A response to Granberg *et al.* (2024)

Erin Hengel[*]

December 2024

Abstract

In this response to Granberg *et al.* (2024), I explain my reasons for using the Flesch Reading Ease, Flesch-Kincaid, Gunning Fog, SMOG and Dale-Chall readability scores in Hengel (2022). I also identify several errors in `textstat`, the program Granberg *et al.* (2024) use to calculate their scores. After correcting these errors, `textstat`'s gender readability gaps and *p*-values for the Flesch Reading Ease, Flesch-Kincaid, Gunning Fog and SMOG scores are very similar to those reported in Hengel (2022). `textstat` also generates Dale-Chall gender readability gaps that are 62–79 percent of the Dale-Chall estimate from Hengel (2022) and remaining variation is entirely due to differences in the familiar word lists used by each program (*e.g.*, `textstat` omits words such as "men's" and "women's" that are disproportionately found in female-authored papers). Meanwhile, Granberg *et al.* (2024)'s alternative scores either generate similarly-sized readability gaps or have been shown to be less powerful predictors of reading comprehension in adult reading material. I conclude by noting that Hengel (2022) neither relies on nor claims to rely on an assumption that readability scores predict scientific quality.

[*] Brunel University of London, erin.hengel@gmail.com.

1 Introduction

Granberg *et al.* (2024) replicate Hengel (2022) using 34 readability measures. They claim to "[uncover] that the measures used by Hengel were extraordinary in terms of their level of significance" (p. 10). In this response, I further explain my reasons for using the Flesch Reading Ease, Flesch-Kincaid, Gunning Fog, SMOG and Dale-Chall scores to proxy for text readability. I also identify several errors in `textstat`, the program Granberg *et al.* (2024) use to calculate their readability scores. Once these errors are corrected, both `textstat` and `Textatistic` (the program employed by Hengel (2022)) generate very similar gender readability gaps. The size and significance of the gaps reported in Hengel (2022) are also very similar to gaps estimated using comparable alternative scores from `textstat`.

As discussed in Section 2, Hengel (2022) includes the Flesch Reading Ease and Dale-Chall scores because they have repeatedly been shown to be the most powerful predictors of reading comprehension in adult reading material (see, *e.g.*, Carver 1974; Caylor 1973; Danielson and Bryan 1963; Farr *et al.* 1951; Harrison 1980; Klare 1963; Powers *et al.* 1958). I added the Flesch-Kincaid and Fog scores because both are popular in economics and finance research—indeed, the Fog index "is one of the most popular readability measures across all fields and it also appears to be the measure of choice in financial research in particular" (Loughran and McDonald 2016, p. 1645). For added robustness, I further included the SMOG score; in contrast to the other scores, it is entirely a function of vocabulary complexity, which has been shown to be one of the most important components for predicting text readability (see, *e.g.*, Chall and Dale 1995, pp. 61–65).

After identifying and correcting several errors in `textstat`'s methods and formulas (Section 3), I find that the gender gaps and p-values for the Flesch Reading Ease, Flesch-Kincaid, Gunning Fog and SMOG scores from Hengel (2022) are very similar to the gaps estimated using comparable scores produced by `textstat`—*i.e.*, scores that determine vocabulary difficulty by counting syllables and words with three or more syllables. `textstat` also generates Dale-Chall gender readability gaps that are 62–79 percent of the size of the Dale-Chall estimate from Hengel (2022). Remaining variation is due to differences in the familiar word lists used by each program—and especially the fact that `textstat` omits several words such as "men's" and "women's" that are disproportionately found in female-authored papers. When I reproduce these scores using a list that combines words from both programs, I find gender gaps that are almost identical to the gap reported in Hengel (2022).

As discussed in Section 2, the remaining scores analysed in Granberg *et al.* (2024) have been shown to be less powerful predictors of reading comprehension in adult reading material, especially in comparison to the Flesch Reading Ease and Dale-Chall formulas. For example, the Spache readability scores rely on a vocabulary complexity component that converges to a constant as text difficulty increases beyond the US eighth-grade level (Harrison 1980), the Wheeler-Smith formula was developed specifically for texts read by primary school-aged children (Wheeler and Smith 1954), and the Scrabble score is not a readability score at all (Benoit 2024). Because these scores are less predictive of readability in advanced texts, they introduce additional classical measurement error relative to the scores analysed in Hengel (2022). As a result, their corresponding gender gaps are generally smaller, although they are almost always in the same direction and frequently even statistically significant at traditional thresholds.

Granberg *et al.* (2024) concludes that "readability scores are poor predictors of scientific quality" (p. 8) based on low correlations between the 34 readability scores they analyse and article citations. While this may be the case, I fail to see the relevance it has to their paper, which they claim is a replication of the main results in Hengel (2022). Hengel (2022) only evaluates the relationship between author gender and abstract readability conditional on paper quality, as proxied for by citations. It does not rely on—nor does it claim to rely on—an assumption that readability predicts scientific quality.

The remainder of this paper proceeds in the following order. Section 2 discusses the reasons why Hengel

(2022) proxies for readability using the Flesch Reading Ease, Flesch-Kincaid, Gunning Fog, SMOG and Dale-Chall scores. In Section 3, I describe the differences between `textstat` and `Textstatistic`. I also identify several errors in `textstat` and one error in `Textstatistic`. Section 4 reproduces Hengel (2022, Table 3, columns 5 and 9) using the 34 readability scores presented in Granberg *et al.* (2024) and corrected for the errors identified in the previous section. Section 5 clarifies that Hengel (2022) does not rely on an assumption that readability predicts scientific quality. Section 6 concludes.

2 Choice of readability scores in Hengel (2022)

Hengel (2022) includes the Flesch Reading Ease and Dale-Chall scores because they have been repeatedly shown to be the most powerful predictors of reading comprehension in adult reading material (see, *e.g.*, Carver 1974; Caylor 1973; Danielson and Bryan 1963; Farr *et al.* 1951; Harrison 1980; Klare 1963; Powers *et al.* 1958). For example, Powers *et al.* (1958) compared the Dale-Chall, Flesch Reading Ease, Gunning Fog, and the Farr-Jenkins-Paterson scores. They concluded that the Dale-Chall and Flesch formulas were more powerful and precise than the other two formulas. Carver (1974) conducted a comprehensive analysis of numerous readability scores—including the RIDE, ARI, SMOG and FRY scores—and found "that the Flesch and Dale-Chall are the most valid readability formulas available" (p. 18).

