Measuring readability

Part 1: The spirit is willing but the Flesch is weak

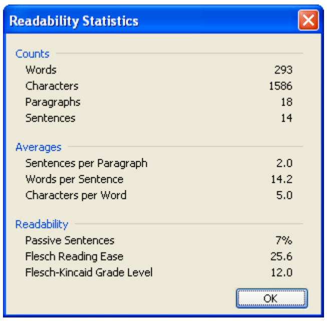
Geoffrey Marnell

This is the first part of a two-part paper exploring issues in readability. Part one assesses the most popular approach to measuring readability, epitomised by the Flesch reading-ease formula (the formula behind the readability scores one finds in Microsoft Word). The conclusion is that the variables that are fed into the Flesch formula are poor measures of readability and thus the formula itself is fundamentally flawed. The second part of this paper (see page 9) argues that the techniques commonly used to validate readability formulas lead to skewed results. Furthermore, redressing the methodological weakness in these techniques necessarily produces an even poorer correlation between independently measured readability scores and Flesch scores. This reinforces the conclusion that such formulas are best discarded.

Readability, like usability, is one of the central themes in the quest for good writing. To maximise readability is a goal that every writer, technical or otherwise, should strive to achieve. To argue otherwise is tantamount to arguing that we do not write to communicate, a bizarre view to say the least.

Most of us have at least a vague idea of what readability is. When we look at an act of parliament, or legal contract, drawn up before Plain English became popular we wonder how anyone could have deciphered it. However difficult a current act of parliament or legal contract might be, the difficulty pales into insignificance when compared to the difficulty inherent in like documents of yesteryear.

But readability is easier to sense than define. Many attempts have been made, some attributing readability to many factors, others to just a handful. Some readability researchers have proposed mathematical formulas to assess the readability of text. Perhaps the most well-known of these researchers is Rudolf Flesch, whose formula is the maths behind the readability scores generated by Microsoft Word (an example of which is shown at the right).



Readability Statistics

Counts	
Words	293
Characters	1586
Paragraphs	18
Sentences	14
Averages	
Sentences per Paragraph	2.0
Words per Sentence	14.2
Characters per Word	5.0
Readability	
Passive Sentences	7%
Flesch Reading Ease	25.6
Flesch-Kincaid Grade Level	12.0

The *Flesch reading-ease formula* is one of a family of readability formulas that associate readability solely with features of words and sentences. They take as variables one or more such features as the average number of words per sentence, the average number of syllables per word, the number of single-syllable words, the number of polysyllables, and the like. These are empirical features that any computer, indeed any person, can calculate. The simplicity of these formulas accounts, in part, for their continuing popularity. There is no need to learn anything about the reader, with all the subjectivity that might entail: just look at the objective, observable features of the words and sentences that readers will encounter.

But do these features of language really constitute the *essence* of readability (such that writers should consider them paramount while they are writing their drafts)? Before we answer this question, let's consider what readability is.

Definitions

Readability has two common meanings, one applying to document design, the other to language. Readability as it is applied to document design is concerned with such matters as line length, leading, white space, font type and the like. Readability as it is applied to language is concerned with the *comprehensibility* or *understandability* of a piece of written text:

> "…the efficiency with which a text can be comprehended by a reader, as measured by reading time, amount recalled, questions answered, or some other quantifiable measure of a reader's ability to process a text..."[1]

Another quote:

> "*Readability* means *understandability*. The more readable a document is, the more easily it can be understood…"[2]

And one more:

> "[Readability is the] sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it an optimal speed, and find it interesting."[3]

In this paper, I am discussing readability as it pertains to language, not as it pertains to document design. And this is the variety of readability that the readability formulas purport to measure.

Why is readability important?

For a start, readability is important because it goes to the very heart of our professional ethics. We want the sick to understand the documentation that accompanies their medicines and the machinist to understand how to operate their machine and not injure themselves and others around them. It is difficult to argue otherwise without inverting the defining goals of our profession. So if formulas can help us assess the readability of our documentation— medical, mechanical or otherwise—then technical writers should embrace them as a core responsibility.

Secondly, the use of readability formulas is gathering popularity, and in places that matter:

> "Today, reading experts use the formulas as standards for readability. They are widely used in education, publishing, business, health care, the military, and industry. Courts [in the USA] accept their use in testimony."[4]

Organisations in the US have been successfully sued by plaintiffs claiming that they have been disadvantaged by an inability to understand certain public documents.[5] To protect themselves against such litigation, many bodies—commercial and

[1] Jack Selzer, "What constitutes a 'readable' technical style?" in PV Anderson, RJ Brockmann & CR Miller (eds), *New Essays in Scientific and Technical Communication: Research, theory and practice*, Baywood, New York, 1983, pp. 71–89. See p. 73.

