

Usability testing reveals more than readability formulas reveal.

Readability Formulas Have Even More Limitations Than Klare Discusses

Janice (Ginny) Redish
 Redish & Associates, Inc.
 6820 Winterberry Lane
 Bethesda, MD 20817
 redish@ari.net

Abstract

A literature review reveals many technical weaknesses of readability formulas (when compared to direct usability testing with typical readers): they were developed for children's school books, not adult technical documentation; they ignore between-reader differences and the effects of content, layout, and retrieval aids on text usefulness; they emphasize countable features at the expense of more subtle contributors to text comprehension.

1.7.5 Document analysis—human factors

Keywords: usability testing, reliability, text comprehension

George Klare's *The Measurement of Readability* is based on a two-part premise:

- Clear, understandable text is a necessary (although not sufficient) characteristic of a usable and useful document.
- Readability formulas, despite their flaws, are valuable assessment tools.

I agree with the first part of Klare's premise but not with the second. We have today much more powerful assessment tools, particularly usability testing. Although not as quick and cheap as just asking the computer to show a readability score, usability testing can be done rapidly and inexpensively. Even testing with a few people is much more informative than a readability score. In fact, as we will see in this commentary, a readability score is seldom useful and can, indeed, be misleading.

Klare's book was an important contribution when it appeared, both because he showed how important it is to be concerned with readability in documents and because he gave us a thorough explanation of the history and research behind many different formulas. Klare himself recognized several limitations of readability formulas. However, his listing of those limitations in chapter one is overshadowed by the rest of the book—which is all about the formulas.

What is a Readability Formula?

We should, perhaps, step back and make sure that we all understand what a readability formula is. A readability formula is a mathematical equation that is meant to *predict* the level of reading ability needed to understand a particular piece of prose. Readability formulas are based on *correlations* with some measure of comprehension (such as scores on a reading test or scores of another formula). They say nothing about the *causes* of any problems people might have in understanding a document. Usability testing, on the other hand, can show you whether people *actually* do have problems with the document, *where* people have problems, and the *types of problems* they have. From that information, you can usually analyze the *causes* of the problems and fix them. Readability formulas give you no help in finding or fixing problems.

How Valid are Readability Formulas for Technical Material for Adult Readers?

No one knows. For technical materials for adults, using any readability formula means generalizing from situations that are decades old and not directly relevant to the audience or type of material.

Grade-level Formulas were Meant for Children's School Books

Readability formulas were originally developed to try to ensure that a school textbook for a particular grade was appropriate for children at that grade level. Even for that purpose, the use of the traditional readability formulas is questionable today. The research on grade-level readability formulas is more than 50 years old; and more than twenty years ago, studies showed that schoolchildren weren't reading on the same level as they had been when the formulas were developed (Jacobson, Kirkland, and Selden, 1978). The formulas are out of date.

Most of us in SIGDOC are writing technical, legal, or business materials for adults. How should we understand the use of a formula meant for school textbooks on materials meant for adults? What does it mean to say "eighth-grade reading level" when referring to adults? An adult who reads at an "eighth-grade level" may be a poor reader but may have a large spoken vocabulary from life experiences far beyond any eighth grader. An adult who still reads at an "eighth-grade level" may have had negative experiences in school and find it much harder to deal with documents than an eighth grader who is reading at grade level. The two audiences (children in school at grade level and adults with a low reading level) are so different that the same readability formula cannot possibly be adequate for both. But readability formulas do not distinguish audiences. Usability testing does. Usability testing will show you how to develop a document that will work for your low-reading-level adult readers far better than a readability formula.

Moreover, as Thomas Duffy points out (Duffy, 1985), the accepted correlation in the grade-

level formulas is that, if 50% of the children at a given grade level got 50% of the questions on a reading passage correct, that passage was considered acceptable at that grade level. Should we be happy if 50% of our readers understand 50% of our documents?

Neither Commonly Used Formula was Developed for Technical Materials

The two most commonly used formulas for materials for adults are the Dale-Chall formula and the Flesch Reading Ease Scale.