Hengel (2022) additionally includes the Flesch-Kincaid and Fog scores. Both scores are popular in economics and finance research and have been used by policy-makers to identify unclear writing. For example, the U.S. Securities and Exchange Commission used both scores (in addition to the Flesch Reading Ease score) to benchmark readability in financial disclosure forms (Cox 2007). Loughran and McDonald (2016) review of the use of readability scores in finance and accounting research and conclude that the Fog index "is one of the most popular readability measures across all fields, and it also appears to be the measure of choice in financial research in particular" (p. 1645).[1]

For added robustness, Hengel (2022) also includes the SMOG score. Most classical readability indices consist of two measures: one capturing sentence structure and another capturing word structure. In contrast, the SMOG score is a function of vocabulary complexity alone, which has been shown to be the more important component for predicting text readability (see, *e.g.*, Chall and Dale 1995, pp. 61–65).

Table 1 lists economics and economics-adjacent research papers that analyse readability using a classical readability score and are cited in Hengel (2022, Appendix D1 p. 1) or Loughran and McDonald (2016). With the exception of Thörnqvist (2015) and Lewis *et al.* (1986), all papers analyse a subset of the scores included in Hengel (2022).

Although I never considered including in Hengel (2022) the additional scores analysed in Granberg *et al.* (2024), almost all either generate similarly-sized gender readability gaps (see Section 4) or are less powerful predictors of reading comprehension in adult reading material (especially in comparison to the Flesch Reading Ease and Dale-Chall formulas). In particular, the Spache readability scores rely on a vocabulary complexity component that converges to a constant as text difficulty increases beyond the US eighth-grade level (Harrison 1980); it therefore struggles to discriminate text difficulty in more advanced materials (Flesch (1943), as cited in Dale and Chall (1948)).[2] Similarly, simplified readability measures that proxy for vocabulary complexity using counts of monosyllabic words or letters per word—including the ARI, Danielson-Bryan and Coleman scores—have consistently proved less valid on advanced reading

[1] The Gunning Fog Index differs from the Flesch Reading Ease and Dale-Chall scores in that it proxies for vocabulary difficulty by counting the number of words with three or more syllables in a passage of text. Gunning (1968) justifies this decision as follows (p. 36). "Among the 1,000 words E. L. Thorndike, the noted educator, found to be used most often, only 36 are of more than two syllables. In Dale's list of 3,000 most familiar words, only one out of 25 is of more than three syllables. On the other hand, among words beyond the 20,000 most often used, two out of every three are of three syllables or more."

[2] Spache readability scores measure vocabulary complexity by counting words on the Dale (1931) list of 769 words well-known to children 6 years and younger.

Table 1: Readability scores used in research in economics and economics-adjacent fields

Study	Readability scores
Biddle *et al.* (2009)	Fog
Enke (2020)	Flesch Reading Ease
Jansen (2011)	Flesch Reading Ease, Flesch-Kincaid
Law and Zaring (2010)	Flesch-Kincaid
Lawrence (2013)	Fog
Lehavy *et al.* (2011)	Fog
Lewis *et al.* (1986)	Dale-Chall, Flesch Reading Ease, Fog, LIX, Kwolek, Fry Graph
Li (2008)	Fog
Loughran and McDonald (2014)	Fog
Lundholm *et al.* (2014)	Fog
Miller (2010)	Fog
Spirling (2016)	Flesch Reading Ease
Thörnqvist (2015)	LIX

Note. Economics and economics-adjacent research papers that analyse readability using a classical readability score and are cited in Hengel (2022, online appendix D.1, p. 1) or Loughran and McDonald (2016). Scores created by the authors of the cited paper are excluded—*e.g.*, in addition to analysing readability using the Fog index, Miller (2010) created and analysed a new readability measure he called the ”Plain English Index”.

materials compared to scores that match words to the Dale-Chall list or count syllables (see, *e.g.*, Carver 1974; Caylor 1973; Danielson and Bryan 1963; Farr *et al.* 1951; Harrison 1980; Powers *et al.* 1958).[3]

Finally, the Scrabble measure reported in Granberg *et al.* (2024) is not a readability score at all. This measure averages the Scrabble scores of all words in a passage of text; its authors caution that “There is no reference for this [score], as we created it experimentally. It’s not part of any accepted readability index!” (Benoit 2024).

3 Differences between `textstat` and `Textatistic`

Granberg *et al.* (2024) use the `textstat` program to calculate readability scores whereas Hengel (2022) relies on `Textatistic`. Although both programs generate roughly similar rankings when applying the same readability formula to a single passage of text, they rarely generate identical scores.[4] This is largely due to differences in how each program counts sentences, words, syllables and the number of words on the Dale-Chall list of familiar words. I describe these differences below.

When applied to the database analysed in Hengel (2022), `textstat` produces slightly less accurate sentence counts compared to `Textatistic`. Although sentence counts matched for the vast majority of observations (9,048 abstracts), for a small number they did not: for 15 abstracts, `Texatatistic` underestimated sentence counts because it failed to count sentences that ended with *etc.* or *et al.*; for 54 observations, `textstat` overestimated sentence counts—sometimes substantially—because it failed to

[3] Indeed, these measures were never intended to exceed the validity of the Dale-Chall and Flesch Reading Ease formulas in more advanced materials. The Wheeler-Smith formula was specifically created to evaluate texts read by primary school-aged children (Wheeler and Smith 1954); the other scores were developed to improve reliability when calculating scores by hand (Thomas *et al.* 1975) or to accommodate early computer programs which struggled to accurately count syllables (Coleman and Liau 1975; Danielson and Bryan 1963).

[4] For example, the Spearman coefficient of correlation between `textstat` and `Textatistic`’s calculation of the Flesch Reading Ease score is 0.92.

accurately identify sentence boundaries when sentences contained many non-alphanumeric characters or did not begin with a capitalised letter (*e.g.*, because the sentence was numbered).[5]

The word counts estimated by textstat and Textatistic differ for 15 percent of observations. Among them, the average word count is 130 for Textatistic and 132 for textstat. This small discrepancy appears to be largely due to differences in how each programme counts symbols (*e.g.*, "$") and number ranges that include hyphens (*e.g.*, "2015–2020").