[2] *Editing Technical Writing*, by Donald C. Samson Jr., Oxford University Press, New York, 1993, p. 58.

[3] E. Dale and J.S. Chall, "The concept of readability", quoted in William H. DuBay, *Smart Language: Readers, Readability, and the Grading of Text*, Impact Information, Costa Mesa, CA, 2007, p. 6.

[4] DuBay, *op.cit.*, p. 5.

[5] For example, Tampa General Hospital and the University of South Florida paid a US$3.8 million settlement to a group of people who claim that a consent form they signed exceeded their reading ability. Cited by DuBay, *ibid.*, p. 2.

government—now have specific guidelines on the minimum readability required of its public documents. And this concern is not limited to documentation authored in the US. My own company, writing documentation to accompany a locally produced product that was to be exported to the US, was asked to ensure that the documentation had a readability score that indicated that it could be fully understood by someone with only an eighth-grade education.

Further, recent changes to tort law in Europe mean that organisations can be held liable for faulty documentation (that is, documentation difficult to understand) just as they can for faults in the product that it accompanies.[6]

So if readability itself is important, going as it does to the very heart of our professional ethics, and if ensuring a minimum readability is increasingly becoming a legal requirement, then readability and its measurement should be of fundamental interest to technical writers. Moreover, if readability is now being widely measured on the basis of readability formulas (and by organisations with clout) we need to be sure that such formulas do indeed provide a reliable and valid measure of readability.

Readability and textual statistics

Numerous attempts have been made to base readability solely on observable features of sentences, without reference to the psychology or prior knowledge of the reader. In fact, the history of readability research is littered with more than 200 text-based readability formulas, each considering as important such features as average number of words per sentence, average number of syllables per word, the number of single-syllable words, the number of polysyllables, the number of words not on some predetermined list of so-called *easy* words, and the like.

For example, the Gunning Fog Index measures, in a sample of 100 words, the average number of words per sentence and the number of words of more than 2 syllables; and the Simple Measure Of Gobbledegook (SMOG) measures the number of words of more than 2 syllables in a sample of 30 words.

Flesch Reading-Ease Formula (FREF)

Probably the most influential of all the readability formulas—text-based or otherwise—is the Flesh reading-ease formula:

$$\text{Reading ease } (RE) = 206.835 - 84.6s - 1.015w$$

where s = the average number of syllables per word and w = the average number of words per sentence.[7]

In most cases, the value of RE will fall within the range 0–100: the higher the value, the more readable the text (or so the theory goes).[8] RE values less than 0 and greater than 100 are possible, but these are ignored by readability practitioners.

The FREF is behind the main readability statistic in Microsoft Word. It has also been tweaked for special uses (such as in the US Navy Readability Indexes), and it provides the raw input for another of the readability statistics that Microsoft Word generates: the Flesch–Kincaid Grade Level (which simply maps ranges of readability scores to particular levels of schooling in the US education system). Because of its

[6] Susan Burton, "A Worldwide Phenomenon", *Intercom*, Sept./Oct. 2007, p. 3.

[7] R. Flesch, "A New Readability Yardstick", *Journal of Applied Psychology*, vol. 32, 1948, issue. 3, pp. 221–233.

[8] "…on a scale between 0 (practically unreadable) and 100 (easy for any literate person)". Flesch, *ibid.*, p. 229

special influence, most of the comments in this paper will be directed at the FREF (although it should be clear that most comments will apply equally to any formula that derives a measure of readability from the properties of text alone).

Pros and cons at first glance

Sentence length

It is undeniable that very long sentences *are* difficult to digest. By the time you have reached the end a multi-clause sentence of, say, 40 or more words, you are often struggling to remember what you had read at the start of the sentence. So it is difficult not to argue that the longer the sentence, the less readable it is.

But there is a fallacy lurking here that we need to be wary of. From the fact that long sentences are difficult to read—and even longer ones even more difficult—it does not follow that maximum readability demands the *shortest* possible sentence. A string of two- or three-word sentences is hardly likely to be maximally readable.

> The cat shook. It sat. It licked. It hissed. Then it slept.

is more difficult to absorb than the following longer version:

> The cat shook and sat. It licked and hissed and then it slept.

The FREF would be more convincing if it set a minimum length for maximally readable sentences (which might or might not be feasible).

Syllable count

On the face of it, polysyllabic words are more challenging than monosyllabic words. Most of us scratch our head, or reach for a dictionary, when trying to read a medical book or an article on quantum physics (texts commonly full of polysyllabic jargon). So a measure that reduces readability as syllable count increases has *prima facie* plausibility.