The Dale-Chall formula measures sentence length and whether each word is on a list of acceptable words. The list was developed in the 1940s through a test with fourth-grade school children. What problems might there be with this for adult users of technical documents? First, many groups that use the Dale-Chall formula improve their scores by arbitrarily adding to the list of acceptable words those words that they *believe* their audiences know. No research has usually been done to show whether readers in that audience really know the words being added to the formula. Second, words may have more than one meaning. "Enter" is an acceptable word on the list; fourth-graders knew it. When the IRS puts their documents through the Dale-Chall formula, the formula does not trip over "enter." But "enter" on a tax form is not the same "enter" that fourth-graders know. (In fact, today, most fourth-graders first meaning for "enter" is likely to be a computer key—and not the "enter" that got on the list in 1948.) How meaningful is that list for today's adults reading today's technical and legal documents?

The Flesch Reading Ease Scale measures sentence length and number of syllables per 100 words. It gives a number from 0 to 100 with higher being better. Insurance laws in several states define clear language as a score of 40 or 50 on the Flesch Reading Ease Scale. However, no research was ever done on the correlation between those scores and users' ability to find what they need or understand what they find in insurance documents. Flesch based his scale on articles in popular magazines, not on technical

material. Moreover, he created the formula by correlations with older comprehension tests and other formulas, not by redoing the research with adult readers.

Before you use a readability formula, think about what, if anything, you are really learning from it.

What is the Case Against Using Readability Formulas?

Klare lists four limitations of readability formulas (pp. 24-25):

1. Formulas measure only style. They do not touch on content, organization, word order, format, or imagery. Nor do they “take into account the differing purposes, maturity, or intelligence of readers.”
2. Even relating to style, they measure only *difficulty*, not mood, tone, persuasive effectiveness, etc.
3. They do not measure difficulty *perfectly*. The score you get may depend on the sample that you choose.
4. They do not tell you that you have *good* style.

These and other limitations are worth exploring in some detail because most of us after considering them will realize that readability formulas are *not* the best way to measure the usability and usefulness of a document. We should begin by taking apart Klare’s first limitation. It has several critical parts.

Readability Formulas Assume All Readers Are Alike

As Jack Selzer and I pointed out some years ago (Redish and Selzer, 1985, p. 49), “people are not text-processing machines.” We wrote, “the underlying assumption of readability formulas—that any text for any reader for any purpose can be measured with the same formula—does not mesh with our current understanding of how people read and under-

...readability formulas are *not* the best way to measure the usability and usefulness of a document.

stand.” Klare says that it is critical for authors to know a lot about their readers (p. 12), but the formulas make no distinctions based on readers’ characteristics. For a usability test, you consider the range among your

users, bring in representatives of different user groups, and watch and listen to different people deal with the document, so that you can take those differences into account in developing a final version.

Readability Formulas Only Measure What Can Be Counted

To have an equation, one must be able to count whatever is going into that equation. Therefore, in constructing the formulas, developers of readability formulas invariably dismiss all the features that cannot be easily counted even though they know that these features influence the readability (usability) of a document.

Moreover, in the final formulas, the developers strive for ease of application as well as for predictive accuracy. Therefore, as I wrote in a review article some years ago, “the features included in the published formulas are usually chosen as much for how easy they are to count as for their predictive value” (Redish, 1980, p. 69). Most formulas select only one or two features to count: sentence length or syllables per word or whether each word is on a list of acceptable words.

Most Of What Makes A Document Usable Is Not Included In Readability Formulas

What cannot be so easily counted? Here are some important aspects of documents that are totally ignored by readability formulas:

- Is the content right for the audience?
- Is the document organized so that users can find what they need?
- Are there any headings? Are the headings meaningful and useful to the audience?
- Is there a table of contents? Is it useful?

- Is there an index? Does it have users' words in it?
- Does the page layout help users find what they need?
- Are there visuals (tables, charts, screen shots, lists) to help users?
- Is the text divided into short sections and paragraphs (chunks) so that users are invited to use them?
- Are the sentences grammatical?
- Are the words ones that these users know?

Given this long list, it is difficult to see why anyone would want to limit the definition of a "readable" document to "short sentences" and "simple words," the features counted by readability formulas.