Syllable-per-word counts are 12 percent higher in textstat than they are in Textatistic. This is due to underlying differences in the third-party libraries used by each program. textstat matches words to the CMU pronunciation dictionary; for words not in the dictionary, it estimates syllable counts using vowel clusters. In contrast, Textatistic counts syllables using the Python package PyHyphen. PyHyphen is based on the C library libhyphen, an implementation of the hyphenation algorithm from Liang (1983).[6]

There are also substantial differences in how Textatistic and textstat identify words on the Dale-Chall list of familiar words. First, Textatistic follows Chall and Dale (1995) and treats numbers in both digit (*e.g.*, "7") and written form (*e.g.*, "seven") as familiar words; in contrast, textstat counts all numbers in digit form as unfamiliar words. Second, the Dale-Chall word lists used by Textatistic and textstat do not perfectly overlap: 15–18 percent of words on one list are not on the other.[7] textstat contains a wider array of adverbs (*e.g.*, "roughly") whereas Textatistic contains more comparative and superlative adjectives (*e.g.*, "higher" and "highest") and irregular verb tenses (*e.g.*, "shown"). textstat also omits several words disproportionately found in female-authored papers, including "woman's", "man's", "women's", "men's", "wife's", "son's" and "whites".[8] To the best of my knowledge, these lists do not exhibit any other noticeable patterns.

Finally, textstat's formulas for the Bormuth GP, Bormuth MC, Fog NRI, Fog PSK and SMOG C scores contain errors. Table A.1 in Appendix A describes these errors and reproduces the corrected formulas.

4 Replicating Granberg *et al.* (2024) with corrected textstat

Tables 2 and 3 reproduce Hengel (2022, Table 3, columns 5 and 9) using the 34 readability scores presented in Granberg *et al.* (2024, Tables A2 and A4). Dale-Chall scores in columns (2)–(3) are calculated using textstat's Dale-Chall word list; scores in columns (4)–(5) combine textstat's and Textatistic's Dale-Chall lists. To calculate each readability score, I use the formulas applied by textstat, corrected to account for the errors identified in Appendix A. Counts of characters, words, syllables, *etc.* are also from textstat with the following two exceptions: (i) I manually adjusted sentence counts to correct for the errors described in Section 3; (ii) following Chall and Dale (1995), I also count numbers in digit form as familiar words.

Dale-Chall gender readability gaps range between 0.054–0.060 in Table 2 and 0.109–0.128 in Table 3 (columns (2)–(3)). These estimates are 62–79 percent of the size of the Dale-Chall estimate from Hengel (2022) as reported in standardised form in Granberg *et al.* (2024).[9] The remaining variation is due to

[5] textstat uses a modified version of the function str_split_boundaries from the R package stringi to split strings at sentence boundaries. This function appears to be sensitive to non-alphanumeric characters and instances where sentences do not begin with a capitalised letter.

[6] Liang (1983)'s algorithm is used by TEX's typesetting system and most open source text processing software (including OpenOffice) to identify where to hyphenate words when split over two lines. Using hyphen locations to identify syllables provides a highly accurate and consistently calculated estimate of syllable counts even for words not included on the CMU word list.

[7] Textatistic and textstat contain 8,490 and 8,133 unique words, respectively. 1,545 words on Textatistic's list are not on textstat's and 1,188 words on textstat's list are not on Textatistic's. Counts were determined after words were converted to lower-case and all fullstops, hyphens and apostrophes were removed.

[8] Bizarrely, textstat includes "husband's" and "daughter's". (Textatistic includes all of these words.)

[9] As Granberg *et al.* (2024) does not report a standardised estimate of the readability gap corresponding to Table 3, column (9) in Hengel (2022), I estimated it myself by running the exact same specification but using a standardised measure

Table 2: Replicating Granberg *et al.* (2024), Table A2 with corrected `textstat`

	Dale-Chall words from `textstat`		Dale-Chall words from `textstat` + `Textatistic`		
Readability score	Coefficient on female	Standard error	Coefficient on female	Standard error	*N*
More sophisticated measures of vocabulary difficulty					
Counts of words on the Dale-Chall list					
Dale-Chall	0.0601	(0.0446)	0.0772	(0.0464)	9,117
Dale-Chall (old)	0.0537	(0.0467)	0.0738	(0.0492)	9,117
Dale-Chall PSK	0.0581	(0.0453)	0.0769	(0.0475)	9,117
Syllable counts					
Flesch	0.1091**	(0.0414)			9,117
Flesch-Kincaid	0.0913**	(0.0440)			9,117
Flesch PSK	0.1068**	(0.0415)			9,117
Strain	0.0639	(0.0479)			9,117
Counts of words with 3+ syllables					
Fog	0.1189***	(0.0401)			9,117
Fog NRI	0.0729	(0.0459)			9,117
Fog PSK	0.1233***	(0.0396)			9,117
Linsear Write	0.1168***	(0.0395)			9,117
nWS	0.0819*	(0.0428)			9,117
nWS 2	0.0884**	(0.0411)			9,117
nWS 3	0.1330***	(0.0390)			9,117
nWS 4	0.1202***	(0.0399)			9,117
SMOG	0.1163***	(0.0433)			9,117
SMOG C	0.1163***	(0.0433)			9,117
Cruder measures of vocabulary difficulty					
Letter counts					
ARI	0.0453	(0.0437)			9,117
Bormuth MC	0.0270	(0.0247)			9,117
Bormuth GP	0.0105	(0.0086)			9,117
Coleman-Liau ECP	0.0185	(0.0438)			9,117
Danielson-Bryan	0.0488	(0.0412)			9,117
Dickes-Steiwer	0.0487	(0.0463)			9,117
Fucks	0.0450	(0.0477)			9,117
Tränkle-Bailer	0.0439	(0.0460)			9,117
Counts of words with 7+ letters					
LIX	0.0786*	(0.0399)			9,117
RIX	0.0764*	(0.0389)			9,117
Counts of monosyllabic words					
Coleman	−0.0213	(0.0668)			9,117
Coleman C2	−0.0169	(0.0670)			9,117
Farr-Jenkins-Paterson	0.0066	(0.0558)			9,117
Wheeler-Smith	0.0253	(0.0473)			9,117
Counts of words on the Dale (1931) list					
Spache	0.0332	(0.0496)			9,117
Spache (old)	0.0342	(0.0498)			9,117
Non-readability measures					
Scrabble	−0.0698	(0.0659)			9,117

Note. Table reproduces Hengel (2022, Table 3, column 5) using the 34 readability scores presented in Granberg *et al.* (2024, Table A2). Readability scores are calculated using counts and formulas from `textstat`, corrected to account for the errors identified in Section 3 and Appendix A.