But the relationship between readability and syllable count is superficial. It fails to acknowledge the influence that frequency of use has on a word's readability. There are numerous two-syllable words that are far more understandable than one-syllable words simply because we use them more often and have used them since childhood. For example:

> Mummy always washes the dishes after breakfast every morning

is obviously more readable than the equally long sentence:

> Electrons jump a level when hit by a photon

even though the former has a higher syllable count (18 as opposed to 13).

Indeed, a sentence can be short and with only monosyllabic words and yet be entirely obscure to the reader. For example:

> The work done was five ergs

To many this sentence is meaningless. Only the scientifically-minded is likely to know that *erg* is a measure of *work* (in much the same way as *litre* is a measure of *volume*). But *The work done was five ergs* has the same number of words and syllables as the eminently readable *The cat sat on the mat*. This suggests that readability is intimately tied to conceptual familiarity (which is hardly a breathtaking discovery).

Textual statistics—such as word length and syllable count—seem barely relevant at all.

More cons

Textual statistics fail to capture poor grammar and nonsense

Poor grammar obviously affects readability. For example:

> Sat the mat the cat on

is a grammatically poor sentence, and yet it has the same number of words and syllables as the much more readable *The cat sat on the mat*.

In fact, the FREF necessarily gives the same readability score however you re-arrange the words in a sentence: with grammar and sense in mind or otherwise. For example, *The on mat the sat cat* has the same number of words and syllables as *The cat sat on the mat* and thus should be equally readable on Flesch's view of readability.

Textual statistics fail to capture poor or unusual punctuation

Punctuation obviously affects readability. For example, if you write

> Have a good holiday.

when you should have written:

> Have a good holiday?

then you are very likely going to confuse or mislead the reader. But the FREF gives these sentences the same score.

The FREF also gives the following sentences the same score:

> Suitable for small business operators
>
> Suitable for small-business operators

Omitting the hyphen in compound adjectives can create ambiguity if the context cannot clarify the intended meaning. (Is the thing in question suitable for business operators of small stature? Or those who run small businesses?) Even if the reader could decipher the intended meaning from the context, they would have done so only with unnecessary effort: an indication of non-maximal readability.

Textual statistics overlook the importance of typographical cueing

The way text is styled or cued can significantly affect readability, and yet style cues are ignored when all that is being considered is sentence length and syllable count. Here are two examples that show the importance of typographical cueing:

> You can learn more about these events by reading How high did he fly and other similar books.

This sentence is ambiguous because the title of the publication the reader is being referred to is not clear. Is it *How high did he fly* or *How high did he fly and other similar books*? But the FREF would give this sentence the same score as the following version:

> You can learn more about these events by reading *How high did he fly* and other similar books.

Typographical cueing is prevalent in technical writing. It is standard practice to use cueing to distinguish *referential* text (being text that mirrors what the reader will see on screen or on a product) from *explanatory* text (being text that describes a process, procedure, concept or the like). Note that the following two sentences, identical in all respects bar the typographical cues, would score the same on the FREF and yet they are obviously not of identical readability:

> Select GST and FBT, deselect invoices only and press Enter
>
> Select **GST** and **FBT**, deselect **invoices only** and press Enter

A reader of the first sentence wouldn't immediately know that there are two options to select, **GST** and **FBT**, rather than one option: **GST and FBT**. And they wouldn't immediately know that there is just one option to deselect rather than potentially many individual invoices.

Textual statistics fail to detect vagueness and non sequiturs

Consider the following passage, taken from documentation that accompanies bookkeeping software:

> Sales are usually allocated to an income account. You should not choose a Trade Debtors account for ordinary sales.

The vagueness and imprecision is all too clear, so to speak. *Usually*, *ordinary*: what do these terms mean? How is the reader to determine from those words whether a particular sale in question should be allocated to an income account, and whether the trade debtors account needs to involved as well?

Consider how that passage could be tightened up:

> Always choose an income account when you are allocating a sale. A trade debtors account should never be used for any sale.

This version has the same number of syllables as the first, but has more words. So it, though obviously less vague than the first, will have a lower FRE score.

Consider now non sequiturs: sentences that begin down one path and end down another. For example:

> Unlike many other areas of business where errors can be adjusted at a later date, employees immediately notice errors in their paycheques.

Whatever FRE score this sentence gets, its meaning is impossible to determine. We can at best guess at it. Indeed, it is easy to concoct non sequiturs that score the highest possible FRE score, such as:

> When the cat sat on the mat, the square root of nine is three.

