Scientometrics (2016) 107:1523–1526
DOI 10.1007/s11192-016-1920-7

Is time up for the Flesch measure of reading ease?

James Hartley[1]

Received: 10 February 2016 / Published online: 21 March 2016

Abstract The Flesch Reading Ease measure is widely used to measure the difficulty of text in various disciplines, including Scientometrics. This letter/paper argues that the measure is now outdated, used inappropriately, and unreliable.

Keyword Flesch readability · Text difficulty

Introduction

The Flesch Reading Ease measure was first published in 1943 (Flesch 1948). Since that date it has been revised several times and it still serves (in various computer-based forms) as one of the most popular measures used today for measuring readability. In the last 3 years, for example, there were at least three separate studies that evaluated the Flesch relative to other readability measures, and each of them concluded that the Flesch was the best one to use (Barbic et al. 2015; Didegah and Thelwall 2013; Zhou et al. 2016). In addition there were numerous other publications by authors—including me—relying on the Flesch to provide their data (e.g., Hartley 2015; Paudyal et al. 2015; van Wesel et al. 2014).

Why then do I now want to call time on the Flesch? Well, the reasons are simple:

1. The Flesch Reading Ease formula is crude and out of date, and it is often used for purposes that it was not intended for.
2. The Flesch does not assess the words in the text in context: indeed, by counting features such as word lengths, sentence lengths and passives, it does not measure the meaning of words at all.

✉ James Hartley
j.hartley@keele.ac.uk

[1] School of Psychology, Keele University, Staffordshire, UK

3. Different (computer-based) versions of the Flesch produce different scores: in a word the results are unreliable.

Let me discuss each of these points in more detail and then reach some conclusions.

The Reading Ease formula is crude and out of date

Probably the most detailed account of the development of the Flesch scale has been provided by Klare (1963). Klare describes several earlier measures of readability before turning to the Flesch and beyond. In all Klare describes 31 such formulae, developed between 1923 and 1959, and there must now be many more. The initial versions of the Flesch were developed in 1943, with two revised versions appearing in 1948. One of these, the Flesch Reading Ease Scale, has become the most frequently used measure of readability to date. In its original form the reading ease score (R.E.) was calculated (by hand) using the new formula. The procedure was as follows:

Systematically select 100 word samples from the material to be rated

Determine the number of syllables per 100 words (wl)

Determine the average number of words per sentence (sl)

Apply in the following reading ease equation:

$$\text{R.E.} = 206.835 - 0.846wl - 1.015sl$$

The resultant score from 0–100 was the entered into a table (see for example, Table 1) to determine the reading ease of the piece. Many authors (including me) have recommended using the Flesch for comparing the readability of both the parts (abstract, introduction, method, results and discussion) and the whole of journal articles (e.g., Hartley 2015).

The Flesch formula does not take into account the meanings of words, only their lengths

The Flesch formula is an example of my second point. This is that the words themselves (and their actual meanings) are not considered as such. Agreed, text that has long sentences and long words within it is likely to be more difficult than text with short words and sentences. But a lot of this depends upon the reader. Texts in nuclear physics might be easy

Table 1 Flesch scores and how they are interpreted

Flesch R.E. score	Reading age	Difficulty	Example
90–100	10–11 years	Very easy	Comics
80–90	11–12 years	Easy	Pulp fiction
70–80	12–13 years	Fairly easy	Popular novels
60–70	14–15 years	Average	Tabloid newspapers
50–60	16–17 years	Fairly difficult	Introductory textbooks
30–50	18–20 years	Difficult	Undergraduates' essays
0–30	Graduate	Very difficult	Academic prose

for nuclear physicists but not for the rest of us. Furthermore, the Flesch measure has been applied to texts in other languages with only minor changes (e.g., see Menoni et al. 2010).

This problem (the lack of meaning) has, of course, been recognised in other readability formulae where the actual vocabulary is taken into account. Probably the most well-known early measure of this kind used the Dale–Chall (Dale and Chall 1948) formula—where those words occurring outside a list of 3000 common words were also counted. But there do not appear to be similar discipline-based formulae for use with specific disciplines.

Different computer-based measures of the Flesch produce different findings

This leads to my last point. Different computer-based measures of the Flesch (and other readability measures) give different scores depending upon how the formulae have been computerised. Such differences can be quite large (e.g., see Harris 1996; Mailloux et al. 1995; Sydes and Hartley 1997; Zhou et al. 2016).

