The readability of original articles in medical journals

Short Title: Readability of medical journals

John C. Hall

School of Surgery and Pathology, The University of Western Australia, Perth WA

CORRESPONDING AUTHOR:

Professor John C Hall,
University Department of Surgery,
Royal Perth Hospital,
Wellington Street,
Perth WA 6000, AUSTRALIA

Tele: 61-8 9224 0228
Fax: 61-8 9224 0204
E-mail: jchall@cyllene.uwa.edu.au
ABSTRACT

**Background:** Readability indices have been developed based on sentence length and the use of long words. One such measure is the Reading Ease Scale developed by Rudolf Flesch. Texts that are easy to read have high scores: texts with scores below 30 are similar to legal contracts. This study uses Flesch scores to evaluate the readability of surgical journals.

**Methods:** Flesch scores were calculated for papers published in the *Archives of Surgery*, the *British Journal of Surgery*, and the *ANZ Journal of Surgery*. The first 30 original articles published in each journal in 2005 were selected for study. Excluded from study were editorials, reviews, and case-reports.

**Results:** The overall median score was 15.1 (0.0 – 29.1). The median scores for the journals were 12.4 (*Archives of Surgery*), 14.4 (*British Journal of Surgery*), and 18.6 (*ANZ Journal of Surgery*). There was only a minor link between Flesch scores and the use of surgical terms.

**Conclusions:** Original articles published in surgical journals contain too many long sentences and complex words. Readability indices are useful tools because they promote the use of simple English. It is realistic for authors to aim for Flesch scores above 30 when creating manuscripts.

**Key Words:** Surgery, publishing, language test, literature, peer review
INTRODUCTION

In the 1930's psychologists began to study how written information is processed by the brain. They concluded that long sentences are bad - it is difficult to retain the meaning of sentences that contain more than 20 words - and readers find it easier to grasp simple words. Only 36 of the 1,000 most often used words have more than two syllables.

Readability indices have been developed based on sentence length and the number of ‘hard words’. Editors are very interested in such indices because good readability fosters popularity, which increases publication volume and profit. The most commonly used measures are the Fog Index devised by Gunning and Flesch’s Reading Ease Scale.

Both of the readability indices are easy to calculate (Table 1). However, the Fog Index contains discretionary judgements about ‘hard words’ that might lead to variations between markers. On the other hand, the Flesch score involves clear-cut calculations and it has been adopted for use in word processing programs. This means that it is possible to obtain Flesch scores that are both precise and objective.

This study uses Flesch scores to evaluate the readability of original articles published in surgical journals. Flesch score range between 0 and 100. The higher the score, the easier it is to read the text. A secondary objective was to scrutinize the relationship between the use of long technical words and the readability score.
METHODS

The journals studied were the *Archives of Surgery*, the *British Journal of Surgery*, and the *ANZ Journal of Surgery*. Digital copies were obtained of the first 30 original papers published in each journal from the start of 2005. Each paper contained original data and a structured abstract. Excluded from study were editorials, reviews, and case-reports.

The ‘Introduction’ and ‘Discussion’ sections of each paper were then isolated for review. They were submitted to two forms of editing. First, minor errors resulting from the downloading of HTML files were removed. Second, the reference numbers in superscript were deleted – a pilot study of 10 papers selected at random found that removal of the references increased the median Flesch score by 2.5 (0.2 – 8.5). The Flesch scores were calculated using the grammar checking facility in Microsoft Word 2002 (Microsoft Corporation, Redmond, Washington, USA).

Twenty articles, the 10 with the highest scores and the 10 with the lowest scores, underwent a further evaluation. The ‘find and replace’ feature in the word processing program was used to reduce the length of three selected words from each article. Each of these words contained three syllables or more. They were then reduced to a stem word e.g., ‘cholecystectomy’ to ‘cyst’, ‘colorectal’ to ‘rectal’. The Flesch scores were determined before and after these changes.