Any formula that attributes maximum readability to that sentence is undoubtedly faulty.

Textual statistics fail to detect transition words

Transition words are binary words: words having two equal or near-equal primary meanings. They are mostly words part-way through a transition from one primary meaning to another. An example is *disinterested*. Its strongest meaning not so long ago was *impartial* and *unbiased*; nowadays we see it used just as often to mean *bored* and *uninterested*. Other transition words are *inhibit* (*impede* or *block*?), *alternate* (*every second* or *alternative*?), *viable* (*able to live independently* or *workable*?), *fortuitous* (*fortunate* or *accidental*?) and *fulsome* (*hearty* or *gross*?).

What is a transition word at one time might not be a transition word at another time. But when a word is in transition, its use is likely to be acutely ambiguous (and chronically ambiguous if the context is of no help). Obviously, a readability formula that concentrates on textual statistics and ignores semantics will fail to detect the ambiguity posed by a transition word.

Textual statistics fail to detect contradiction and inconsistency

Any analysis of readability that considers textual features only and not meaning cannot detect outright contradiction and undisciplined use of vocabulary, both of which obviously detract from maximal readability.

> Roma is the capital of Italy…But since Turin, the capital of Italy, is…
>
> Roma is the capital of Italy…But since Roma, the capital of Italy, is…

Though both passages have identical FRE scores, the former is more likely to have you scratching your head than the latter.

A related issue is terminological inconsistency. Technical writers are schooled to shun elegant variation in favour of the *principle of single and distinct denotation*: one term, one concept; one concept, one term. The reason is straightforward: if you start using term B to refer to something you've just introduced with term A, you run the risk that your reader will think you are talking about two things rather than one. And this obviously impedes readability.

> The options available on the menu bar…From the menu strip choose…
>
> The options available on the menu bar…From the menu bar choose…

The latter is likely to be more comprehensible (that is, more readable) than the former, even though both share the same readability score.

Textual statistics fail to acknowledge syllabic variation

How many syllables comprise a word is not always clear cut. The number depends on how the word is *pronounced*, not how it is written, and this, of course, can vary between the different English languages. For US speakers, *temporarily* has five syllables and *medieval* two; but for many others the syllable count is four and three respectively. Here are some more words that have acceptable variations in the number of syllables pronounced (mostly dependent on region):

- comparable
- secretary
- laboratory
- extraordinary
- veterinary
- medicine
- gaseous
- library
- glacial

Any formula that emphasises the importance of syllabic count but considers only *written* language rather than spoken must make assumptions about how words are pronounced. Such assumptions obviously introduce further imprecision into the calculation.

Conclusion

Sentence length and syllable count (and any variation on these) cannot *define* or *cause* readability. Readability is a much more complex concept. At the very least, a measure of readability must take into account the reader as well as what is read (such as their domain knowledge).

Thus the Flesch reading-ease score and its cognates are a poor measure—indeed, are *no* measure—of readability. The prestige such formulas are given by their presence in popular word processing software should not mislead us into thinking that they are currency of much worth.

The second part of this paper will explore whether textual statistics might still be useful as an *indicator* or *predictor* of readability. It will also examine the techniques used to validate readability formulas. The conclusion will be that quantitative readability formulas of the Flesch variety are fundamentally flawed. Technical writers should therefore resist the gathering pressure to judge the readability of their writing on the basis of such formulas.

Part 2: Validation and its pitfalls

This is the second part of a two-part paper exploring issues in readability. Part one (above) considered the most popular approach to measuring readability, epitomised by the Flesch reading-ease formula (the formula behind the readability scores one finds in Microsoft Word). The conclusion was that the variables that get fed into the Flesch formula are poor measures of readability and thus the formula itself is fundamentally flawed. This part of the paper argues that the techniques commonly used to validate readability formulas lead to skewed results. Furthermore, redressing the methodological weakness in these techniques necessarily produces an even poorer correlation between independently measured readability scores and Flesch scores. This reinforces the conclusion that such formulas are best discarded.

The first part of this paper considered text-based readability formulas, the most influential of which is the Flesch reading-ease formula (FREF), the formula that gives the readability scores in Microsoft Word. The arguments in that part should dispel the notion that sentence length and syllable count—the only variables considered by the FREF—*define* or *cause* readability. But could the score on the FREF still be a good *indicator* or *predictor* of readability? An analogy: we do not define the concept of temperature in terms of the height of mercury in a thermometer. Rather, temperature is defined as the degree of hotness or coldness (and sometimes in terms of the average kinetic energy of the particles in a body). Nonetheless, the height of mercury in a thermometer has been found to be a very reliable indicator of temperature. Indeed, the correlation between temperature and mercury level is as strong as can be. So, might textual statistics— sentence length and syllable count—be a good indicator of readability even though readability cannot be defined in terms of them? If so, a writer might reasonably use the FREF to evaluate the relative readability of various drafts of a document even though they might not write with textual statistics as their overriding guide.