Table 3: Replicating Granberg *et al.* (2024), Table A3 with corrected `textstat`

	Dale-Chall words from `textstat`		Dale-Chall words from `textstat` + `Textatistic`		
Readability score	Coefficient on female	Standard error	Coefficient on female	Standard error	N
More sophisticated measures of vocabulary difficulty					
Counts of words on the Dale-Chall list					
Dale-Chall	0.1276**	(0.0563)	0.1557**	(0.0582)	5,774
Dale-Chall (old)	0.1093*	(0.0584)	0.1425**	(0.0603)	5,774
Dale-Chall PSK	0.1211**	(0.0577)	0.1520**	(0.0597)	5,774
Syllable counts					
Flesch	0.1391***	(0.0506)			5,774
Flesch-Kincaid	0.1415***	(0.0466)			5,774
Flesch PSK	0.1449***	(0.0494)			5,774
Strain	0.1209**	(0.0450)			5,774
Counts of words with 3+ syllables					
Fog	0.1398***	(0.0476)			5,774
Fog NRI	0.1170**	(0.0449)			5,774
Fog PSK	0.1404***	(0.0478)			5,774
Linsear Write	0.1362***	(0.0459)			5,774
nWS	0.1195**	(0.0538)			5,774
nWS 2	0.1208**	(0.0522)			5,774
nWS 3	0.1393***	(0.0483)			5,774
nWS 4	0.1400***	(0.0477)			5,774
SMOG	0.1347**	(0.0510)			5,774
SMOG C	0.1347**	(0.0510)			5,774
Cruder measures of vocabulary difficulty					
Letter counts					
ARI	0.1071**	(0.0437)			5,774
Bormuth MC	0.0296	(0.0254)			5,774
Bormuth GP	0.0008	(0.0005)			5,774
Coleman-Liau ECP	0.0422	(0.0566)			5,774
Danielson-Bryan	0.1085**	(0.0459)			5,774
Dickes-Steiwer	0.1117**	(0.0427)			5,774
Fucks	0.1063**	(0.0438)			5,774
Tränkle-Bailer	0.1053**	(0.0431)			5,774
Counts of words with 7+ letters					
LIX	0.1465***	(0.0513)			5,774
RIX	0.1389***	(0.0474)			5,774
Counts of monosyllabic words					
Coleman	0.0478	(0.0694)			5,774
Coleman C2	0.0627	(0.0709)			5,774
Farr-Jenkins-Paterson	0.1041	(0.0640)			5,774
Wheeler-Smith	0.1081**	(0.0506)			5,774
Counts of words on the Dale (1931) list					
Spache	0.1049**	(0.0502)			5,774
Spache (old)	0.1051**	(0.0497)			5,774
Non-readability measures					
Scrabble	−0.0293	(0.0691)			5,774

Note. Table reproduces Hengel (2022, Table 3, column 9) using the 34 readability scores presented in Granberg *et al.* (2024, Table A3). Readability scores are calculated using counts and formulas from `textstat`, corrected to account for the errors identified in Section 3 and Appendix A.

differences in the Dale-Chall familiar word lists used by `textstat` and `Textstatistic`: when I reproduce the three Dale-Chall scores using a list that combines words from both programs (columns (4)–(5) of Tables 2 and 3), I find a gender difference that is 86–97 percent of the coefficient size from Hengel (2022); *p*-values are likewise very similar.

The second and third sets of readability scores shown in Tables 2 and 3 determine vocabulary difficulty by counting syllables and words with three or more syllables. In Table 2, all but the Strain and Fog NRI indices are significant at traditional thresholds; in Table 3, all scores are highly significant. The average effect size and *p*-value in Table 2 are 0.104 and 0.035; in Table 3, they are 0.133 and 0.011.[10] For comparison, the average effect sizes for the Flesch, Flesch-Kincaid, Gunning Fog and SMOG scores reported in Hengel (2022) are 0.116 and 0.140; average *p*-values are 0.032 and 0.011.

As discussed in Section 2, the scores in the second panels of Tables 2 and 3 have generally been shown to be less powerful predictors of reading comprehension in adult reading material (particularly in relation to the Flesch Reading Ease and Dale-Chall scores). Among measures that determine vocabulary difficulty by counting letters or words on the Dale (1931) list of familiar words, gender gaps are always positive.[11] While they are smaller and more noisily estimated in Table 2, they are larger and generally statistically significant at traditional thresholds in Table 3. Gender gaps are smallest (and even negative in two instances in Table 2) among measures that determine vocabulary difficulty by counting monosyllabic words.

In contrast, the gender gap is always positive and significant when the LIX or RIX readability scores are used as dependent variables. Both scores gauge vocabulary difficulty by counting words with seven or more letters in a passage of text. Compared to letter counts, monosyllabic word counts and counts of words on the Dale (1931) list, this may be a more precise proxy for word complexity in advanced reading materials. As a result, gender gap estimates produced by the LIX and RIX scores may contain less classical measurement error relative to gaps produced by the other scores in the second panels of Tables 2 and 3.

The final panels in Tables 2 and 3 show the gender gap in the average Scrabble score among abstracts analysed in Hengel (2022). These results suggest that abstracts in papers authored by women have a slightly lower average Scrabble score compared to abstracts authored by men. Because Scrabble rewards long words and words spelled with uncommon letters—*e.g.*, X and Q—this negative result is arguably consistent with Hengel (2022)'s general conclusion that female-authored abstracts published in top economics journals are more readable. Nevertheless, it should be interpreted cautiously: not only is the gender gap noisily estimated, but the Scrabble score was created experimentally by `textstat`'s authors (Benoit 2024) and is not a readability score at all (see Section 2).

