ExELL Explorations in English Language and Linguistics

ISSN 2303-4858

CC BY

14.1 (2026): 21–45
https://doi.org/10.51558/2303-4858.2026.14.1.21
Original scientific article

Research writing in the sciences and humanities: The readability issue

Alvin Ping Leong

Nanyang Technological University, Singapore

ABSTRACT

Academic writing has generally been regarded as being difficult to read. While this is a common sentiment about science writing, recent research has suggested that it is true of humanities writing as well. However, comparative studies involving both genres are few. This study addressed this gap by comparing the readability of 2,000 research articles from the major sub-fields in the sciences and humanities. Nine readability formulas were used, and a normed average of the readability scores was computed for ease of comparison. The results revealed that the readability of science and humanities articles was in the 12-14 grade-level range. The humanities articles were found to be generally less readable than the science articles. The least readable sub-fields were history and chemistry, and the most readable sub-fields were classics and technology. This study offers a guide to the norms in the sub-fields of each discipline. Given that research articles are intended for a specific target audience, further discussion is needed as to whether the readability scores reflect a grade level that is considered appropriate.

ARTICLE HISTORY

Received
25 June 2026
Accepted
1 September 2026

KEYWORDS

readability; academic writing; science writing; humanities writing; readability formulas.

1. Introduction

The difference between science writing and humanities writing, due in large part to the distinct research approaches in the respective disciplines, has long been observed (Grech, 2019; Hogsette, 2019). Michalikova (2022) notes that the focus of research work in the sciences generally involves experiments, calculations, and simulations, which are separate from extensive reading or writing. In such a research orientation, writing serves primarily as a means to record what was done and found. By contrast, in the humanities, research work heavily involves both reading and writing (Michalikova, 2022). The

qualitative nature of humanities research often results in a writing style that is divergent, where different perspectives are presented and alternative interpretations explored. This view of science and humanities writing is commonplace, appearing in numerous descriptions of academic writing, as exemplified in the following:

> In the sciences, new knowledge is accepted by experimental proof. Science writing reinforces this by highlighting a gap in knowledge, presenting a hypothesis related to this gap, and then reporting experimental findings to support this. The humanities, on the other hand, rely on case studies and narratives while claims are accepted on strength of argument. (Hyland, 2008: 549)

Recent corpus-based comparative studies offer some support for this broad characteristic of science and humanities writing. Leong (2021, 2023), for instance, found that humanities writing had a more complex clause structure, displaying a more extensive use of projected and embedded clauses. In a later study, he also found that humanities writing used more thematized finite clauses, and science writing, more non-finite thematized clauses (Leong, 2025). This suggests that humanities writing prefers a more elaborate writing style focusing on interpretive depth and the integration of diverse perspectives from other scholars in the field (Leong, 2025). Science writing, on the other hand, values brevity and a more condensed style of writing.

These findings are insightful, but they also raise questions about the readability of science and humanities writing. After all, the use of different writing styles and linguistic features is bound to affect how easily texts are processed by the reader. On the one hand, it could be argued that the condensed nature of science writing makes it more difficult to read. However, the longer sentences and more complex use of clauses in humanities writing could pose similar readability issues. Indeed, sentiments about the readability of academic texts in general have not been encouraging. Descriptions about science writing range from the generic "difficult to read" to the more extreme "inaccessible" and "impenetrable" (Singh Chawla, 2020; Jones & Medina, 2021; Lingard & Watling, 2024). Humanities writing has also been criticized as being obscure and convoluted. Bruce Cole, who was chairman of the National Endowment for the Humanities (2001–2009), referred to it as being "needlessly opaque" with "unfathomable theory-speak" (Cole, 2016). Such views on scholarly writing are supported by a recent study by *The Economist*, which found that doctoral abstracts across disciplines had become harder to read over the past 80 years, with this trend being most pronounced in the case of the humanities (*Academic Writing*, 2024).

The study by *The Economist* is a useful gauge, but it is limited in its analysis of only doctoral abstracts. Where research articles are concerned, Hartley et al. (2004) found that science articles had lower readability scores than humanities articles. In the main, however, studies involving research articles are few and far between, and so the question concerning the readability of science and humanities writing in this area remains an open one. Further, both *The Economist* and Hartley et al. (2004) relied on a single measure of readability—the Flesch Reading Ease score. Given that scholarly publications are the mainstay of academic work, a more careful look at research articles using other readability measures is needed for a fuller understanding of science and humanities writing.

To address this gap, this study investigated the readability of science writing and humanities writing based on a corpus of 2,000 research articles in the major sub-fields of each broad discipline. Nine readability formulas were used to compute the normed average for each article and for each sub-field.

The rest of the paper is organized as follows. Section 2 describes the common readability formulas in use and related studies involving academic texts. Section 3 details the corpus of research articles and the analytical procedure. The findings are presented and discussed in Section 4. The concluding section outlines the implications of the study and recommends areas for further research.

2. Readability and related studies

2.1. *Readability formulas*

According to an oft-cited definition, readability is the "ease of understanding or comprehension due to the style of writing" (Klare, 1963: 1). The notions "ease" and "comprehension" are naturally subjective, but attempts have nevertheless been made over the years to quantify them using linguistic features such as word/sentence lengths and vocabulary lists (e.g., Flesch, 1948; O'Hayre, 1966; Chall & Dale, 1995). It is important to note at this stage that readability formulas make no claims that the scores accurately reflect the understandability of the text, since not all textual features (e.g., writing style, connotative meaning) can be easily reduced to an algorithm. Notwithstanding this limitation, and in the absence of other quantifiable measures, readability formulas remain a useful baseline approximation of the accessibility of any given text. Further, as some studies have shown, readability formulas tend to be more accurate for higher-ability individuals (Begeny & Greene, 2014), making them a useful, albeit imperfect, tool for assessing scholarly articles.

Klare's definition draws attention to two considerations—ease and comprehension—that have characterized the design of readability formulas over the years. Both are inextricably tied to the education level of the reader, since what is considered easy or difficult to understand depends on the reader's education level. A word such as "chastise" might appear difficult to a fourth grader but should pose few problems to a twelfth grader. Accordingly, readability formulas generate scores that are equated with either the American grade level or a reading scale. The grade level corresponding to a readability score is straightforward: the grade level is simply the readability score, rounded to the nearest integer. For instance, a grade-level score of 10.81 corresponds to Grade 11, and a score of 9.18 to Grade 9, and so forth.

Reading scales, on the other hand, differ depending on the formula that is used. Table 1, adapted from Scott (2024), lists the reading levels for each reading-scale score based on three readability formulas—Flesch Reading Ease (FRE), Linear Write (LW), and new Dale-Chall (NDC).

Table 1: Reading-scale scores and the corresponding difficulty levels

FRE	LW	NDC	Difficulty
	≥ 120	≥ 58	Extremely easy
	115–119	54–57	Extremely easy
≥120	110–114	50–53	Very easy
110–119	105–109	45–49	Fairly easy
100–109	100–104		Easy
90–99	95–99	40–44	Easy
80–89	90–94		Easy
70–79	85–89	34–39	Average
60–69	80–84		Above average
60–69	75–79	28–33	Slightly difficult
50–59	70–74		Slightly difficult
	65–69	22–27	Fairly difficult
	60–64		Fairly difficult
30–49	55–59	16–21	Difficult
	50–54		Difficult
	45–49		Very difficult
	40–44		Very difficult
0–29	< 40	≤ 15	Extremely difficult

Over the years, a number of readability formulas have been proposed (see Klare, 1974), each with its own purpose and calculation basis. Present-day online calculators rely on the more well-known of these formulas. The specifics of these nine commonly used formulas are presented below as a backdrop to the review of related studies in Section 2.2, and the analysis undertaken in this present study. The formulas are divided into reading-scale and grade-level formulas. A full list of the abbreviations is provided in the appendix.