Readability Formulas Do Not Work On Forms, Web Pages, Or Documents With Lots Of Lists

Readability formulas assume that you are writing prose paragraphs. They count sentence length by going from period to period. If you use bulleted lists to chunk your material and lay your text out with white space, readability formulas will say you have long sentences. Yet, usability studies have consistently shown the value of lists and white space as aids to locating and understanding information.

Readability formulas were never meant to say anything about forms or other primarily visual material. As we come to understand the need to design information more visually on the web, readability formulas will have even less value for judging the usability of web pages. A good web page might have not a single complete prose sentence—and that would make a readability formula totally inappropriate as a measure of either readability or usability for web sites.

Readability Formulas Are Not Very Reliable

When we consider the other three limitations in Klare's own list (above and pages 24-25 of his chapter one), we realize that Klare is making two other points:

1. Readability formulas don't help you write well. They just count a few measures that

might relate to one aspect of reading—difficulty getting through a sentence. (This is from limitations two and four in Klare's list.)

2. The score you get from one part of a document may differ from the score you get in another part of the document. (This is from limitation three in Klare's list.)

In addition to the lack of reliability from passage to passage within a document, research has shown that the same passage may come out at very different grade levels on different formulas.

Improving Comprehension Does Not Correlate Well With Improving Readability Scores

If readability formulas were really measuring how well readers work with text, they would always improve when you revise a passage and increase scores on a comprehension test. But that isn't the case. In another publication, Klare (1976) reviewed 36 studies that attempted to improve comprehension by improving readability scores. Only about half succeeded, and to improve comprehension they had to change the readability scores by an average of 6.5 grade levels.

Even more telling, in a major study of legal documents, Charrow and Charrow (1979) tested how well people waiting for jury duty understood traditional jury instructions. After analyzing the problems that these potential jurors had with the instructions, the Charrows revised the instructions and did a second test. Comprehension went up. But the Charrows also looked at readability scores for both sets of instructions. In many cases, their revisions got better comprehension scores but worse readability scores. (This happened primarily because they added words to show the relationships among the information items in the instructions.)

So Why Do People Use Readability Formulas?

Readability formulas are text-based. They are easy to use. They don't require getting

real readers. They give you a number. And, therefore, they are seductive. But stop and think: What does the number you are getting mean? In fact, for adult readers of technical and legal documents, the number doesn't mean much, if anything.

If you do use a readability formula and your document gets a very poor score, that probably indicates that people will have problems with it. It probably has overly long sentences and long or unfamiliar words—and documents with overly long sentences often have many other problems. The poor score is a red flag that the document was probably developed without a focus on users, without any consideration of who the users are, what they know, how it should be organized for them, etc., etc. But the poor score doesn't tell you what else is wrong with the document (besides long sentences and long or unfamiliar words); nor does it give you any hints on how to fix it.

Therefore, using the formula for anything beyond seeing a poor score as a red flag is problematic for two reasons:

1. A good score does not mean you have a usable or useful document.
2. Rewriting to get a better score is misusing the formula.

A Good Score Does Not Mean You Have A Usable Or Useful Document

The most serious problem with using a readability formula is that a good score does not mean you have a usable or useful document. A good readability score barely touches the beginning of what makes a document work for users. As we have already seen, much else may be wrong with a document besides the length of the sentences or the length of the words. (See, for example, the list earlier in this commentary of ten questions that readability formulas do not address.)

Moreover, a short sentence may be difficult for readers. Even a short word may be difficult. To the Flesch Reading Ease Scale, these two sentences have identical scores:

- I wave my hand.
- I waive my rights.

They are not at the same level of difficulty for most readers.

A super short sentence may be more difficult than a longer one that shows the relationships among the elements in the sentences. Compare these passages (from Redish and Selzer, 1985, p. 49):

- He is the defendant. He is fifteen years old. He is in his teens. Someone says he stole from the store.
- The defendant is a fifteen-year old teenager who is accused of shoplifting.

The second, longer sentence is actually easier to understand, although it has the poorer readability score.

Rewriting To Get A Better Score Is Misusing The Formula

Klare, Flesch, Gunning, and all the other developers of readability formulas insist that the formulas are not to be used for revision. Remember that readability formulas are only *correlations*; they do not indicate *causes*. Klare explained it best (Klare, 1979) when he suggested that expecting comprehension to improve by writing to a readability formula is like lighting a match under a thermometer to warm up a room. The temperature on a thermometer is an index of how warm the room is. Lighting a match under the thermometer will make the index value go up, but the room won't get any warmer.