This is indeed the claim of contemporary proponents of the textual analysis of readability. There is now widespread admission that many factors—not just sentence length and syllable count—contribute to readability and that writers should acknowledge those features when they are writing. But proponents argue that textual statistics are still the *best* indicator of readability, that if you calculate the correlation between other features of readability and comprehensibility, the value you get is no better than if you calculated the correlation between textual variables and comprehensibility. In other words, an analysis of sentence length and syllable count is just as good as, but far simpler than, more complicated analyses of language.

> 'Critics of the formulas…rightly claim that the formulas use only "surface features" of text and ignore other features like content and organization. The research shows, however, that these surface features—the readability variables—with all their limitations have remained the best predictors of text difficulty as measured by comprehension tests.'[9]

But *best* does not imply *good*. At one time, the *best* way we had of estimating the number of stars in the universe was to look at the night sky and count them. But that,

[9] William H. DuBay, *Smart Language: Readers, Readability, and the Grading of Text*, Impact Information, Costa Mesa, CA, 2007, p. 79

obviously, was not a very *good* technique. So how might we determine that the FREF gives a *good* way of indicating or predicting readability?

Any creditable test must have two important attributes: *reliability* and *validity*. A test is reliable if you get the same or similar result each time you apply it to the same subject. Obviously, the FREF is a very reliable test. Every time it is applied to the very same piece if text, it will give the very same score.

What of validity? A test is considered valid if it actually measures what it purports to measure. A test that assigned intelligence to a person solely on the basis of university results would obviously be less valid (if valid at all) than one that assigned intelligence on the basis of numerical, verbal and logical skills.

The validity of a test is usually determined by comparing the results it gives against those given by an independent test widely accepted as being a good measure of whatever is being tested. A strong correlation between the results is considered to give the test in question validity.

So is the observed correlation between readability *as independently measured by reputable tests* and readability *as measured by the FREF* strong enough to warrant the use of the FREF as a valid indicator or predictor of readability?

How correlations are derived

In a typical readability research project, participants are given a number of texts and their comprehension of these texts is assessed. Two methods are widely used:

- Participants are given a number of questions about each text, and the number of correct answers supplied is used to assign a level of difficulty to a text.
- Participants undergo a cloze test whereby they fill in words that have been deliberately omitted from the texts. (In a typical cloze test, every fifth word is omitted.) The number of correct words added to a text is used to assign a level of difficulty to that text.

Once each text in a bank of texts has been graded, the relevant textual statistics in each are calculated and fed into a readability formula. Researchers then determine the correlation between the level of difficulty of a text (as determined by a comprehension test or cloze test) and the score given by the readability formula.

The observable correlations

Correlation is simply a measure of how one variable changes when another variable changes. The most widely used formula for determining what is called the correlation coefficient, r, produces a value between −1 and +1. If $r = +1$, when the value of one variable is high, the value of the other variable is also high; and when it is low, the other variable is correspondingly low. If $r = -1$, when the value of one variable is high, the value of the other variable is low, and vice versa. Values between −1 and +1 indicate a less than perfect correlation, that is, the relationship between the two variables is loose and it is impossible always to infer with confidence what will happen to one variable when the other variable changes. And if $r = 0$, the two variables are completely independent: when the value of one is high, the value of the other is sometimes high and sometimes low, and equally so.

So how do the readability values generated by the Flesch reading-ease formula (FREF) correlate with the levels of difficulty determined by comprehension and cloze tests? In his paper introducing the FREF, Flesch reported a correlation coefficient of

0.7047.[10] Is such a value strong enough to validate the FREF? A number of theorists have argued that it is not.[11] It might be true that such a value tells us that an inference from a high score on the FREF to high comprehensibility has a greater probability of being correct than incorrect; but it is also true that that inference will sometimes be incorrect. In other words, there will be cases where a high FREF score occurs together with low comprehensibility (as the many counter-examples given in the first part of this paper prove). In other words, you could never tell from an FREF score alone just how readable a text is. You could only guess.

How serious is this for proponents of readability formulas? Suppose we were to learn that there is only a correlation of 0.7 between the actual number of words in a document and the number of words given by Microsoft Word's word-count feature. Would we continue using that feature? I suspect we wouldn't. Perhaps, then, we should view Microsoft Word's Flesch-based readability scores in a similar light.