In the early days (when the Flesch was created) some of the rules for entering the text depended on the person doing it, and on the then mechanics of the day. Thus, for example, researchers were advised to 'systematically select' samples of 100 words, to count the number of syllables by hand (not always an easy task) and to apply the formula by hand. Today Flesch scores are computed on the basis of mechanically counting items and standardising the data to produce a score that can range from 0 to 100. However, different computer programs apparently treat similar items differently, so that Flesch scores vary slightly depending upon which computer programs are being used. Thus, for example, some programs might count the letters in a word, whereas some might count the syllables. Similarly, some might treat 'don't' and 'first-grade' as single words, whereas others might treat them as two. Again, some programs might treat semi-colons and colons as full stops and some might not. Differences such as these originate from the days when typewriters were used to input the text and items were separated by using the space-bar.

Some conclusions: What should we do now?

Perhaps it is now time to abandon the notion of one measure and one computer program being suitable for all purposes. It should not be impossible today to have different computer-based measures for different tasks. For example, we could have:

- Discipline related measures
- Ability related measures
- Age related measures
- And combinations of these.

Such programs would have greater validity than the Flesch but they would still fail to take into account the effects of other variables that affect readability. These other variables include page-size and orientation, line-spacing, type-sizes and typefaces, the presence (or absence) of headings, tables, diagrams, graphs and possibly video clips (Badarudeen and Sabharwal 2010; Hartley 1994). Only tests such as the 'cloze' test (Taylor 1953)—where you count words with more than three syllables—or the 'SMOG' test (simple measure of gobbledygook, McLaughlin 1969),—where you respond to the texts in question by supplying every omitted—say—5th—word) can take variables such as these into account. But, of course, asking readers to do such things takes time. And we wouldn't want to do that, would we?

Acknowledgments I am extremely grateful to Guillaume Cabanac for submitting this paper to *Scientometrics* for me.

References

Badarudeen, S., & Sabharwal, S. (2010). Assessing readability of patient information materials. *Clinical Orthopaedics and Related Research, 468*, 2572–2580.

Barbic, S. P., et al. (2015). Readability assessment of psychiatry journals. *European Science Editing, 41*(1), 3–10.

Dale, E., & Chall, J. S. (1948). The concept of readability. *Elementary English, 26*, 19–26.

Didegah, F., & Thelwall, M. (2013). Which factors help authors produce the highest impact research? Collaboration, journal and document properties? *Journal of Informetrics, 7*(4), 861–873.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233.

Harris, R. (1996). Variation among style checkers in sentence measurement. *Text Technology, 6*(2), 80–90.

Hartley, J. (1994). *Designing instructional text* (3rd ed.). London: Kogan Page.

Hartley, J. (2015). Making writing readable. *The Psychologist, 28*(4), 254–255.

Klare, G. R. (1963). *The measurement of readability*. Ames: Iowa State University Press.

Mailloux, S. L., Johnson, M. E., Fisher, D. G., & Pettibone, T. J. (1995). How reliable is computerized assessment of readability? *Computers in Nursing, 13*(5), 221–225.

McLaughlin, G. (1969). SMOG grading—A new readability formula. *Journal of Reading, 22*, 639–646.

Menoni, V., Lucas, N., Leforestier, J. F., Dimet, J., Doz, F., Chatellier, G., et al. (2010). The readability of information and consent forms in clinical research in France. *PLoS One, 5*, e1056.

Paudyal, P., Capel-Williams, G. M., Griffiths, E., Theadom, A., Frew, A. J., & Smith, H. E. (2015). Readability, presentation and quality of allergy-related patient information: A cross sectional and longitudinal study. *Journal of Allergy and Therapy, 6*, 213. doi:10.4172/2156121.1000213.

Sydes, M., & Hartley, J. (1997). A thorn in the Flesch: Observations on the unreliability of computer-based readability formulae. *British Journal of Educational Technology, 28*(2), 143–145.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30*, 415–433.

van Wesel, M., Wyatt, S., & ten Haaf, J. (2014). What a difference a colon makes: How superficial factors influence subsequent citation. *Scientometrics, 98*, 1601–1605.

Zhou, S., Green, P. A. & Jeong, H. (2016, in preparation). How consistent are the best-known readability equations in estimating the readability of design standards? (Copies available from pagreen@umich.edu).