Data were described using the median value and the absolute range. Comparative statistics were not used in this descriptive study.
RESULTS

Figure 1 details the Flesch scores for the three surgical journals. Overall, the median score was 15.1 (0.0 – 29.1) - the median scores for each of the journals were 12.4 (*Archives of Surgery*), 14.4 (*British Journal of Surgery*), and 18.6 (*ANZ Journal of Surgery*). Six papers had a score of zero. Although the formula indicates that it is possible to have a negative Flesch score; in practice, negative scores are rated as zero.

The replacement of long technical terms with stem words increased the median scores. For the lowest scoring papers it increased from 0.0 (0.0 – 4.3) to 10.9 (2.4 – 16.9). Whilst, for the highest scoring papers it increased from 26.4 (25.2 – 29.1) to 29.9 (28.7 – 36.0). It was also noted that some articles tended to repeat the same surgical term. One paper with an average sentence length of 25.5 words used the term ‘parathyroid’ once in every 22.5 words.

\[^4\]
DISCUSSION

The following quote from Sir Winston Churchill has a Flesch score of 95 - "We shall go on to the end. We shall fight in France. We shall fight in the seas and oceans. We shall fight on the beaches, in the fields, in the streets, and in the hills. We shall never surrender." But, it is hard to maintain this style without the text becoming staccato and losing its high impact.

What is a suitable Flesch score for the surgical literature? It is generally recommended that writers should aim for a score between about 60 and 70. However, Table 2 suggests that it might be more realistic to aim for scores between 30 and 50. After all, many non-medical writers have such scores when conveying opinions that involve the use of complex terms. Even TS Eliot, whose poems tended to be allegorical to the point of obscurity, was able to maintain a Flesch score of above 40 when explaining his ideas about scholarship.

This study of surgical articles found a median Flesch score of 15. In comparison, reviews of articles published in the British Medical Journal, the Journal of the American Medical Association, and the Annals of Internal Medicine have reported mean Flesch scores of about 30. This is of major concern. Texts with Flesch scores below 30 have been compared to legal contracts. But, legal documents need to define complex situations. There is no reason for the surgical literature to mimic legal documents.

Editorial processes only have a marginal effect on the readability of articles. Roberts and others looked at manuscripts sent to the Annals of Internal Medicine. The mean
Flesch score was 28.9 on submission and this only rose to 29.1 at the time of publication. The authors concluded that the papers ‘remained difficult to read at publication’. Editorial staff do not have the time or the resources to re-write papers. Well written papers have a greater chance of being accepted for publication. Poorly written papers are more likely to be rejected.

Why do Flesch scores vary between journals? This study observed that the American journal had the lowest median Flesch scores. Also, Weeks and Wallace\textsuperscript{11} found that articles published in the \textit{British Medical Journal} were easier to read than those published in \textit{Journal of the American Medical Association}, as indicated by higher mean Flesch scores (31.5 verses 27.8). This raises two issues. First, these differences are marginal. They would not be evident to the readers. Second, the comparisons are not reliable - the papers were not chosen at random, only one journal from each country was selected for study, there is no allowance for the nature and quality of the content, and not enough papers have been selected for study.

This study looked at the effect of contracting specific surgical terms into stem words. For papers with low Flesch scores, this resulted in a median increase in the score from zero to 10.9. This is well short of being acceptable. It is inevitable that clearing the text of such words will improve the Flesch scores. But, the changes are relatively minor. There is more
to the poor readability of surgical articles than the use of complex surgical terms. It is the style of writing that is the problem.

How should readability scores be used? They can be used on sections of text to identify problem areas. They signal the need for further editing. Table 3 demonstrates how editing can improve the readability of a passage of text. Of course, the final judge on whether such changes are acceptable is the reader. But, there is more to readability than the use of short sentences and simple words. Writing to generate a low readability score can result in a jerky and impoverished text. This is why the use of readability indices is probably of greatest benefit during the final drafting stage.

Computerized assessments of readability vary according to the choice of software program. So, it is advisable to calculate readability scores using a standard procedure.