This is, however, a little too harsh. Supporters of readability formulas admit that assessing readability is not an exact science (whereas counting words is). And we should not expect data that can only be analysed statistically to be susceptible to the same precision possible with purely numerical analyses. That might be so, but we can still ask just how useful a correlation coefficient of 0.7 really is.

Directional volatility (aka *normalised Kendall tau distance*)

To answer this question, let's first consider another statistic, one I'll call *directional volatility*. This statistic indicates how likely it is that you will be wrong in predicting that a change in one variable (up or down) corresponds to a similar change (up or down) in another variable.[12] This is especially relevant to writers, because if the FREF is to be useful we need to be confident that we have actually improved the readability of a document if our latest draft scores higher on the FREF than a previous draft. Directional volatility provides us with a measure of that confidence.

Table 1: Sample scores

Text	FREF score	Test score
1	30	30
2	35	25
3	40	40
4	45	35
5	50	50
6	55	45
7	60	65
8	65	50
9	70	50
10	75	45

Consider the data in Table 1. This table shows, for a set of 10 texts, the score on a comprehension or cloze test and the corresponding score on the FREF. To simplify calculating the directional volatility, the pairs of scores have been arranged so that one set of scores is in numerical order. (In this case it is the FREF scores that are so arranged, but it doesn't matter which.) Start by considering each possible combination of two FREF values: (30, 55), (60, 70) etc. (In a sample of 10 scores, there are 45 such combinations. You can imagine these combinations as representing FREF scores on draft 1 and on draft 2 of a document.) For each combination, note whether the difference between the second score and first score is positive or negative. (Because we arranged the Flesch scores in ascending numerical order, we know without observing the scores that the difference will always be positive.) Now match each combination of FREF scores with the

[10] R. Flesch, "A New Readability Yardstick", *Journal of Applied Psychology*, vol. 32, 1948, issue 3, p. 225.

[11] See, for example, J. S. Chall, *Readability: An Appraisal of Research and Application*, Ohio State University Press, Columbus, 1958.

[12] This statistic is also called the *normalised Kendall tau distance*. I'll use the simpler, more descriptive and less alienating term *directional volatility*.

corresponding pair of comprehension test scores: (30, 55) corresponds to (30, 45); (60, 70) corresponds to (65, 50); and so on. Note now how many positive differences between pairs of FREF scores correspond with negative differences between corresponding pairs of test scores. In other words, how often do FREF scores go in one direction (up or down) while comprehension scores go in the opposite direction.

For variables that are positively correlated (as in Table 1), the number of directional mismatches (13 in this case) as a percentage of the total number of possible combinations (45) is the directional volatility of the sample.[13] In so far as the sample is representative, this is a measure of how likely you are to err in predicting that a rise in FREF score corresponds to a rise in actual comprehension as determined by the independent tests. And for the data in Table 1, the directional volatility is 29%. That is, the probability that an observed increase in Flesch scores does *not* correspond to an increase in comprehension is 0.29.

Directional volatility and correlation

By indicating how confident you can be that you have actually improved the comprehensibility of the document by increasing the Flesch score, directional volatility is an important statistic in assessing the validity of the FREF. For the FREF to be valid it must have a low directional volatility. But, as far as I can determine from the literature, no Flesch validation study has considered the directional volatility of sampled scores. Instead we are simply offered the correlation coefficient.

But here's the rub: a correlation coefficient does not correlate well with directional volatility. In Table 1 above and Table 2 below, we have three sets of data in which the correlation coefficient is 0.7 (the same value that Flesch reported)[14]. But the directional volatility varies widely: 29% for the data in Table 1, and 15.5% and 44.4% respectively for the data in Table 2.

Table 2: Identical correlation coefficients (0.7) but widely varying directional volatility (15.5% and 44.4% respectively)

Data set 1			Data set 2		
Text	**FREF score**	**Test score**	**Text**	**FREF score**	**Test score**
1	30	30	1	30	70
2	35	31	2	35	66
3	40	32	3	40	62
4	45	33	4	45	58
5	50	34	5	50	54
6	55	35	6	55	95
7	60	36	7	60	91
8	65	37	8	65	90
9	70	38	9	70	89
10	75	32	10	75	88

So it is possible, with a correlation coefficient of 0.7, for the probability that an observed increase in Flesch scores does not correspond to an increase in comprehension to be as high as 0.444. In that case you would be wrong nearly half the time in predicting that draft 2 of a document was more comprehensible than

[13] In negative correlation, directional volatility would be the complement of the discrepancies (45 − 13) as a percentage of the possible combinations.