Reading-scale readability formulas

(a) FLESCH READING EASE (FRE) was developed by Flesch (1948). It focuses on the average sentence length in words and the average number of syllables per word.

$$\text{FRE} = 206.835 - 1.015\frac{\text{Total}_{\text{words}}}{\text{Total}_{\text{sentences}}} - 84.6\frac{\text{Total}_{\text{syllables}}}{\text{Total}_{\text{words}}}$$

(b) LENSEAR WRITE (LW) differs from FRE by focusing on single-syllable words. Such words, characterized as the "most natural to the language" (O'Hayre, 1966: 7), are regarded as the hallmark of plain English. The formula, however, excludes commonly occurring words such as the definite article "the" and the verbs "is", "are", "was", and "were", even though they are single-syllable words.

$$\text{LW} = \frac{100}{\text{Total words}} \times (\text{Words}_{\text{syllable}=1} + (3 \times \text{Total sentences}))$$

(c) The NEW DALE-CHALL (NDC) formula uses a list of 3,000 words, pegged at the fourth grade level, to assess the reading ease of a text (Chall & Dale, 1995). NDC is new in the sense that the word list is an expanded version of the original list compiled in the 1940s. Words that lie outside the list are considered unfamiliar words.

$$\text{NDC} = 64 - 95\frac{\text{Unfamiliar words}}{\text{Total words}} - 0.69\frac{\text{Total words}}{\text{Total sentences}}$$

Grade-level readability formulas

(d) The AUTOMATED READABILITY INDEX (ARI) differs from the other readability measures by relying on characters (e.g., letters) rather than syllables. Smith and Senter (1967) found a close relationship between the letters used in a word and the syllables it contains. Using letters is also computationally simpler than counting syllables.

$$\text{ARI} = 0.5\frac{\text{Total words}}{\text{Total sentences}} + 4.71\frac{\text{Total characters}}{\text{Total words}} - 21.43$$

(e) The COLEMAN-LIAU (CL) formula echoes ARI in using letter counts per word rather than syllable counts. At the time, using mechanical counting devices to estimate syllable counts for each word could not always return accurate results as compared to letter counts (Coleman & Liau, 1975). In this respect, CL is similar to ARI in its basic setup. Differences in the coefficients and constants are due to the different corpora used by both sets of researchers in their analyses.

$$\text{CL} = 5.88\frac{\text{Total letters}}{\text{Total words}} - 29.6\frac{\text{Total sentences}}{\text{Total words}} - 15.8$$

(f) The GUNNING FOG (GF) formula was developed by Gunning (1952), who was of the view that lengthy sentences and polysyllabic words contributed to the "fog" in writing, making texts hard to read. Words containing three or more syllables are regarded as complex words, but these exclude proper nouns (e.g., "Switzerland"), simple compound words (e.g., "afternoon"), and words inflected for tense or plurality (e.g., "constructed").

$$\text{GF} = 0.4\left(\frac{\text{Total words}}{\text{Total sentences}} + 100\frac{\text{Total complex words}}{\text{Total words}}\right)$$

(g) SMOG, an acronym for "simple measure of gobbledygook", was developed by McLaughlin (1969). Like GF, it focuses on polysyllabic words (≥ three syllables) as an approximate measure of semantic difficulty. Unlike GF, however, it recognizes that the relationship between word difficulty and reading difficulty is not linear, as the other formulas appear to suggest. McLaughlin (1969: 640) notes that "[a] slight difference in word or sentence length between two passages does not indicate the same degree of difference in difficulty for hard passages, as it does for easy passages." For this reason, the square root is introduced into the formula.

$$\text{SMOG} = 3.1291 + 1.0430\sqrt{(\text{Words}_{\text{syllables}\geq 3}) \times \frac{30}{\text{Total sentences}}}$$

(h) By contrast, FORCAST, an acronym derived from a blend of the developers' names (Ford, Caylor, and Sticht), emphasizes single-syllable words rather than polysyllabic ones (Caylor et al., 1973). It is also the only formula in this list that does not take sentences or sentence lengths into account, making it ideal to assess texts where sentences need not be used (e.g., lists, labels).

$$\text{FORCAST} = 20 - 15\frac{\text{Words}_{\text{syllable}=1}}{\text{Total words}}$$

(i) The FLESCH-KINCAID (FK) formula is a reconfiguration of FRE. As seen earlier, FRE is a reading-scale index, which requires an additional step to convert its

scores to a corresponding grade level. Kincaid et al. (1975) reformatted the FRE formula to directly produce grade-level scores, thus removing the need for such a step.

$$\text{FK} = 0.39\frac{\text{Total words}}{\text{Total sentences}} + 11.8\frac{\text{Total syllables}}{\text{Total words}} - 15.59$$

There are two points to note as regards the readability formulas. First, the formulas typically rely on word length and sentence length as the core textual features to assess readability (FORCAST, as noted earlier, is an exception). These features are favored because they can be easily and objectively measured. This is not to say that other features such as syntactic complexity, coherence, or writing style are not relevant, but they can be difficult to quantify and may make readability formulas less accessible for general use.

Second, as readability formulas are regression equations, differences in the coefficients and constants in the formulas reflect the different corpora analyzed by the developers. These range from texts appropriate for different education grades (e.g., CL, SMOG) to materials used in the military (e.g., ARI, FK). For this reason, corpus-based studies, particularly the more recent work, have recognized the need to include at least two formulas to obtain a more accurate understanding of readability. We turn now to take a closer look at these studies.

2.2. *Studies on the readability of research writing*

Research on the readability of academic texts has tended to involve the medical and health sciences. The urgency in ensuring that health-related texts are accessible is understandable, considering how the general public often relies on online articles to educate themselves, make decisions, or share information (Fick et al., 2025). This, however, is sometimes compromised because of the use of jargon (Wen & Yi, 2023).

One obvious solution to this conundrum is to simplify the language. Swanson et al. (2024), for instance, found that substituting hypernyms for complex words and phrases improved readability, reducing the grade level of texts from the college level to the middle/high-school level. Much effort has also been made to compare the readability of lay summaries and traditional abstracts in science articles. Intuitively, lay summaries should be easier to read since they are intended for a wider audience. The results, however, are mixed. Using four readability formulas (ARI, GF, FK, and CL), Wen and Yi (2023) found that lay summaries required a higher grade level to read than the corresponding abstracts. As the authors themselves admitted, this finding was surprising and went against common expectations, although it should also be noted that the anomaly could be due in part to the restricted corpus used by them—only texts from a single journal were considered.

In other studies involving texts from different journals, more congruent findings were reported. In the field of autism research, lay summaries were found to be more readable than abstracts, although the former were still viewed as being difficult for non-specialist audiences (Wen et al., 2023). Stricker et al. (2020) and Anderson et al. (2022), working with texts from psychology and cystic fibrosis, respectively, found similar

results. The improved scores for lay summaries were consistently reflected across different readability indices such as FRE, FK, and SMOG, suggesting that these plain-language texts are effective in communicating research to non-specialist readers (Stricker et al., 2020).

While concerns about making lay summaries accessible to the general public are well justified, this still leaves unanswered the question as to why science writing itself remains difficult to read (Hochhauser, 2003). Indications from a number of studies suggest that it is *advantageous* for scientists to keep their writing technical. In the peer review process, for instance, it was found that preprints written using more technical language were regarded more favorably by reviewers (Sun et al., 2024). Several studies have also shown that less-readable articles were more likely to be cited (Hartley et al., 2002; Stremersch et al., 2007; Ante, 2022). As Ante (2022: 1) wryly notes:

> If readability indeed drives citations, the results imply that scientists have an incentive to (artificially) reduce the readability of their abstracts in order to signal quality and competence to readers—both to get noticed at all and to attract more citations. This may mean a prisoner dilemma in academic (abstract) writing, where authors intentionally but unnecessarily complicate the way in which they communicate their work.

It is arguably for this reason that science writing is not just difficult to read, but has become more so over time. In a diachronic study, Plavén-Sigray et al. (2017) analyzed over 700,000 abstracts published from 1881 to 2015 from a diverse mix of scientific disciplines, ranging from molecular biology to psychology. The results revealed that science writing in general had become increasingly more difficult to read. The FRE scores dipped from a mean of 40.14 in 1881 to 10.12 in 2015. Plavén-Sigray et al. (2017) also used the older, grade-level version of the Dale-Chall formula and found that the grade levels rose from 9.53 in 1881 to 12.99 in 2015 due to the increased use of jargon. That is to say, following Table 1, science writing has evolved from being difficult to extremely difficult to read, with a jump of two to three grade levels.