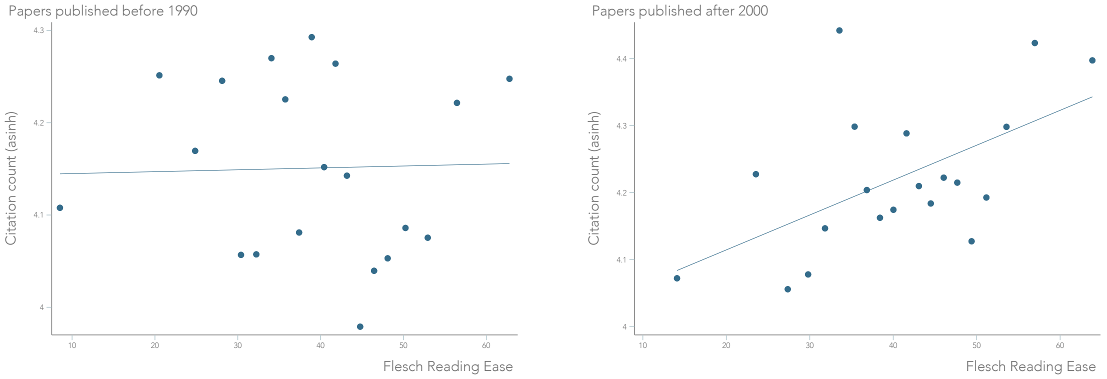
Table B.1 (Appendix B) replicates Table 3 but excludes papers published in the *American Economic Review Papers & Proceedings* (563 observations). On average, gender gaps are smaller (and less significant) than those shown in Table 3 but larger (and more significant) than those from Table 2. In Appendix C, I replicate Tables 2, 3 and B.1 using word, character, syllable, *etc.* counts from `Textatistic`. Gender gaps from both programs are very similar.

of the Dale-Chall score from Hengel (2022) as the dependent variable. The coefficient on female ratio from that regression is 0.16 (standard error 0.06).

[10] If the Flesch, Flesch-Kincaid, Fog and SMOG scores are omitted, the average effect size and *p*-value are 0.102 and 0.043 for Table 2 and 0.131 and 0.012 for Table 3.

[11] In contrast to most other formulas, the Bormuth MP score combines two measures of vocabulary difficulty: one counting the number of words not on the Dale-Chall list of familiar words and a second counting the number of characters per words (Bormuth 1969). Because the former term—which is always a number between zero and one (and on average is 0.39 in the data analysed in Hengel (2022))—is cubed, vocabulary complexity is almost entirely determined by the latter term. (The average weighted value of the count of characters per words is 0.2 whereas the average weighted value of the number of words not on the Dale-Chall list of familiar words is 0.01.) Additionally, the Bormuth GP figure should be viewed with caution—its average value among the abstracts analysed in Hengel (2022) is absurdly high (see Appendix A for further details).

Figure 1: Bottom half of Figure D.1 in Hengel (2022)



Notes. These two graphics are reproduced from Hengel (2022, online appendix D.1, Figure D.1, p. 2). They plot abstracts' Flesch Reading Ease scores against their articles' (asinh) citation counts for the samples of papers analysed in Hengel (2022) (excluding papers published in the *AER Papers & Proceedings*). The left–hand graph includes only articles published before 1990; the right-hand graph includes only articles published after 2000.

5 The correlation between readability and scientific quality

Hengel (2022) evaluates the relationship between author gender and abstract *readability* conditional on paper "quality", as proxied for by citations. It does not rely on—nor does it claim to rely on—an assumption that readability predicts scientific quality.[12] Nevertheless, I did receive several seminar queries about this relationship, so I included in Hengel (2022)'s online appendix D.1 two graphs plotting the Flesch score against (asinh) citations: one for the sample of papers published before 1990 and another for the sample of papers published between 2000–2015. (To conserve space, I omitted the graph for papers published between these two periods.) I reproduce these graphs in Figure 1.

Figure 1 makes clear that the data analysed in Hengel (2022) are ambiguous about the relationship between readability and citations. I further emphasise this ambiguity when referencing Figure 1 in the text (Hengel 2022, online appendix D.1 p. 1):

> *Evidence from other studies linking readability and citations is, however, weaker (Berniger et al., 2017; Laband and Taylor, 1992; Lei and Yan, 2016). My own data suggest a positive relationship in papers published after 1990—and particularly those published post-2000—but no relationship before that (Figure D.1).*

While I was pleased to learn that Granberg *et al.* (2024) are exploring the existence and robustness of a relationship between readability and scientific quality, it is not clear to me how or why this is relevant to their paper, which claims to be a replication of the main results in Hengel (2022).

6 Conclusions

Granberg *et al.* (2024) replicate Hengel (2022) using 34 readability measures from the program `textstat`. In this response, I explain my reasons for using the Flesch Reading Ease, Flesch-Kincaid, Gunning Fog, SMOG and Dale-Chall scores to proxy for text readability. I also identify several errors in the methods and formulas that `textstat` uses to calculate its scores. After correcting these errors, I find that the gender gaps and p-values for the Flesch Reading Ease, Flesch-Kincaid, Gunning Fog and SMOG scores from `Textatistic` (the program used in Hengel (2022)) are very similar to the gaps produced

[12] In contrast, Hengel *et al.* (2024) evaluates the relationship between gender and citations among publications in economics and finance journals. Their results—which rely on the assumption that (asinh) citations adequately proxy for scientific quality—suggest that female-authored papers are higher quality than male-authored papers, conditional on publication.

by comparable scores from `textstat`. `textstat` also generates Dale-Chall gender readability gaps that are 62–79 percent of the size of the Dale-Chall estimate from `Textatistic`; all remaining variation is entirely due to differences in the familiar word lists used by each program. Meanwhile, the additional scores analysed in Granberg *et al.* (2024) either generate similarly-sized gender readability gaps or are less powerful predictors of reading comprehension in adult reading material.

Granberg *et al.* (2024, p. 8) conclude by suggesting that "readability scores are poor predictors of scientific quality". I fail to understand how this is relevant to a replication of Hengel (2022). Hengel (2022) evaluates the relationship between author gender and abstract readability conditional on paper quality, as proxied for by citations. It does not rely on—nor does it claim to rely on—an assumption that readability predicts scientific quality.

References

Benoit, Kenneth (2024). *Calculate readability — `textstat_readability`*. Date accessed: 2024-10-22. URL: https://quanteda.io/reference/textstat_readability.html.

Biddle, Gary C., Gilles Hilary, and Rodrigo S. Verdi (2009). "How does financial reporting quality relate to investment efficiency?" *Journal of Accounting and Economics* 48 (2-3), pp. 112–131.

Bormuth, John R. (1969). *Development of readability analysis*. Tech. rep. ED029166. Department of Health, Education, and Welfare.

Carver, Ronald P. (1974). *Improving reading comprehension: measuring readability*. Tech. rep. AIR-30801-5/74-FR. American Institutes for Research.

Caylor, John S. (1973). *Development of a simple readability index for job reading materials*. Tech. rep. EC 076 707. Human Resources Research Organization.

Chall, Jeanne S. and Edgar Dale (1995). *Readability Revisited*. Brookline Books.

Coleman, Meri and T. L. Liau (1975). "A computer readability formula designed for machine scoring". *Journal of Applied Psychology* 60 (2), pp. 283–284.