It is easy to get more information about readability scores. Typing ‘readability’ into the Help menu of a word processing program will provide the core facts. Whilst searching on the internet will reveal an diverse array of interesting sites. It is also important to note that some programs make it difficult to re-evaluate the same document. A useful strategy is to open the draft document, copy the relevant text, open a new file and paste the text into it, calculate the readability score, close the temporary file, and then continue to revise the draft document. Incidentally, the Flesch score for this article - minus the references and illustrations - is 41.2.
In conclusion, original articles in the surgical journals are difficult to read. This is due to the use of long sentences and complex words. However, this situation could be improved if authors used readability scores when composing text. It is realistic for authors to aim for Flesch scores above 30 when creating manuscripts.
ACKNOWLEDGEMENTS

The author thanks David Clark (Newcastle Breast Centre, New South Wales) for his encouragement to produce this article.
REFERENCES


Table 1. How to calculate two Indices of Readability

Select a block of text that contains more than 200 words. Calculate the average number of words per sentence (regard colons and semicolons as full-stops). Count the number of words with three or more syllables (hard words) - don't count words that are capitalized, combinations of short words, or verbs where the third syllable is a terminal ‘ed’ or ‘es’.

*The Fog Index:*

\[
\text{Fog Index} = 0.4 (\text{ASL} + \% \text{ of Hard Words})
\]

*The Flesch Score:*

\[
\text{Flesch Score} = 206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW})
\]

Where:

\[\text{ASL} = \text{The average sentence length}\]

\[\text{ASW} = \text{The average number of syllables per word}\]
Table 2. Readability scores for selected passages of text.

<table>
<thead>
<tr>
<th>Title</th>
<th>Flesch Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ernest the Fierce Mouse – ‘ideal for 6-8 year olds’.</td>
<td>78.0</td>
</tr>
<tr>
<td>Haworth - a travel essay by Virginia Woolf.</td>
<td>62.8</td>
</tr>
<tr>
<td>Notes from a Small Island – travel writing by Bill Bryson.</td>
<td>62.1</td>
</tr>
<tr>
<td>What is Science? - an essay by George Orwell.</td>
<td>46.8</td>
</tr>
<tr>
<td>Can Education be Defined? - a lecture delivered by TS Eliot at the University of Chicago in 1950.</td>
<td>43.9</td>
</tr>
<tr>
<td>Media Release Archives by the Royal Australasian College of Surgeons for May 2005 (<a href="http://www.surgeons.org">http://www.surgeons.org</a> – accessed 30 May 2005)</td>
<td>32.9</td>
</tr>
<tr>
<td>His Brain, Her Brain by Larry Cahill on SCIENTIFIC AMERICAN.com (<a href="http://www.sciam.com">http://www.sciam.com</a> – accessed 8 June 2005)</td>
<td>32.3</td>
</tr>
<tr>
<td>Our System of Government on the Australian Prime Minister’s Website (<a href="http://www.pm.gov.au">http://www.pm.gov.au</a> - accessed 30 May 2005)</td>
<td>27.7</td>
</tr>
</tbody>
</table>
Table 3. An example of text revision resulting in a higher Flesch score.

Original Text:\textsuperscript{14}

The proposal published recently in the UICC TNM supplement 2003, to subdivide T1 tumours into T1a (diameter up to 10 mm) and T1b (diameter between 11 and 20 mm), will increase the confusion because in the fifth edition the suffix ‘a’ denotes unifocality and the suffix ‘b’ multifocality. The authors recommend reinstitution of the T classification of the fifth edition until evidence-based data are available to justify its modification. (Flesch score = 1.3)

Revised Text:

It has been proposed to subdivide T1 tumours into T1a (diameter up to 10 mm) and T1b (diameter between 11 and 20 mm). This is confusing because ‘a’ and ‘b’ have indicated unifocality and multifocality. We suggest that there should be no change unless it is based on good evidence. (Flesch score = 42.7)
Figure 1. Flesch Reading Ease Scores (FRES) for selected text from the first 30 original articles published in 2005 within the Archives of Surgery (Arch Surg), British Journal of Surgery (Br J Surg), and the ANZ Journal of Surgery (ANZ J Surg).