This decrease in readability appears to be true not merely for science writing, but research writing in general. In a study similar in purpose to Plavén-Sigray et al.'s (2017) work, *The Economist* analyzed 347,000 abstracts from doctoral dissertations published from 1812 to 2023 in the sciences, social sciences, and humanities (*Academic Writing*, 2024). The FRE scores revealed that the abstracts had become harder to read in all three broad disciplines, with the shift being the most marked in the case of the humanities. The study by *The Economist* provides an update to an earlier work by Hartley et al. (2004), who also used the FRE formula and found that humanities writing was more readable than writing in the sciences and social sciences. Hartley et al.'s (2004) study, however, considered only three journals, one for each major discipline, and the corpus involved only extracts from the introductions of articles published in 2001–2002. The investigation by *The Economist* suggests that the current situation is quite different.

Other efforts at comparing the readability of scholarly texts across disciplines involved the work of Okulicz-Kozaryn (2013) and a follow-up study by Lei (2016). Okulicz-Kozaryn (2013: 679) focused on only two word classes, adjectives and adverbs, arguing somewhat forcefully that such words were "unnecessary clutter." In his sample of about 1,000 articles, he found that the natural sciences had the lowest proportions of

adjectives and adverbs, and the social sciences had the highest, about 15% more than the former. The humanities were ranked behind the social sciences. The low proportions of adjectives and adverbs (taken as a single group) in the natural sciences were viewed as a positive attribute, leading Okulicz-Kozaryn (2013) to advocate that scientists should remove as many of the adjectives and adverbs as possible to ensure readability. Lei (2016) expanded on Okulicz-Kozaryn's (2013) study by analyzing adjectives and adverbs separately, rather than as a single group. The corpus was also far larger, comprising 6,459 texts. Lei (2016) found that the sciences used more adjectives but fewer adverbs than the humanities, thus providing a more nuanced understanding of the occurrences of these two word classes. Lei (2016) also included a rudimentary measure of readability by calculating the mean length of words (according to characters), and found that the use of adjectives and adverbs did not correlate with this readability index.

As we have seen, readability studies have tended to focus on science writing. Research interest in humanities writing, by contrast, has been far more modest. Further, although humanities texts were included in the studies by *The Economist*, Hartley et al. (2004), Okulicz-Kozaryn (2013), and Lei (2016), the corpus used in each case was overly restricted. *The Economist* relied on abstracts, and Hartley et al. (2004) on extracts from only a single section of journal articles. The analyses by Okulicz-Kozaryn (2013) and Lei (2016) were also confined to just two word classes. Even Plavén-Sigray et al.'s (2017) diachronic study involved only abstracts, rather than full texts. Yet, abstracts and full texts have fundamentally different functions and linguistic features (Cohen et al., 2010). Abstracts cannot be taken to represent research writing as a whole.

To address this gap, the present study analyzed 2,000 full articles published in journals representing the major sub-fields in the sciences and humanities. It sought to find out the readability of research articles in these two major disciplines, and what this implies for research writing in general. A suite of nine readability formulas was used. I present next details about the corpus and method of analysis used in this study.

3. Methodology

3.1. *Corpus*

The corpus, containing close to 19 million words, comprised 2,000 research articles in the sciences and humanities. Following the classification in Scimago (2024), five sub-fields in each broad discipline were identified, and within each sub-field, two journals were selected. The journals were highly regarded according to available ranking lists (e.g., Scopus, Scimago). For instance, in Scimago, Nature Genetics was ranked first in the field of genetics, and Opuscula was second in the field of classics.

These journals represent a range of research orientations, reflecting what Becher and Trowler (2001: 44) term academic tribalism, where sub-fields are conceived of "having recognizable identities." Each broad discipline, that is to say, is not a collection of homogeneous sub-fields, but one where marked distinctions between sub-fields exist. In science, for instance, the selected biology journals emphasized experimental work on cells, cancer, and genetics, among others. By contrast, the physics and technology

journals were far more quantitative in nature, with the former focusing on matter and energy, and the latter on computational systems. Similarly, in the humanities, while both the literature and philosophy journals covered issues at the conceptual level, philosophical arguments involved a wide range of real or hypothetical phenomena, and not merely the written text, which is fundamental to literary studies.

One hundred research articles from each journal were then downloaded for analysis. These articles were published from 2014 to 2024 and were the most recent publications at the time of writing. The reason for the wide year gap is that the humanities articles, unlike those in the sciences, were published less frequently. Table 2 provides a breakdown of the sub-fields in the sciences and humanities, and the journals in each sub-field. The humanities articles were generally longer, exceeding the science articles by more than half a million words in total.

Table 2: Basic details of the corpus

Discipline	Journal	Pub. year	Articles	Words
Sciences				
Biology	Cell	2024	100	1,405,810
	Nature Genetics	2024	100	948,296
Chemistry	JACS Au	2024	100	623,117
	Nature Chemistry	2024	100	631,412
Technology	ACM Transactions on Graphics	2024	100	1,029,348
	Mathematical Programming Computation	2020–2024	100	939,641
Geology	Earth Systems and Environment	2023–2024	100	778,098
	Remote Sensing of Environment	2024	100	1,068,182
Physics	Nature Physics	2024	100	566,997
	Physical Review X	2024	100	1,191,885
Sub-total			1,000	9,182,786
Humanities				
Classics	Annual of the British School at Athens	2018–2024	100	1,526,642
	Opuscula	2014–2024	100	1,362,217
History	Journal of Global History	2020–2024	100	966,599
	Journal of Heritage Tourism	2022–2024	100	807,427
Literature	Journal of World Literature	2021–2024	100	719,615
	Language & Literature	2019–2024	100	881,907
Philosophy	Minds and Machines	2022–2024	100	1,003,124
	Noûs	2022–2024	100	1,097,299
Performing Arts	International Journal of Performance Arts and Digital Media	2020–2024	100	757,954
	Journal of Dance Education	2021–2024	100	593,106
Sub-total			1,000	9,715,890
Grand total			2,000	18,898,676

3.2. *Method of analysis*

The articles were analyzed using the nine readability formulas outlined in Section 2.1. Readability scores were generated via an online calculator developed by Scott (2024), a tool used in studies on various text types such as web articles (Uzunçıbuk et al., 2024) and educational articles (Miskiewicz et al., 2024). The calculator [retrieved June 3, 2025, from Scott (2024)] was manually configured to ignore cardinal numbers so that numerical data, particularly in science articles, were excluded from the analysis. Other features that were ignored included dates, web addresses, references, author information, and acknowledgments. All other textual segments, including those in tables, diagrams, and supplementary notes, were included for analysis. Titles, headings, and captions were also retained because they form part of the main text and contribute to the overall reading experience. Captions are important as they describe the contents of tables and diagrams; a badly written caption may result in confusion or misinterpretation.

Apart from the nine readability scores for each article, the calculator also computed the normed average of the scores in terms of a grade level. This grade-level average, which Scott (2024) terms the "average reading level consensus" (hereafter ARLC), is crucial as it brings together both the reading-scale and grade-level formulas. As we have seen, reading scales and grade levels move in opposite directions—a text is considered more readable if it has a *higher* reading-scale score (see Table 1) but a *lower* grade-level score. To solve this problem, Scott (2024) first converted each band in the reading scales of FRE, LW, and NDC (Table 1) to a matching grade based on the difficulty level of the text. For instance, the bands "70–79" (FRE), "85–89" (LW), and "34–39" (NDC), indicating a text of average difficulty, correspond to Grade 7. For each text, these converted scores are then added to the other grade-level scores (i.e., ARI, CL, GF, etc.). The ARLC score is the simple average of the total. It reduces reading-scale and grade-level scores to a single statistic and metric (grade level) to ease comparisons between texts.

Figure 1 shows a partial screenshot of the readability scores for a humanities article. For the sake of brevity, only three of the nine scores are shown in the screenshot; the ARLC score is listed at the top.

Statistical analyses were conducted using Real Statistics Resource Pack (Zaiontz, 2025), a Microsoft Excel add-in. Cronbach's alpha was computed to evaluate the extent to which the reading-scale and grade-level readability formulas, respectively, ranked and scaled the texts consistently. The Student's t-test was used to determine if the observed differences in mean scores between science and humanities articles were statistically significant. Welch's ANOVA was used to compare mean readability scores between sub-fields in each major discipline. For significant ANOVA results, the Games-Howell test was used as the follow-up test. The significance level for all statistical tests was α=.05.