Cox, Christopher (2007). *Closing remarks to the second annual corporate governance summit*. Delivered at USC Marshall School of Business, Los Angeles, California, 23 March.

Dale, Edgar (1931). "A comparison of two word lists". *Educational Research Bulletin* 10 (18), pp. 484–489.

Dale, Edgar and Jeanne S. Chall (1948). "A formula for predicting readability". *Educational Research Bulletin* 27 (1), pp. 11–20.

Danielson, Wayne A. and Sam Dunn Bryan (1963). "Computer automation of two readability formulas". *Journalism and Mass Communication Quarterly* 40 (2), pp. 201–206.

Enke, Benjamin (2020). "Moral values and voting: Trump and beyond". *Journal of Political Economy* 128 (10), pp. 3679–3729.

Farr, James N., James J. Jenkins, and Donald G. Paterson (1951). "Simplification of Flesch Reading Ease formula". *Journal of Applied Psychology* 35 (5), pp. 333–337.

Flesch, Rudolf (1943). "Marks of readable style". PhD thesis. New York Teachers College Columbia University.

Granberg, Mark, Joakim Jansson, and Yifan Yang (2024). "A comment on "Publishing while female" by Hengel (2022)". Mimeo.

Gunning, Robert (1968). *The Technique of Clear Writing*. 2nd ed. New York: McGraw Hill.

Harrison, Colin (1980). *Readability in the classroom*. Cambridge University Press.

Hengel, Erin (2022). "Publishing while female. Are women held to higher standards? Evidence from peer review." *The Economic Journal* 132 (648), pp. 2951–2991.

Hengel, Erin, Euyoung Moon, and Richard Tol (2024). "Gender and equality at economics and finance journals". Mimeo.

Jansen, David Jan (2011). "Does the clarity of central bank communication affect volatility in financial markets? Evidence from Humphrey-Hawkins testimonies". *Contemporary Economic Policy* 29 (4), pp. 494–509.

Klare George, R. (1963). *The Measurement of Readability*. Iowa State University Press.

Law, David S. and David Zaring (2010). "Law versus ideology: the Supreme Court and the use of legislative history". *William and Mary Law Review* 51 (5), pp. 1653–1747.

Lawrence, Alastair (2013). "Individual investors and financial disclosure". *Journal of Accounting and Economics* 56 (1), pp. 130–147.

Lehavy, Reuven, Feng Li, and Kenneth Merkley (2011). "The effect of annual report readability on analyst following and the properties of their earnings forecasts". *Accounting Review* 86 (3), pp. 1087–1115.

Lewis, N. R. *et al.* (1986). “Accounting report readability: the use of readability techniques”. *Accounting and Business Research* 16 (63), pp. 199–213.

Li, Feng (2008). “Annual report readability, current earnings, and earnings persistence”. *Journal of Accounting and Economics* 45 (2-3), pp. 221–247.

Liang, Franklin Mark (1983). “Word hy-phen-a-tion by com-put-er”. PhD thesis. Stanford University.

Loughran, Tim and McDonald (2014). “Measuring readability in financial disclosures”. *Journal of Finance* 69 (4), pp. 1643–1671.

Loughran, Tim and Bill McDonald (2016). “Textual analysis in accounting and finance: a survey”. *Journal of Accounting Research* 54 (4), pp. 1187–1230.

Lundholm, Russell J., Rafael Rogo, and Zhang Jenny Li (2014). “Restoring the Tower of Babel: how foreign firms communicate with U.S. investors”. *The Accounting Review* 89 (4), pp. 1453–1485.

Miller, Brian P. (2010). “The effects of reporting complexity on small and large investor trading”. *Accounting Review* 85 (6), pp. 2107–2143.

Powers, R. D., W. A. Sumner, and B. E. Kearl (1958). “A recalculation of four adult readability formulas”. *Journal of Educational Psychology* 49 (2).

Spirling, Arthur (2016). “Democratization and linguistic complexity: the effect of franchise extension on parliamentary discourse, 1832–1915”. *Journal of Politics* 78 (1), pp. 120–136.

Thomas, Georgelle, R. Derald Hartley, and J. Peter Kincaid (1975). “Test-retest and inter-analyst reliability of the Automated Readability Index, Flesch Reading Ease score and the Fog count”. *Journal of Reading Behavior* 7 (2), pp. 149–154.

Thörnqvist, Tomas (2015). “Sophistication, news and individual investor trading”. Mimeo.

Wheeler, Lester R. and H. Smith Edwin (1954). “A practical readability formula for the classroom teacher in the primary grades”. *Elementary English* 31 (7), pp. 397–399.

Appendices

A Errors in textstat formulas

Table A.1 describes errors in textstat's calculation of five scores: the Bormuth GP, the Bormuth MC, the Fog NRI, the Fog PSK and the SMOG C. The second column reproduces the corrected formula, and the third column describes the original error. The source for the correction is listed in the final column.

I additionally note that the Bormuth GP score should be used with caution. Because several of its terms are squares and cubes of figures that are not necessarily between 0 and 1 in magnitude, it can produce difficult to interpret results. For example, the average Bormuth GP score among the abstracts analysed in Hengel (2022) is 491,625. If interpreted literally, this suggests that the average abstract in that sample requires almost 500,000 years of schooling to understand.

