they require some college education to comprehend. In contrast, the readability formulae rated more than half the news articles and consumer education materials to be at or below 12th grade (high-school) levels. An analysis of the formulae's grades for documents of each expert rating level (RL), showed similar results (Table 5).

RL	FKGL		SMOG	
	Mean	SD	Mean	SD
1	7.88	2.52	10.54	1.73
2	7.43	1.89	10.62	1.68
3	8.59	2.46	11.35	2.02
4	10.71	2.12	13.02	1.76
5	11.54	2.85	13.84	2.38
6	12.02	3.72	14.26	2.92
7	13.97	3.58	15.72	2.78

Table 5. Descriptive statistics for FKGL and SMOG grades (by rating level).

The underestimation appeared to be more pronounced for the more difficult documents. For the set of documents rated 7 by the experts (i.e. the most difficult documents), the minimum FKGL grade was 6.01 (maximum was 20.66).

[7] Standard Deviation
[8] First Quartile
[9] Third Quartile

Discussion

We developed a gold standard to evaluate the readability of health texts by employing a panel of health experts. As mentioned in the Background, there is no gold standard for evaluating existing readability formulae or developing new ones in the health domain. Compared to the Cloze test, the expert panel approach makes it feasible to assess a much larger sample of documents. In lieu of using documents of certain types (e.g. blogs or journal articles) as the gold standard for easy or difficult materials, expert ratings do not rely on the presumption of a particular type of document being inherently difficult or easy, and recognize the varying levels of difficulty within a type.

As expected, two commonly used readability formulae (i.e. FKGL and SMOG) had statistically significant but not very strong correlation with the gold standard expert ratings. This finding is consistent with the results reported by Gemoets et al. [8] and suggests that there is room for improvement in terms of health readability assessment.

This study can help identify characteristics that differentiate the documents deemed easy (rating 1-3) from those considered hard. These features will be of interest to those developing readability assessment tools and health content creators. In addition, the corpus when made publicly available can be used to validate and compare new readability formulae.

We need to point out that the FKGL and SMOG grades reported here were calculated using the available formulae. However, given the differences in rules governing annotation of words and sentences and the non-trivial task of counting syllables, the grades may vary slightly from other implementations.

This study has several limitations. The gold standard we created relies on expert knowledge, rather than direct user testing. Ideally, we would like to test each document on a representative sample of users. However, adult health care consumers are extremely diverse. Given the formidable cost of testing a large number of documents on a large number of patients, we believe the experts' opinion can serve as a reasonable proxy.

We have included several different types of documents in the text corpus but not *all* the types of documents that consumers read (e.g. blogs and informed consent forms). The text corpus also has only a small number of very easy (level 1) documents, despite the inclusion of materials targeting kids.

As a next step, we intend to further validate the gold standard by recruiting more patients. We also intend to use this gold standard to test the health-specific readability measures that we have developed and are in the process of refining.

Acknowledgements

This work is supported by the NIH grant R01 LM07222 and R01 DK 075837. We would like to thank Chantal Friedman for coordinating this study and our panel members Dorothy Curtis, Ann Furey, Cara Helfner, Helen Osborne, Andrea Penney and Katie Weinger for reviewing the documents.

References

1. Nielsen-Bohlman L, Panzer AM, Kindig DA. Health Literacy: A Prescription to End Confusion. Washington, DC: National Academy Press 2004.
2. Osborne H. Health Literacy From A To Z : Practical Ways To Communicate Your Health: Jones & Bartlett Pub 2004.
3. Halliday MA, Hasan R. Cohesion in English. London: Longman 1976.
4. 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):1-43.
5. Doak CC, Doak LG, Root JH. Teaching Patients With Low Literacy Skills. 2nd ed: Lippincott Williams & Wilkins 1996.
6. Mosenthal P, Kirsch I. A New Measure of Assessing Document Complexity: The PMOSE/IKIRSCH Document Readability Formula. Journal of Adolescent & Adult Literacy. 1998;41(8):638–57.
7. Rosemblat G, Logan R, Tse T, Graham L. How Do Text Features Affect Readability? Expert Evaluations on Consumer Health Web Site Text. MEDNET; 2006; Toronto, CA; 2006. p.
8. Gemoets D, Rosemblat G, Tse T, Logan R. Assessing Readability of Consumer Health Information: An Exploratory Study. Medinfo. 2004;11(Pt 2):869-73.
9. Kim H, Goryachev S, Rosemblat G, Browne A, Keselman A, Zeng-Treitler Q. Beyond Surface Characteristics: A New Health Text-Specific Readability Measurement. AMIA 2007.
10. G. Leroy, T. Miller, G. Rosemblat, and A. Browne. A Balanced Approach to Health Information Evaluation: A Vocabulary-based Naïve Bayes Classifier and Readability Formulas. Journal of the American Society for Information Science and Technology. Forthcoming 2008.
11. Taylor WL. Recent Developments in the Use of "Cloze Procedure". Journalism Quarterly. 1956 33, 42-48.