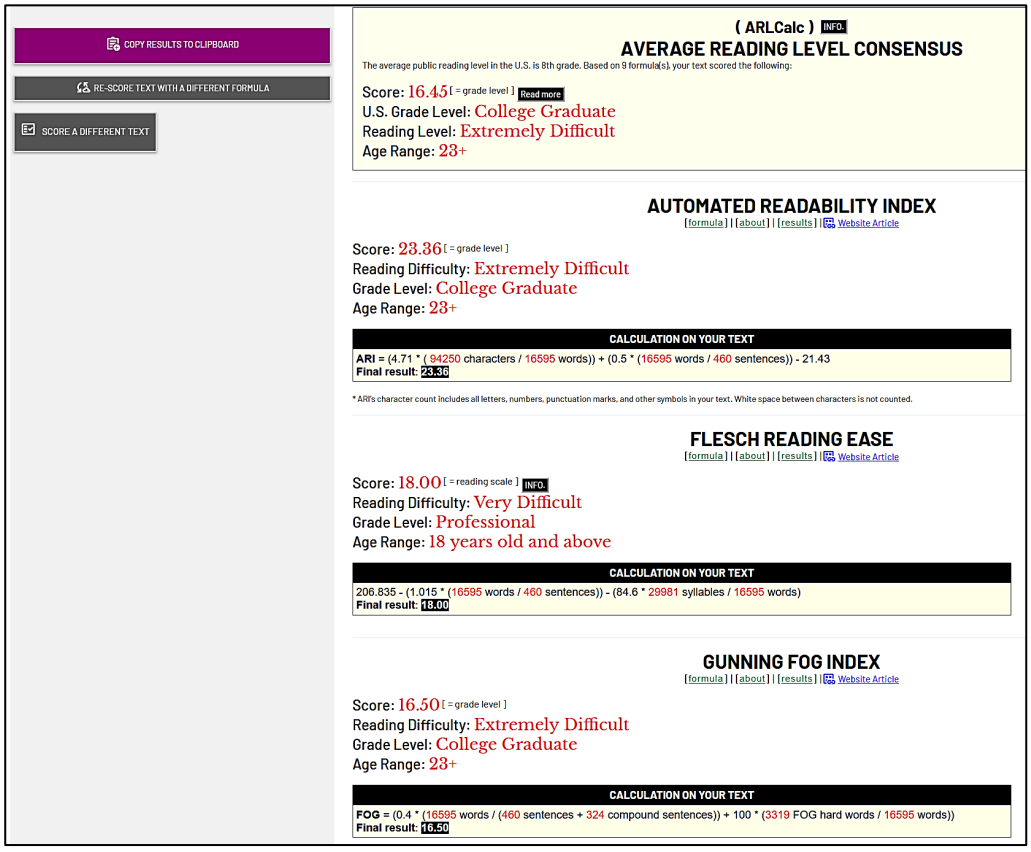
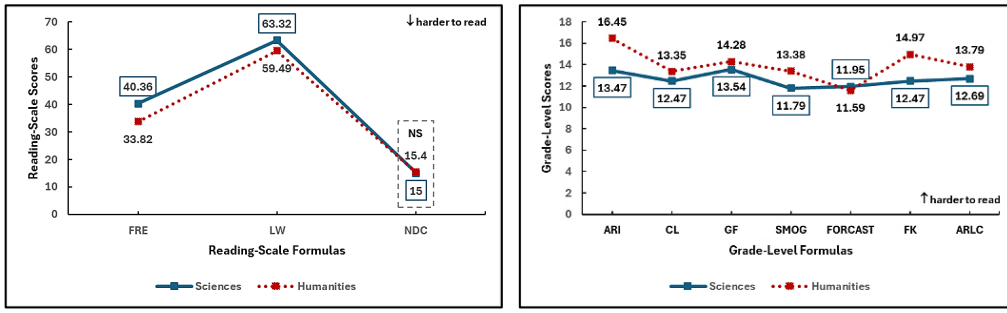


Figure 1: Partial screenshot of readability results using the online calculator

4. Findings and discussion

4.1. *Readability of science and humanities research articles*

Readability scores for the science and humanities articles are shown in Figure 2. Two sets of results are presented, one for the reading-scale scores and the other for the grade-level scores. The values for Cronbach's alpha showed consistency in the way the reading-scale formulas ($\alpha_{science}$=0.92, $\alpha_{humanities}$=0.88) and grade-level formulas ($\alpha_{science}$=0.96, $\alpha_{humanities}$=0.95) ranked and scaled the texts. In the diagram, the science scores are boxed to distinguish them from the humanities scores. Apart from NDC, all the observed differences in means were found to be highly significant ($p<.001$). The statistical non-significance of the NDC scores is marked off by the dashed box "NS" (for "not significant") in Figure 2(a).



(a) Reading-scale scores (b) Grade-level scores

Figure 2: Mean readability scores for science and humanities articles

The ARLC scores in Figure 2(b) reveal that the humanities articles (13.79, *SD*=1.88) were more difficult to read as compared to the science articles (12.69, *SD*=2.15; $t(1998)=12.11$, $p<.001$, 95% CI [0.92, 1.27], $d=0.54$). This finding is true for seven of the nine formulas as well, indicating a general consensus among the formulas.

Of interest are the FRE scores—40.36 (science) and 33.82 (humanities). These are noticeably higher than the scores reported in related studies. Plavén-Sigray et al. (2017), for instance, found that the mean FRE scores for their most recent science abstracts were 10.12. The findings from the study by *The Economist* are harder to pin down, as there is no official report detailing the (breakdown of the) specific results. Instead, the findings are mentioned in broad terms as part of an online article. The mean FRE scores for doctoral abstracts in the sciences, social sciences, and humanities appear to be about 18 in the 2020s (*Academic Writing*, 2024). As can be seen, the scores from both these studies are considerably lower than the FRE scores in Figure 2(a), implying less readability.

This discrepancy in the FRE scores is likely due to the different text types analyzed. Both Plavén-Sigray et al. (2017) and *The Economist* relied on only abstracts, as opposed to this present paper, which analyzed full manuscripts. As abstracts are a shortened version of the full paper, the language tends to be more condensed with a greater occurrence of technical terms. Full papers, on the other hand, have a greater variation in word/sentence length and word choice. Given these differences, the readability of abstracts and full papers is likely to be dissimilar. Yeung et al. (2018) have shown that abstracts tend to be less readable than full papers, a result which is consistent with the discrepancy noted in the FRE scores here.

We move next to NDC and FORCAST, which produced results that did not conform to the general trend. The NDC scores were found to be statistically not significant, and the FORCAST scores showed that the science articles were harder to read as compared to the humanities articles. We begin with NDC. As this formula relies on a pre-determined list of familiar words, the statistically non-significant result for the scores is perhaps not surprising. This is because the corpus comprised research-based scholarly work, which naturally contained specialized vocabulary and formal language intended for the professional reader. In this regard, there was little separating the articles in the two major disciplines—according to Table 1, the difficulty level for both hovered at the

border between "very difficult" and "extremely difficult." This is illustrated in (1–2) below; in all textual examples in this paper, the major discipline, sub-field, and article number are indicated in brackets.

(1) *Quantum algorithms, notably quantum principal component analysis (qPCA), the Harrow, Hassidim, and Lloyd (HHL) algorithm, and least-squares quantum support vector machines (qSVM), offer exponential speedups over classical counterparts in processing large datasets, solving linear systems, and machine learning tasks. (Science, Geology, Earth058)*

(2) *Traditionally epistemologists have taken doxastic states to come in three varieties – belief, disbelief, and suspension. Recently many epistemologists have taken our doxastic condition to be usefully represented by credences – quantified degrees of belief. (Humanities, Philosophy, Noûs038)*

Where word complexity is concerned, it is perhaps tempting to think of it as a science-writing problem, due in part to complaints in both academic circles and the popular media about this genre being full of jargon (e.g., Singh Chawla, 2020). However, as (2) indicates, the wording in humanities writing can be just as complex. Words such as "epistemologists", "doxastic", and "credences" are seldom used and would likely be incomprehensible even to adult readers.