Table A.1: Errors in `textstat` formulas

Score	Corrected formula	Error in `textstat`	Source
Bormuth MC	$0.886593 - 0.083640 \times \frac{\text{no. characters}}{\text{no. words}} + 0.161911 \times \left(\frac{\text{no. Dale-Chall words}}{\text{no. words}}\right)^3 - 0.021401 \times \frac{\text{no. words}}{\text{no. sentences}} + 0.000577 \times \left(\frac{\text{no. words}}{\text{no. sentences}}\right)^2 - 0.000005 \times \left(\frac{\text{no. words}}{\text{no. sentences}}\right)^3$	`textstat` replaced the coefficient on the fourth term with −0.21401 (it should be −0.021401).	Bormuth (1969, Table 15, p. 162)
Bormuth GP	$4.275 + 12.881 \times \text{Bormuth MC} - 34.934 \times \text{Bormuth MC}^2 + 20.388 \times \text{Bormuth MC}^3 + 26.194 \times \text{Cloze} - 2.046 \times \text{Cloze}^2 - 11.767 \times \text{Cloze}^3 - 44.285\,(\text{Bormuth MC} \times \text{Cloze}) + 97.620\,(\text{Bormuth MC} \times \text{Cloze})^2 - 59.538\,(\text{Bormuth MC} \times \text{Cloze})^3$	`textstat` replaced the Cloze Criterion Score with character counts in the eighth term of the equation.	Bormuth (1969, Figure 2, p. 170)
Fog NRI	$\frac{1}{2} \times \left(\frac{\text{no. 1–2-syllable words} + 3 \times \text{no. 3+-syllable words}}{\text{no. sentences}} - 3\right)$	`textstat` incorrectly divided the number of sentences by the number of words.	Kincaid *et al.* (1975, p. 14)
Fog PSK	$3.0680 + 0.0877 \times \frac{\text{no. words}}{\text{no. sentences}} + 0.0984 \times 100 \times \frac{\text{no. 3+-syllable words}}{\text{no. words}}$	`textstat` multiplied the number of words per sentences figure by 3.0680; this terms should be added.	Powers *et al.* (1958, p. 101)
SMOG C	$0.9986 \times \sqrt{30 \times \frac{\text{no. 3+-syllable words}}{\text{no. sentences}}} + 5 + 2.8795$	`textstat` added 5 in the square-root term; it should be added outside of it.	McLaughlin (1969, p. 643)

B Robustness

Table B.1 replicates Table 3 but excludes papers published in the *Papers and Proceedings* issue of the *American Economic Review* (*AER P&P*). On average, gender gaps are smaller (and less significant) than those shown in Table 3 but larger (and more significant) than those from Table 2, which does not control for tertiary *JEL* codes and includes all articles in the database analysed in Hengel (2022).[1]

[1] The samples analysed in Tables 3 and B.1 only includes abstracts published after 1990, when *JEL* codes were substantially revised.

Table B.1: Replicating Granberg *et al.* (2024), Table A3 with corrected `textstat`, excluding *P&P*

	Dale-Chall words from `textstat`		Dale-Chall words from `textstat` + `Textatistic`		
Readability score	Coefficient on female	Standard error	Coefficient on female	Standard error	*N*
More sophisticated measures of vocabulary difficulty					
Counts of words on the Dale-Chall list					
Dale-Chall	0.0817	(0.0598)	0.1112*	(0.0621)	5,211
Dale-Chall (old)	0.0554	(0.0593)	0.0903	(0.0620)	5,211
Dale-Chall PSK	0.0706	(0.0601)	0.1031	(0.0626)	5,211
Syllable counts					
Flesch	0.0795	(0.0475)			5,211
Flesch-Kincaid	0.1035**	(0.0464)			5,211
Flesch PSK	0.0907*	(0.0471)			5,211
Strain	0.1035**	(0.0471)			5,211
Counts of words with 3+ syllables					
Fog	0.0989**	(0.0448)			5,211
Fog NRI	0.1000**	(0.0460)			5,211
Fog PSK	0.0968**	(0.0446)			5,211
Linsear Write	0.0949**	(0.0445)			5,211
nWS	0.0495	(0.0484)			5,211
nWS 2	0.0517	(0.0467)			5,211
nWS 3	0.0889*	(0.0439)			5,211
nWS 4	0.0983**	(0.0447)			5,211
SMOG	0.0962*	(0.0502)			5,211
SMOG C	0.0962*	(0.0502)			5,211
Cruder measures of vocabulary difficulty					
Letter counts					
ARI	0.0742*	(0.0438)			5,211
Bormuth MC	0.0091	(0.0257)			5,211
Bormuth GP	0.0008	(0.0006)			5,211
Coleman-Liau ECP	−0.0206	(0.0529)			5,211
Danielson-Bryan	0.0633	(0.0446)			5,211
Dickes-Steiwer	0.0902**	(0.0441)			5,211
Fucks	0.0901*	(0.0460)			5,211
Tränkle-Bailer	0.0827*	(0.0443)			5,211
Counts of words with 7+ letters					
LIX	0.1024**	(0.0488)			5,211
RIX	0.1070**	(0.0469)			5,211
Counts of monosyllabic words					
Coleman	−0.0121	(0.0669)			5,211
Coleman C2	0.0044	(0.0698)			5,211
Farr-Jenkins-Paterson	0.0505	(0.0617)			5,211
Wheeler-Smith	0.0833	(0.0520)			5,211
Counts of words on the Dale (1931) list					
Spache	0.0663	(0.0548)			5,211
Spache (old)	0.0696	(0.0544)			5,211
Non-readability measures					
Scrabble	−0.0629	(0.0805)			5,211

Note. Table replicates Table 3 but excludes papers published in the *AER P&P*.

C Calculating readability using `Textatistic`

In Tables C.1, C.2 and C.3, I replicate Tables 2, 3 and B.1 (respectively) using word, character, syllable, *etc.* counts from `Textatistic`, the program used to generate readability scores in Hengel (2022). The resulting gender gaps are very similar to the gaps produced by `textstat`.

Table C.1: Replicating Granberg *et al.* (2024), Table A2 with corrected `Textatistic`

	Dale-Chall words from `textstat`		Dale-Chall words from `textstat` + `Textatistic`		
Readability score	Coefficient on female	Standard error	Coefficient on female	Standard error	*N*
More sophisticated measures of vocabulary difficulty					
Counts of words on the Dale-Chall list					
Dale-Chall	0.0620	(0.0444)	0.0789*	(0.0462)	9,117
Dale-Chall (old)	0.0562	(0.0459)	0.0760	(0.0486)	9,117
Dale-Chall PSK	0.0603	(0.0449)	0.0788*	(0.0472)	9,117
Syllable counts					
Flesch	0.1233***	(0.0434)			9,117
Flesch-Kincaid	0.0951**	(0.0444)			9,117
Flesch PSK	0.1174***	(0.0431)			9,117
Strain	0.0665	(0.0467)			9,117
Counts of words with 3+ syllables					
Fog	0.1302***	(0.0449)			9,117
Fog NRI	0.0759	(0.0472)			9,117
Fog PSK	0.1360***	(0.0446)			9,117
Linsear Write	0.1278***	(0.0425)			9,117
nWS	0.1006**	(0.0435)			9,117
nWS 2	0.1016**	(0.0423)			9,117
nWS 3	0.1497***	(0.0441)			9,117
nWS 4	0.1319***	(0.0448)			9,117
SMOG	0.1326***	(0.0474)			9,117
SMOG C	0.1326***	(0.0474)			9,117
Cruder measures of vocabulary difficulty					
Letter counts					
ARI	0.0494	(0.0434)			9,117
Bormuth MC	0.0316	(0.0244)			9,117
Bormuth GP	0.0106	(0.0087)			9,117
Coleman-Liau ECP	0.0276	(0.0416)			9,117
Danielson-Bryan	0.0531	(0.0416)			9,117
Dickes-Steiwer	0.0511	(0.0457)			9,117
Fucks	0.0444	(0.0476)			9,117
Tränkle-Bailer	0.0468	(0.0456)			9,117
Counts of words with 7+ letters					
LIX	0.0809**	(0.0400)			9,117
RIX	0.0771*	(0.0394)			9,117
Counts of monosyllabic words					
Coleman	0.0511	(0.0633)			9,117
Coleman C2	0.0556	(0.0627)			9,117
Farr-Jenkins-Paterson	0.0691	(0.0522)			9,117
Wheeler-Smith	0.0594	(0.0461)			9,117
Counts of words on the Dale (1931) list					
Spache	0.0346	(0.0496)			9,117
Spache (old)	0.0354	(0.0497)			9,117