Long sentences are not a problem just because they are long. Length is only a corollary of several linguistic aspects that make sentences difficult. (See Redish and Selzer, 1985; Selzer, 1983; Holland and Campbell, 1982; and others for further discussion of what makes sentences difficult for readers.) Klare writes (p. 25), "the temptation simply to substitute an easy word for a hard one, to cut sentences in half, or to eliminate prepositional phrases even when they are needed may produce exactly the opposite of the desired result." And, as we have seen, the sentences may be only a minor part of what is making the document difficult for people.

The problem is that the temptation is likely to be too great when just a few changes will get

a better score. The temptation may be especially great when each draft can be quickly measured right in the word processor and when the writer is required to achieve a particular grade level or readability score.

What Is Better Than A Readability Formula?

As Jack Selzer and I wrote (Redish and Selzer, 1985, p. 49), “readability formulas are a simplistic answer to a very complex problem.” The methodology that allows you to deal with the complexity of a real document for real users is usability testing. In a usability test, representative users work with the document (or computer interface or other product or part of a product) while trained observers watch and listen and take notes. By analyzing what people actually do with a document, you can determine whether there are problems, where the problems are, and how to fix them. You can see problems in content, tone, organization, page or screen layout, typography, and other aspects of information design, as well as style and word choice. Since Klare wrote *The Measurement of Readability*, we have learned a great deal about how to evaluate documents with real people relatively quickly, inexpensively, and easily. Resources abound to help you with usability evaluations, including Dumas and Redish (1993, revised 1999), Nielsen (1993), Rubin (1994), Schriver (1997), Hughes (1999).

References

- Charrow, R. P. and Charrow, V. R. (1979). Making legal language understandable: A psychological study of jury instructions. *Columbia Law Review*, 79, 1306-1374.
- Dumas, J. S. and Redish, J. C. (1999). *A Practical Guide to Usability Testing*. Exeter, England: Intellect (revised edition); originally published by Ablex, 1993.
- Duffy, T. M. (1985). Readability formulas: What's the use? In T. M. Duffy and R. M. Waller (Eds.), *Designing Usable Texts* (pp. 113-143). NY: Academic Press.
- Holland, V. M. and Campbell, L. J. (1982). Understanding the language of public documents because readability formulas don't. In R. J. Di Pietro (Ed.), *Linguistics and the Professions* (pp. 157-171). Hillsdale, NJ: Erlbaum.
- Hughes, M. (1999). Rigor in usability testing. *Technical Communication*, 46(4), Fourth Quarter, 488-494.
- Jacobson, M. D., Kirkland, C. E., and Selden, R. W. (1978). An examination of the McCall-Crabbs standard test lessons in reading. *Journal of Reading*, 22, 224-230.
- Klare, G. R. (1979). *Readability Standards for Army-Wide Publications*. Fort Benjamin Harrison, IN: U. S. Army Administrative Center.
- Klare, G. R. (1976). A second look at the validity of readability formulas. *Journal of Reading Behavior*, 8, 129-152.
- Klare, G. R. (1963). *The Measurement of Readability*. Ames, IA: Iowa State University Press.
- Nielsen, J. (1993). *Usability Engineering*, San Diego, CA: Academic Press.
- Redish, J. C. (1980). Readability. In D. B. Felker (Ed.), *Document Design: A Review of the Relevant Research* (pp. 69-93). Washington, DC: American Institutes for Research.
- Redish, J. C. and Selzer, J. (1985). The place of readability formulas in technical communication. *Technical Communication*, 32(4), Fourth Quarter, 46-52.
- Rubin, J. (1994). *Handbook of Usability Testing*. NY: John Wiley & Sons.
- Schriver, K. A. (1997). *Dynamics in Document Design*. NY: John Wiley & Sons.
- Selzer, J. (1983). What constitutes a “readable” technical style? In P. V. Anderson, J. Brockmann, and C. R. Miller (Eds.), *New Essays in Technical and Scientific Communication: Research, Theory, and Practice* (pp. 71-89). Farmington, NY: Baywood.