Of the nine formulas, FORCAST was the only one that returned scores showing an opposite result. Science articles (11.95, SD=0.52) had a higher grade-level mean score than humanities articles (11.59, SD=0.67; t(1998)=13.60, p<.001, 95% CI [0.31, 0.42], d=0.61). One might suspect that this is due to the fact that FORCAST excludes sentences and sentence lengths in its computation. In this respect, some accounts (e.g., Scott, 2024) have noted that FORCAST is more suitable for assessing passages where the orthographic sentence need not apply. Examples of such passages include titles, section headings, and captions. However, as there is no marked difference in the way these text elements appear in both the science and humanities articles, they are unlikely to have contributed to the overall result. Further, in the case of captions, the period mark is conventionally inserted at the end of the caption, as seen in (3–4) below, regardless of whether the caption itself is grammatically a sentence. An automated calculating system would simply regard these lines as sentences. Hence, the effect of "non-sentences" in nudging the result in the opposite direction appears minimal.

(3) *Cellular localization of immune and lysosomal proteins. (Sciences, Biology, NatGen003)*

(4) *General view of the temple from the west, looking along its longitudinal axis. (Humanities, Classics, Opuscula042)*

A more likely explanation is the number of monosyllabic words used in each discipline. The FORCAST formula lays emphasis on such words – the lower the number of monosyllabic words, the higher the grade-level score, and vice versa. As shown in Table 3, the mean percentage of monosyllabic words to total words was lower in the science articles (53.67; SD=3.48) than in the humanities articles (56.09, SD=4.44; t(1998)=13.60, p<.001, 95% CI [2.08, 2.78], d=0.61). Going by this criterion alone, it would thus appear that the science articles were less readable. However, when other factors are

taken into account—such as the longer sentence lengths in the humanities articles (26.62 vs. 20.56; $t(1998)=20.57$, $p<.001$, 95% CI [5.48, 6.63], $d=0.92$)—the scores in Figure 2(a–b) suggest that the humanities articles were actually harder to read.

Table 3: Monosyllabic and polysyllabic (≥3 syllables) words and mean sentence lengths in the sciences and humanities

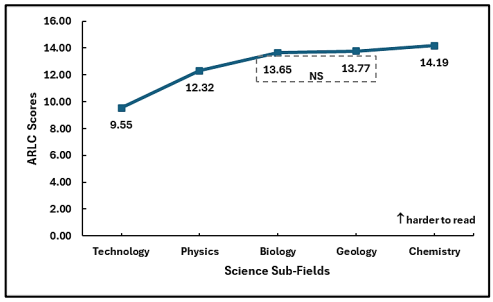
	Mono. Words (%)	SD	Poly. Words (%)	SD	Sent. length (words)	SD
Sciences	53.67	3.48	20.73	2.99	20.56	6.68
Biology	52.80	2.76	21.21	2.27	23.04	3.03
Chemistry	52.87	3.29	22.54	3.16	24.94	2.87
Technology	54.81	4.07	18.42	2.25	10.83	4.78
Geology	52.58	3.34	21.01	2.84	24.47	3.90
Physics	55.27	2.84	20.48	2.76	19.52	5.63
Humanities	56.09	4.44	20.62	4.24	26.62	6.49
Classics	58.48	4.01	17.14	3.44	23.65	6.23
History	52.31	4.00	24.49	4.14	31.00	8.45
Literature	57.03	3.53	19.50	2.87	27.39	5.63
Philosophy	57.61	4.27	19.86	3.43	25.86	4.66
Performing Arts	55.03	3.44	22.12	3.16	25.18	4.03

The NDC and FORCAST exceptions highlight the importance of using *multiple* formulas to guard against the bias of individual formulas. Relying on just a single catch-all metric may give misleading results. Using only the FORECAST or NDC formula in this case may lead one to mistakenly conclude that science writing is less readable than, or is as hard to read as, humanities writing. But seven other readability formulas and the ARLC scores suggest a different result.

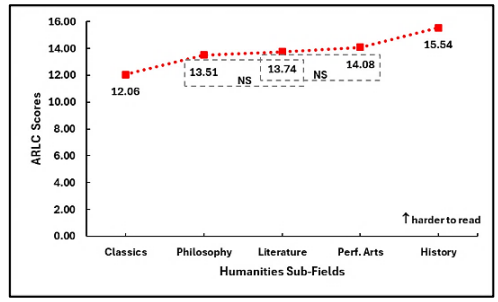
4.2. *Readability of science and humanities research articles by sub-fields*

As the sciences and humanities are broad disciplines, a closer look is needed to investigate the readability levels of articles written in the various sub-fields, and whether this has any implications for the broad conclusion reached in Section 4.1.

The mean ARLC readability scores for the sub-fields in the two major disciplines are shown in Figure 3; the scores are arranged in ascending order.



(a) ARLC scores for science sub-fields



(b) ARLC scores for humanities sub-fields

Figure 3: Mean ARLC scores for sub-fields in the sciences and humanities

Within each discipline, ANOVA results were statistically significant (science: $F(4, 487.68)=313.74$, $p<.001$, $\omega^2=0.62$; humanities: $F(4, 494.86)=109.16$, $p<.001$, $\omega^2=0.35$). The Games-Howell post-hoc test for pairwise comparisons revealed that all sub-fields within each discipline were significantly different from each other ($p<.001$), with the exception of the scores between (a) biology and geology, (b) literature and philosophy, and (c) literature and the performing arts. Figure 3 thus shows the ranking order of the sub-fields in each discipline, at least where the most and least readable sub-fields are concerned.

Comparing between the disciplines, the Student's *t*-test also revealed that the ARLC scores for several science sub-fields were higher than some humanities sub-fields. These sub-fields are listed in Table 4.

Table 4: Science sub-fields with higher ARLC scores than humanities sub-fields

Science sub-field	**Humanities sub-field**	
Biology (13.65, *SD*=0.78)	Classics (12.06, *SD*=2.01)	$t(398)=10.43$, $p<.001$, 95% CI [1.29, 1.89], $d=1.04$
Chemistry (14.19, *SD*=0.94)	Classics (12.06, *SD*=2.01)	$t(398)=13.57$, $p<.001$, 95% CI [1.82, 2.44], $d=1.36$
	Literature (13.74, *SD*=1.35)	$t(398)=3.80$, $p<.001$, 95% CI [0.21, 0.67], $d=0.38$
	Philosophy (13.51, *SD*=1.42)	$t(398)=5.61$, $p<.001$, 95% CI [0.44, 0.91], $d=0.56$
Geology (13.77, *SD*=0.99)	Classics (12.06, *SD*=2.01)	$t(398)=10.78$, $p<.001$, 95% CI [1.40, 2.02], $d=1.08$
	Philosophy (13.51, *SD*=1.42)	$t(398)=2.11$, $p=.036$, 95% CI [0.02, 0.50], $d=0.21$

In this light, the broad finding in Section 4.1 that humanities writing is harder to read than science writing now needs to be moderated by taking into account the difficulty level of texts in specific sub-fields.

4.2.1. *Science sub-fields*

Figure 3 suggests that texts in the technical sub-fields (i.e., technology, physics) are generally easier to read. This may appear counter-intuitive to some since the quantitative nature of such articles can be daunting. Consider, for instance, the following passage from a technology article.

(5) *Next, we introduce the cut formulation, for which we need the concept of induced subgraphs. Given a graph* $G = (V, E)$ *and a vertex subset* $V' \subset V$*, the induced subgraph on* V' *is* $G[V'] = (V', E')$ *with* $E' \coloneqq \{\{u, v\} \in E : u, v \in V'\}$*. (Sciences, Technology, MPC078)*

To a person who is not well versed in mathematics, (5) would be almost incomprehensible. It represents what some would regard as impenetrable writing. The ARLC score for (5), though, was 9.59, indicating that the language itself could be read by someone with at least a 10th-grade level of education. Thus, a possible reason why

some might find (5) difficult to read is not the words *per se*, but the mathematical symbols and equations sprinkled throughout the passage. Indeed, ARI, which focuses on characters such as letters and symbols, regarded the same passage as being considerably more difficult to read, giving it a score of 12.37.