[14] Correlation coefficients can be calculated easily with the CORREL function in Microsoft Excel.

draft 1 on the basis solely of an improvement in FREF score. That should hardly inspire confidence in a correlation coefficient of just 0.7.

Of course, these three data sets are fictitious, but they do prove that correlation alone is not a sufficient indicator of predictive reliability. Directional volatility does a better job at that. But Flesch provided us only with the correlation coefficient and no other statistic that might indicate predictive reliability. Subsequent commentators seem also to have overlooked directional volatility. But until the directional volatility of actual test data is calculated, the usefulness of the FREF as a predictor of readability remains in doubt. And perhaps in recognition of the fact that a correlation coefficient as low as 0.7 is compatible with significantly volatile data, current proponents of readability formulas are beginning to temper their enthusiasm for them, now admitting that they can only be rough guides:

> '[Readability formulas]…are not perfect predictors. They provide probability statements or, rather, rough estimates…of text difficulty.'[15]

It would be a welcome clarification were Microsoft to include a disclaimer to this effect on the **Readability Statistics** dialog in Microsoft Word.

Subsequent research

When Flesch introduced the FREF sixty years ago, he reported a correlation coefficient of 0.7047. While some follow-up studies found a similar result, others reported results much lower: 0.64 and 0.5.[16] And in a 1998 study of tourism texts, Woods *et al.* found a correlation between text difficulty as assessed by cloze tests and text difficulty as determined by the FREF of just 0.13.[17] *This is close to indicting that there is no correlation at all*. The authors likewise found no high correlation between cloze scores and scores on a number of other popular text-based readability formulas (FRY, SMOG and FORCAST). They also noted that the four readability formulas gave widely inconsistent results, with some formulas scoring a text as considerably difficult that others scored as relatively easy. They concluded that:

> 'The readability tests examined in the present study gave very inconsistent results and none of the tests did a very good job at predicting readers' responses [to the cloze tests]…The results do not support the use of the readability tests analysed in this study (FRY, SMOG, FORCAST or Flesch's Ease of Reading test).' [18]

Limitations of the method

But whatever correlation coefficient is obtained, the validation methodology readability researchers adopt—using comprehension and cloze tests—necessarily *overstates* the actual coefficient. The reason? Unavoidable sampling errors.

Any study of scientific merit does not arbitrarily limit its data sampling. If you want to establish that all bodies fall with the same rate of acceleration, you do not limit your experiments to, say, metal objects. But the problem with relying only on the results of comprehension or cloze tests to determine the correlation between reading difficulty and readability statistics is that the data sample is *necessarily* limited. This is

[15] DuBay, o*p. cit.* p. 110.

[16] Jack Selzer, "What constitutes a 'readable' technical style?" in PV Anderson, RJ Brockmann & CR Miller (eds), *New Essays in Scientific and Technical Communication: Research, theory and practice*, Baywood, New York, 1983, p. 73. p. 75. Others have reported coefficients of 0.64 and 0.7 (as in DuBay, *op. cit*., p. 57).

[17] B. Woods, G. Moscardo & T Greenwood, "A Critical Review of Readability and Comprehensibility Tests", *The Journal of Tourism Studies*, vol. 9, no. 2, December 1998, pp. 49–61.

[18] *ibid*., p. 58

because those who devise a comprehension or cloze test can only use texts that are fully comprehensible (at least to them); otherwise they would not be able to determine if an answer provided by a testee is the correct answer. Indeed, they would find it difficult, if not impossible, to devise sensible questions in the first place.

But to avoid skewing the results, the data sampled must be extended to include texts that are *in*comprehensible. Such texts are not difficult to find; but they can also be concocted. One could write passages that are hopelessly ambiguous, or hopelessly vague; but it is much easier simply to concoct nonsense strings, such as *The door is in love* and *Honesty is the largest integer less bilious than the smartest flooring wart*. (There's nothing devious about concocting a sample of incomprehensible texts. Even the comprehensible texts used in readability research are concocted to ensure that a wide range of average comprehension scores is obtained.)

Samples of nonsense strings don't need to be tested for comprehension before they are subjected to the FREF. We know, by definition, that they have a comprehension value of zero (for they are nonsense). Now if the FREF is a valid indicator of readability (and thus of comprehension) then nonsense strings should get an FREF score of zero. (Recall that FREF scores range from 0 to 100, with zero indicating that the text is incomprehensible and 100 that it is fully comprehensible to any literate person.)