Technically, however, symbols and equations are *not* words, and the ARLC score here explains in part the lower grade-level numbers for technology and physics articles. As noted earlier, ARLC is the normed average of the scores generated by the nine formulas used in this study. While it includes the ARI score, it also includes other grade-level scores, which in the case of (5) ranged from 7.24 to 10.45, reflecting the general agreement among these other formulas that the words used in the example are not overly challenging. In fact, if we disregard the symbols and equations, the message in the text is relatively straightforward. Omitting these elements lowered the ARI score to 8.21, and the ARLC score, in consequence, to 8.71. In the amended version (5a), bolded letters are used in place of the mathematical symbols and equations:

(5a) *Next, we introduce the cut formulation, for which we need the concept of induced subgraphs. Given a graph* ***'A'*** *and a vertex subset* ***'B'****, the induced subgraph on* ***'C'*** *is* ***'D'*** *with* ***'E'****.*

It might be argued that symbols and abbreviated terms are also frequently found in other sub-fields such as chemistry and biology, as witnessed in examples (6–7) below. Would the use of such terms not reduce the grade-level scores as well?

(6) *The Al–W bond length in 1·$W(CO)_5$ amounts to 258.5(1) pm, which is longer than in a carbazolylaluminylene tungsten pentacarbonyl complex (253.6(1) pm), but shorter than those reported for (TMEDA)(R)Al→$W(CO)_5$-type compounds (TMEDA = tetramethylethylenediamine; R = Cl: 264.5(2) pm; R = Et: 267.0(1) pm; R = ^{t}Bu: 274.1(4) pm). (Sciences, Chemistry, NatChem090)*

(7) *The DdmDE anti-plasmid system was discovered in a Vibrio cholerae strain of the O1 El Tor biotype (7PET), which is the primary cause of the seventh pandemic. From a representative 7PET strain A1552, two plasmid defense systems were identified and named DdmABC and DdmDE, respectively. DdmABC can defend against both plasmid invasion and bacteriophage infection. In contrast, DdmDE is mainly responsible for eliminating plasmids, particularly small, multicopy plasmids, by degradation. (Sciences, Biology, Cell093)*

The ARLC scores for these two passages were high: 15.84 for (6) and 13.95 for (7). However, unlike (5), the high scores do not appear to be due merely to the use of abbreviations, but due to the language itself. Both passages contain numerous unfamiliar, polysyllabic words, e.g., "carbazolylaluminylene," "anti-plasmid," "bacteriophage." Even removing the formulas and abbreviations did not improve the ARLC scores by much; the ARLC score was lowered to 14.92 for (6a), and to 13.52 for (7a). The scores remained high, at the collegiate level.

(6a) *The* ***'A'*** *bond length in* ***'B'*** *amounts to 258.5(1) pm, which is longer than in a carbazolylaluminylene tungsten pentacarbonyl complex (253.6(1) pm), but shorter than*

*those reported for **'C'**-type compounds (TMEDA = tetramethylethylenediamine; **'D'**: 264.5(2) pm; **'E'**: 267.0(1) pm; **'F'**: 274.1(4) pm).*

(7a) *The **'A'** anti-plasmid system was discovered in a Vibrio cholerae strain of the **'B'** biotype (**'C'**), which is the primary cause of the seventh pandemic. From a representative **'C'** strain **'D'**, two plasmid defense systems were identified and named **'E'** and **'F'**, respectively. **'E'** can defend against both plasmid invasion and bacteriophage infection. In contrast, **'F'** is mainly responsible for eliminating plasmids, particularly small, multicopy plasmids, by degradation.*

It should be added that the explanation above remains a conjecture, restricted to the specific examples raised in this sub-section. No global re-analysis of the physics and technology articles in the corpus was attempted in this study. It was also not practical, given that the re-analysis would have required the identification and replacement of all symbols and equations in each text.

4.2.2. *Humanities sub-fields*

The higher ARLC score for the humanities can perhaps be attributed to the nature of the discipline itself. As noted in commentaries and guidebooks, the humanities "use writing as a tool to reflect upon life, and to tell how life should, or should not, be lived" (*Writing in the Humanities*, n.d.). Words, that is to say, are the primary means to describe the spectrum of human experiences, and insofar as these experiences are diverse and complex, it stands to reason that the words would also be similarly complex (see Hyland, 2000).

This is not to say that the phenomena explored and described in the sciences are less complex. As shown in (6–7), technical terms referring to chemical compounds, organisms, and the like, can be difficult to read. Table 4 also shows that some science sub-fields such as chemistry and geology had higher grade-level scores than classics and philosophy. On the whole, as the variation in the humanities ARLC scores was less marked than the science ARLC scores (Figure 3), humanities articles can be taken to be generally harder to read.

Interestingly, in the use of symbolism, some humanities articles in the corpus were not dissimilar to the science articles. Equations are not normally associated with humanities writing, and yet, such equations were found in a number of philosophy articles as an efficient way to express logical reasoning. A somewhat extreme example is given in (8).

(8) *We assume the agent's beliefs about which state is actual are represented by a subjective probability function, $Pr(\cdot)$, and the agent's preferences over outcomes are represented by a cardinal utility function, $u(\cdot)$. We also assume that states and outcomes are indexed in ascending preference order, such that $u(x_i) \leq u(x_{i+1})$ for $i = 1, \ldots, n-1$. Then the risk-weighted expected utility of f is defined as*

$$REU(f) \coloneqq u(x_i) + \sum_{k=2}^{n} \left[r\left(\sum_{i=k}^{n} \Pr(S_i) \right) \cdot \left(u(x_k) - u(x_{k-1}) \right) \right]$$

(Humanities, Philosophy, Noûs008)

To the uninitiated, (8) is perhaps as inaccessible as (5). However, unlike the technology article, the ARLC score was considerably higher at 13.72, due in part to the use of polysyllabic words such as "subjective", "probability", and "utility", among others. In the corpus, symbolic notations and equations were not pervasively found in the philosophy articles. Where they did occur, they were used chiefly to set the context or to clarify connections between concepts. The preference, instead, was for arguments and discussions to be elaborated more fully in the writing. For this reason, the mean ARLC score for philosophy (13.51, SD=1.42) was higher than the same for technology (9.55, SD=1.72; t(398)=25.09, p<.001, 95% CI [3.65, 4.27], d=2.51) and physics (12.32, SD=1.82; t(398)=7.30, p<.001, 95% CI [0.87, 1.52], d=0.73).

This preferred approach contributes to what Leong (2025: 50) terms an "elaborate style of writing" in the discipline. This in turn suggests that differences among the humanities sub-fields could be due simply to the extent to which unfamiliar or difficult words are used. This appears to be true in the case of (9–10), taken from the history and classics sub-fields.

(9) *The proliferation of diaspora policies by homeland states, especially since the 1990s, has seen a strategic shift in the conceptualisation of diaspora. The diaspora is now recognised as an 'ongoing transnational network' or 'transnational community', inclusive of a variety of generations of members. Accordingly, belonging to a diaspora community is maintained through a range of transnational engagements, policies, and strategies, such as through diaspora tourism. (Humanities, History, HerTour057)*

(10) *Harpocrates is typically depicted as a naked infant/an early-age child, although on occasion he is portrayed wearing cloth loosely draped around his waist area or with a cloak flowing down his back. He appears in various poses, seating, reclining or standing. He is typically represented in the act of making a characteristic gesture, either touching or approaching his lips with the index finger of his right hand as a sign of his infancy. Harpocrates may be depicted with curly hair, or with his head almost entirely shaved, though preserving his typical Horus side-lock. (Humanities, Classics, BSA077)*

As was done for the earlier examples, both passages were scored for readability. The difference between the ARLC scores for (9) and (10)—16.28 vs. 12.52, respectively—is wide, approximately three education grades apart. This difference is doubtless attributed in part to the longer and more complicated words used in (9), such as "proliferation", "conceptualisation", and "transnational." Indeed, Table 3 shows that the percentage of polysyllabic words (≥3 syllables) was far higher in history articles (24.49, SD=4.14) than in classics articles (17.14, SD=3.44; t(398)=19.32, p<.001, 95% CI [6.60, 8.10], d=1.93). One might also suspect that the lower ARLC score for (10) is due to the design of one of the formulas used in the computation. As noted in Section 2.1, the GF formula regards proper nouns containing three or more syllables (e.g., "Harpocrates") as non-complex words.

The question as to why these two sub-fields, with a shared focus on historical phenomena and people, occupy opposite ends of the readability cline could well be due to their different research paradigms. As Kazarov and Petrechko (2019) explain, while

both classics and history require a close examination of historical sources, classics places a heavy emphasis on the language itself to clarify the original author's view. History, on the other hand, reconstructs and interprets the past, often requiring the historian "to criticize the source, i.e., to determine the degree of its reliability" (Kazarov & Petrechko, 2019: 14) so as to make informed judgments about development patterns and inter-relationships between historical events. This may explain in part the presence of more complex and polysyllabic words in history texts. A deeper investigation into the research approach and textual features of articles in these sub-fields would help to shed further light on how they shape the choice and use of words.

5. Conclusion and implications

This study compared the readability of science and humanities research articles based on ARLC, the normed average of scores generated by nine formulas. A summary of the key findings is as follows.

(a) Humanities articles were found to be harder to read than science articles. The ARLC scores for humanities and science articles were 13.79 and 12.69, respectively. On balance, the humanities articles were harder to read because of the discipline's more extensive use of words.

(b) Among the humanities sub-fields, classics articles had the lowest mean ARLC score (12.06), and history articles had the highest (15.54). Among the science sub-fields, technology articles had the lowest ARLC score (9.55), and chemistry articles had the highest (14.19).

It is also suggested that the use of symbolic notations and equations appeared to have an impact in the science sub-fields of technology and physics, since removing the notations and equations improved the readability scores. In chemistry, biology, and geology, however, the ARLC scores remained high, even when these notations and equations were disregarded. In the humanities, symbolism was used less pervasively, and so had a minimal impact on readability.

More than two decades ago, in response to the perceived "unreadable" writing style of medical articles, Hochhauser (2003) commented that further research on this topic was unnecessary since any future work would only come to the same conclusion. Hochhauser's reservation is understandable, but such studies remain important for us to understand more fully what the norms are in different disciplines and what the readability scores collectively imply about scholarly writing in general.

As we have seen, the readability scores for both the science and humanities articles are in the 12–14 grade-level range. While the scores are statistically different, they are separated by just one education grade. The average reader of articles in either discipline would need to have at least a 12th- or 13th-grade education. In many education systems, this translates to the final year of high school or the first year of university. The broader question that needs to be addressed, then, is whether this grade level of readership is appropriate. Based on oft-expressed opinions (e.g., Woolston, 2020), it does not appear

to be so, which begs the follow-up question—what would count as an appropriate grade level?

To say that a text needs to be written at a particular grade level is to claim that the text is intended primarily for those with that grade level of education. Thus, the question about the appropriate grade level for research articles is essentially one about the *intended* readership of research articles. Who are these readers? Hashmi (2015) suggests that research articles are meant for experts in the field. A study by Mohammadi et al. (2015) lends some support to this view. The authors found that the majority of readers from various disciplines were doctoral students, academics, and professionals. Put differently, research articles are intended for people with advanced interest and knowledge in the field, rather than the general reader. This is not to deny that the general public also reads such articles, but the numbers are likely to be far lower. The public instead relies more on the media (Science Media Centre, 2023).

In this light, the 12–14 grade-level scores for science and humanities articles appear reasonable. Nevertheless, could academic writing be made simpler? This is possible, and certainly something to aim for. After all, texts that are overly difficult to read may impose an unnecessary cognitive burden on readers and so hamper the efficient dissemination of knowledge. However, the nature of academic research and communication may constrain it. As Sun et al. (2024: 10) aptly note, "[w]riting in simple and clear language is easier for readers to understand, but it may be less practical for scientific papers laden with technical content." Sun et al. (2024) go on to suggest some strategies to make research articles more accessible to non-specialists: using simplified summaries and highlighting important sections written in plain English. These strategies may help to bridge the gap between the specialist community writing research articles and the wider community who may be keen to find out more about their research. Recent advances in artificial intelligence (AI) technology have also made it easier for authors to use AI tools to simplify their writing appropriately.

A clear limitation of the present study is the selection of just two major disciplines. In future work, the social sciences will need to be included as well. As the social sciences straddle the sciences and humanities, investigating how (or whether) the readability of articles in this discipline and its sub-fields differ from the findings reported here will shed more light on the nature of academic writing.

More work is also needed to finetune current measures of readability. The current formulas, which rely on words and sentences, cannot adequately account for the role of other linguistic features in affecting the readability of texts, such as coherence, cohesion, and syntactic complexity, among others. With the growing pervasiveness of AI and developments in natural language processing, there is an urgent need to develop more comprehensive models of readability (see Crossley et al., 2023).

A further issue to explore is the perceptions of the target readers themselves. While quantitative metrics are useful in providing a mechanical measure of readability, the views of readers are perhaps even more important. This is because readability formulas cannot measure everything. Certain linguistic features such as coherence or the author's writing style cannot be easily quantified. This is where the views of readers are critical.

Doctoral students, researchers, and fellow academics will be in the best position to judge what the current challenges are and how research articles can be made more understandable.

References

Academic writing is getting harder to read—the humanities most of all. (2024). *The Economist* https://www.economist.com/science-and-technology/2024/12/18/academic-writing-is-getting-harder-to-read-the-humanities-most-of-all

Anderson, Hannah L., John E. Moore, Beverley C. Millar (2022). Comparison of the readability of lay summaries and scientific abstracts published in CF Research News and the Journal of Cystic Fibrosis: Recommendations for writing lay summaries. *Journal of Cystic Fibrosis 21*(1): E11–E14. https://doi.org/10.1016/j.jcf.2021.09.009

Ante, Lennart (2022). The relationship between readability and scientific impact: Evidence from emerging technology discourses. *Journal of Informetrics 16*(1): 101252. https://doi.org/10.1016/j.joi.2022.101252

Becher, Tony, Paul R. Trowler (2001). *Academic Tribes and Territories: Intellectual Enquiry and the Culture of Disciplines* (2nd edition). Buckingham: Society for Research into Higher Education & Open University Press.

Begeny, John C., Diana J. Greene (2014). Can readability formulas be used to successfully gauge difficulty of reading materials? *Psychology in the Schools 51*(2): 198–215. https://doi.org/10.1002/pits.21740

Caylor, John S. Thomas G. Sticht, Lynn C. Fox, J. Patrick Ford (1973). *Methodologies for Determining Reading Requirements of Military Occupational Specialties*. Alexandria, VA: Human Resources Research Organization.

Chall, Jeanne Sternlicht, Edgar Dale (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge, MA: Brookline Books.

Cohen, K. Bretonnel, Helen L. Johnson, Karin Verspoor, Christophe Roeder, Lawrence E. Hunter (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics 11*: 492. https://doi.org/10.1186/1471-2105-11-492

Cole, Bruce (2016). What's wrong with the humanities? *Public Discourse.* https://www.thepublicdiscourse.com/2016/02/16248/

Coleman, Meri, T. L. Liau (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology 60*(2): 283–284. https://doi.org/10.1037/h0076540

Crossley, Scott, Stephen Skalicky, Cynthia Berger, Ali Heidari (2023). Assessing readability formulas in the wild. Dascalu, Mihai, Patrizia Marti, Francesca Pozzi, eds. *Polyphonic Construction of Smart Learning Ecosystems*. Singapore: Springer Nature, 91–101. https://doi.org/10.1007/978-981-19-5240-1_6

Fick, Julian, Luca Rudolph, Friederike Hendriks (2025). Jargon avoidance in the public communication of science: Single- or double-edged sword for information evaluation? *Learning and Instruction*, *98*: 102121. https://doi.org/10.1016/j.learninstruc.2025.102121

Flesch, Rudolf Franz (1948). A new readability yardstick. *Journal of Applied Psychology 32*(3): 221–233. https://doi.org/10.1037/h0057532

Grech, Victor (2019). Writing a scientific paper (WASP): An overview of differences in styles between the sciences and the humanities. *Early Human Development 129*: 96–99. https://doi.org/10.1016/j.earlhumdev.2018.12.008

Gunning, Robert (1952). *The Technique of Clear Writing*. New York, NY: McGraw-Hill.