However, it should be clear that there can be no correlation whatsoever between the comprehension scores of nonsense strings and corresponding FREF scores. This is because we can concoct any number of nonsense strings with few words and few syllables (which would score high on the FREF: the first example above scores 100) and any number of nonsense strings with many words and many syllables (which would score low on the FREF: the second example above scores 46.6). And if we are conducting the experiment scientifically, we do need to include all types of nonsense strings: short, long, monosyllabic and polysyllabic. Thus a zero actual comprehension score can be matched to any value in the 0–100 range of FREF scores, *and this indicates a correlation coefficient of zero*.

So, for maximally comprehensible texts, the best correlation coefficient that testers can find between comprehension scores and FREF scores is 0.7. And the only correlation coefficient possible for incomprehensible texts is zero. It follows that if correlation testing were to include samples all types of texts—comprehensible and incomprehensible— the real correlation coefficient must be even less than the already unconvincing 0.7 that Flesch reported (and even direr than the insignificant 0.13 that Woods and her co-researchers found).

Are the formulas superfluous?

Recall that studies to establish a correlation between scores on comprehension tests and scores on the FREF use texts that are maximally comprehensible to the testers. If they didn't, the answers testees give would have no value in determining a text's degree of difficulty. But how will those devising a test know that a candidate text is fully comprehensible? They can't subject it to a comprehension or cloze test to assess whether it is suitable to be subject to a comprehension or cloze test. Nor can they subject it to a readability formula, since the validity of the formula is the very thing they are trying to prove. So for testers to determine whether a text is fully comprehensible there must be some other criterion available for them to use, that is, a criterion other than a score on a comprehension, cloze or Flesch test.

So the methodology boils down to this: you start with a reputable criterion or test of comprehensibility (which, on our earlier definitions, would be a test of readability) to select fully comprehensible texts for a comprehension or cloze test. You then use the results from the comprehension or cloze test to validate a readability formula, finding that the best correlation coefficient you can find is about 0.7 (which is unavoidably over-stated). That is, you start with a criterion that must give a good measure of readability and use it to validate a formula that can at best give just a rough estimate of readability. *Wouldn't it be so much better just to use the initial criterion as our test of readability and forget about readability formulas altogether?*

Paradoxically, some proponents of readability formulas come close to adopting this same view. For example, after a long and favourable consideration of many formulas, the most that DuBay can say is:

> 'When adjusting a text to the reading level of an audience, using a formula gets you started, but there is still a way to go. You have to bring all the methods of good writing to bear.'[19]

But if we still have to bring all the methods of good writing to bear—that is, to concentrate on all those features of text that are essential for maximal readability—why don't we just concentrate on those features and forget about the additional task of applying the FREF?

DuBay's quite sensible advice would not be welcomed by those who argue that for workaday uses we need a simple proxy measure of readability, such as a text-based formula, because directly assessing readability is just too difficult and time-consuming. I suspect that those who take this line may have gone through schooling during that Dark Age of English language teaching: the latter quarter of the twentieth century. Assessing readability might be a multi-faceted task, but it is not especially difficult. Any reputable language manual is a good start. Indeed, many of the examples given in the first part of this paper to discredit the conceptual linking of readability and textual statistics can guide us here, enabling us to distil some of the general features of readability.

Those examples suggest that three of the most important attributes of readability are familiarity, clarity and consistency, not one of which is taken into account by text-based readability formulas. Respect for the reader, economy of explanation and typographical cueing also contribute to readability and are ignored by the formulas. Devising a text-based, reader-independent algorithm that will assign an objective value to each of these attributes is highly unlikely, and probably impossible. (For a start, familiarity is as much a function of the reader as of the text.) It may be time, then, to dispose of text-based readability formulas on the scrap heap of over-zealous quantification.

In conclusion

The purpose of this paper has been to counsel caution in the gathering swell to make readability a necessary and overriding concern of technical writers. Readability *is* a fundamental goal, and moves to make it a legally binding feature of product documentation (discussed in the first part of this paper) are to be commended. But assessing readability requires a studied appreciation of language and careful analysis. Allowing oneself to be seduced by the simplicity of text-based formulas of the type we see in Microsoft Word is no shortcut. I hope the arguments set out in

[19] DuBay, *op. cit.*, p. 112. This seems to contradict DuBay's earlier claim that when you consider all the features of a text, textual statistics are still the best predictor of readability. See page 1 above.

both parts of this paper will encourage you to challenge those clients and employers who insist that you submit your documentation to the bumkin calculus of Flesch-type formulas.

Geoffrey Marnell teaches technical writing and editing in the English Department at the University of Melbourne. He is also the founder and managing director of Abelard Consulting Pty Ltd, a documentation consultancy providing technical writing services, technical writer placement services, and training in technical and scientific writing.