Hartley, James, Eric Sotto, Claire Fox (2004). Clarity across the disciplines: An analysis of texts in the sciences, social sciences, and arts and humanities. *Science Communication 26*(2): 188–210. https://doi.org/10.1177/1075547004270164

Hartley, James, Eric Sotto, James Pennebaker (2002). Style and substance in psychology: Are influential articles more readable than less influential ones? *Social Studies of Science 32*(2): 321–334. https://doi.org/10.1177/0306312702032002005

Hashmi, Aamir Rafique (2015). In defence of academic research. *The Straits Times*. https://www.straitstimes.com/opinion/in-defence-of-academic-research

Hochhauser, Mark (2003). Why are unreadable articles still being written? *BMJ 326*: 711. https://doi.org/10.1136/bmj.326.7391.711

Hogsette, David S. (2019). *Writing that Makes Sense: Critical Thinking in College Composition* (2nd edition). Eugene, OR: Resource Publications.

Hyland, Ken (2000). *Disciplinary Discourses: Social Interactions in Academic Writing*. London: Longman.

Hyland, Ken (2008). Genre and academic writing in the disciplines. *Language Teaching 41*(4): 543–562. https://doi.org/10.1017/S0261444808005235

Jones, Cherice Escobar, Genesis Barco Medina (2021). Teaching racial literacy through language, health, and the body: Introducing bio-racial rhetorics in the writing classroom. *College English 84*(1): 58–77. https://doi.org/10.58680/ce202131452

Kazarov, Sarkis, Oleh Petrechko (2019). Philological and historical research methods of classical studies. *East European Historical Bulletin* (13): 8–15. https://doi.org/10.24919/2519-058x.13.188648

Kincaid, J. Peter, Robert P. Fishburne Jr, Richard L. Rogers, Brad S. Chissom (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy Enlisted Personnel*. Millington, TN: Naval Air Station Memphis.

Klare, George Roger (1963). *The Measurement of Readability*. Ames, IA: Iowa State University Press.

Klare, George Roger (1974). Assessing readability. *Reading Research Quarterly 10*(1): 62–102. https://doi.org/10.2307/747086

Lei, Lei (2016). When science meets cluttered writing: Adjectives and adverbs in academia revisited. *Scientometrics 107*(3): 1361–1372. https://doi.org/10.1007/s11192-016-1896-3

Leong, Alvin Ping (2021). Writing in the sciences and humanities: A clause-complex perspective. *Word 67*(2): 137–158. https://doi.org/10.1080/00437956.2021.1909866

Leong, Alvin Ping (2023). Discipline-specific writing and embedded clauses: The case of cell biology and classics. *Word 69*(4): 362–379. https://doi.org/10.1080/00437956.2023.2270875

Leong, Alvin Ping (2025). Marked themes in academic writing: A comparative look at the sciences and humanities. *Text & Talk 45*(1): 33–53. https://doi.org/10.1515/text-2022-0188

Lingard, Lorelei, Chris Watling (2024). The writer's voice repertoire: Exploring how health researchers accomplish a distinctive 'voice' in their writing. *Medical Education 58*(5): 523–534. https://doi.org/10.1111/medu.15298

McLaughlin, G. Harry (1969). SMOG grading—A new readability formula. *Journal of Reading 12*(8): 639–646.

Michalikova, Martina (2022). Differences in academic writing & publishing between STEM and humanities. *Writing Scientist*. https://writingscientist.com/stem-humanities/

Miskiewicz, Michael, Salvatore Capotosto, Kenny Ling, Frederick Hance, Edward Wang (2024). Readability analysis of patient education material on rotator cuff injuries from the top 25 ranking orthopaedic institutions. *JAAOS: Global Research and Reviews 8*(5): e24.00085. https://doi.org/10.5435/JAAOSGlobal-D-24-00085

Mohammadi, Ehsan, Mike Thelwall, Stefanie Haustein, Vincent Larivière (2015). Who reads research articles? An altmetrics analysis of Mendeley user categories. *Journal of the Association for Information Science and Technology 66*(9): 1832–1846. https://doi.org/10.1002/asi.23286

O'Hayre, John (1966). *Gobbledygook has Gotta Go.* Washington, D.C.: U.S. Department of the Interior.

Okulicz-Kozaryn, Adam (2013). Cluttered writing: Adjectives and adverbs in academia. *Scientometrics 96*(3): 679–681. https://doi.org/10.1007/s11192-012-0937-9

Plavén-Sigray, Pontus, Granville James Matheson, Björn Christian Schiffler, William Hedley Thompson (2017). Research: The readability of scientific texts is decreasing over time. *eLife*, *6*: e27725. https://doi.org/10.7554/eLife.27725

Science Media Centre (2023). Summary report—Ipsos survey on science and the media. https://www.sciencemediacentre.org/wp-content/uploads/2023/11/PDF-summary-report.pdf

Scimago (2024). Scimago journal and country rank. http://www.scimagojr.com/

Scott, Brian (2024). Readability formulas. Retrieved June 3, 2025, https://readabilityformulas.com/

Singh Chawla, Dalmeet (2020). Science is getting harder to read. *Nature*. https://www.nature.com/nature-index/news/science-research-papers-getting-harder-to-read-acronyms-jargon

Smith, E. A., R. J. Senter (1967). *Automated Readability Index*. Dayton, OH: Aerospace Medical Research Laboratories.

Stremersch, Stefan, Isabel Verniers, Peter C. Verhoef (2007). The quest for citations: Drivers of article impact. *Journal of Marketing 71*(3): 171–193. https://doi.org/10.1509/jmkg.71.3.171

Stricker, Johannes, Anita Chasiotis, Martin Kerwer, Armin Günther (2020). Scientific abstracts and plain language summaries in psychology: A comparison based on readability indices. *PLOS One 15*(4): e0231160. https://doi.org/10.1371/journal.pone.0231160

Sun, Zhuanlan, Dongjin He, Yiwei Li (2024). How the readability of manuscript before journal submission advantages peer review process: Evidence from biomedical scientific publications. *Journal of Informetrics 18*(3): 101547. https://doi.org/10.1016/j.joi.2024.101547

Swanson, Karl, Shuhan He, Josh Calvano, David Chen, Talar Telvizian, Lawrence Jiang, Paul Chong, Jacob Schwell, Gin Mak, Jarone Lee (2024). Biomedical text readability after hypernym substitution with fine-tuned large language models. *PLOS Digital Health 3*(4): e0000489. https://doi.org/10.1371/journal.pdig.0000489

Uzunçıbuk, Hande, Maria Maddalena Marrapodi, Gamze Gökulu, Reem A. Alansari, Vincenzo Ronsivalle, Marco Cicciù, Giuseppe Minervini (2024). Assessment of the readability of online orthodontic educational resources related to 'orthodontic treatment': A cross-sectional study. *Seminars in Orthodontics 30*(4): 463–468. https://doi.org/10.1053/j.sodo.2024.01.015

Wen, Ju, Sike He, Lan Yi (2023). Easily readable? Examining the readability of lay summaries published in Autism Research. *Autism Research 16*(5): 935–940. https://doi.org/10.1002/aur.2917

Wen, Ju, Lan Yi (2023). Comparing lay summaries to scientific abstracts for readability and jargon use: a case report. *Scientometrics 128*(10): 5791–5800. https://doi.org/10.1007/s11192-023-04807-1

Woolston, Chris (2020). Words matter: Jargon alienates readers. *Nature 579*: 309. https://doi.org/10.1038/d41586-020-00580-w

Writing in the Humanities (n.d.). Wikibooks. https://human.libretexts.org/Bookshelves/Composition/Introductory_Composition/Rhetoric_and_Composition_(Wikibooks)/05%3A_Advanced_Topics/5.02%3A_Writing_in_the_Humanities

Yeung, Andy W. K., Tazuko K. Goto, W. Keung Leung (2018). Readability of the 100 most-cited neuroimaging papers assessed by common readability formulae. *Frontiers in Human Neuroscience 12*: 308. https://doi.org/10.3389/fnhum.2018.00308

Zaiontz, Charles (2025). Real statistics using Excel. http://www.real-statistics.com/

Author's address

Language and Communication Centre
School of Humanities, Nanyang Technological University
48 Nanyang Avenue, Singapore 639818, Singapore
e-mail: alvin.leong@ntu.edu.sg

Appendix

Abbreviations for the names of readability formulas

ARI	Automated readability index
ARLC	Average reading level consensus
CL	Coleman-Liau
FK	Flesch-Kincaid
FORCAST	Blend of the developers' names (Ford, Calyn, and Sticht)
FRE	Flesch reading ease
GF	Gunning Fog
LW	Lensear Write
NDC	New Dale-Chall
SMOG	Simple measure of gobbledygook