Note. Table replicates Table 2 but uses count data from `Textatistic` to calculate readability scores.

Table C.2: Replicating Granberg *et al.* (2024), Table A3 with corrected `Textatistic`

	Dale-Chall words from `textstat`		Dale-Chall words from `textstat` + `Textatistic`		
Readability score	Coefficient on female	Standard error	Coefficient on female	Standard error	*N*
More sophisticated measures of vocabulary difficulty					
Counts of words on the Dale-Chall list					
Dale-Chall	0.1277**	(0.0562)	0.1556**	(0.0581)	5,774
Dale-Chall (old)	0.1095*	(0.0576)	0.1424**	(0.0595)	5,774
Dale-Chall PSK	0.1212**	(0.0573)	0.1519**	(0.0592)	5,774
Syllable counts					
Flesch	0.1291**	(0.0535)			5,774
Flesch-Kincaid	0.1306**	(0.0492)			5,774
Flesch PSK	0.1342**	(0.0526)			5,774
Strain	0.1159**	(0.0458)			5,774
Counts of words with 3+ syllables					
Fog	0.1531***	(0.0493)			5,774
Fog NRI	0.1208**	(0.0464)			5,774
Fog PSK	0.1552***	(0.0494)			5,774
Linsear Write	0.1502***	(0.0483)			5,774
nWS	0.1359**	(0.0551)			5,774
nWS 2	0.1370**	(0.0529)			5,774
nWS 3	0.1581***	(0.0494)			5,774
nWS 4	0.1538***	(0.0493)			5,774
SMOG	0.1492***	(0.0539)			5,774
SMOG C	0.1492***	(0.0539)			5,774
Cruder measures of vocabulary difficulty					
Letter counts					
ARI	0.1050**	(0.0437)			5,774
Bormuth MC	0.0279	(0.0245)			5,774
Bormuth GP	0.0008	(0.0005)			5,774
Coleman-Liau ECP	0.0398	(0.0534)			5,774
Danielson-Bryan	0.1064**	(0.0454)			5,774
Dickes-Steiwer	0.1102**	(0.0426)			5,774
Fucks	0.1046**	(0.0434)			5,774
Tränkle-Bailer	0.1037**	(0.0432)			5,774
Counts of words with 7+ letters					
LIX	0.1468***	(0.0515)			5,774
RIX	0.1388***	(0.0476)			5,774
Counts of monosyllabic words					
Coleman	0.0695	(0.0717)			5,774
Coleman C2	0.0834	(0.0737)			5,774
Farr-Jenkins-Paterson	0.1197*	(0.0666)			5,774
Wheeler-Smith	0.1165**	(0.0528)			5,774
Counts of words on the Dale (1931) list					
Spache	0.1050**	(0.0507)			5,774
Spache (old)	0.1053**	(0.0501)			5,774

Note. Table replicates Table 3 but uses count data from `Textatistic` to calculate readability scores.

Table C.3: Replicating Granberg *et al.* (2024), Table A3 with corrected `Textatistic`, excluding *P&P*

	Dale-Chall words from `textstat`		Dale-Chall words from `textstat` + `Textatistic`		
Readability score	Coefficient on female	Standard error	Coefficient on female	Standard error	*N*
More sophisticated measures of vocabulary difficulty					
Counts of words on the Dale-Chall list					
Dale-Chall	0.0825	(0.0601)	0.1118*	(0.0623)	5,211
Dale-Chall (old)	0.0568	(0.0589)	0.0915	(0.0616)	5,211
Dale-Chall PSK	0.0716	(0.0601)	0.1040	(0.0626)	5,211
Syllable counts					
Flesch	0.0789	(0.0516)			5,211
Flesch-Kincaid	0.1004*	(0.0499)			5,211
Flesch PSK	0.0895*	(0.0516)			5,211
Strain	0.1009**	(0.0477)			5,211
Counts of words with 3+ syllables					
Fog	0.1270**	(0.0493)			5,211
Fog NRI	0.1096**	(0.0481)			5,211
Fog PSK	0.1273**	(0.0492)			5,211
Linsear Write	0.1172**	(0.0479)			5,211
nWS	0.0747	(0.0502)			5,211
nWS 2	0.0774	(0.0483)			5,211
nWS 3	0.1256**	(0.0486)			5,211
nWS 4	0.1271**	(0.0493)			5,211
SMOG	0.1213**	(0.0557)			5,211
SMOG C	0.1213**	(0.0557)			5,211
Cruder measures of vocabulary difficulty					
Letter counts					
ARI	0.0710	(0.0439)			5,211
Bormuth MC	0.0067	(0.0246)			5,211
Bormuth GP	0.0008	(0.0007)			5,211
Coleman-Liau ECP	−0.0246	(0.0484)			5,211
Danielson-Bryan	0.0612	(0.0444)			5,211
Dickes-Steiwer	0.0878*	(0.0441)			5,211
Fucks	0.0885*	(0.0457)			5,211
Tränkle-Bailer	0.0802*	(0.0444)			5,211
Counts of words with 7+ letters					
LIX	0.1008**	(0.0491)			5,211
RIX	0.1057**	(0.0473)			5,211
Counts of monosyllabic words					
Coleman	0.0063	(0.0674)			5,211
Coleman C2	0.0222	(0.0706)			5,211
Farr-Jenkins-Paterson	0.0632	(0.0630)			5,211
Wheeler-Smith	0.0880	(0.0534)			5,211
Counts of words on the Dale (1931) list					
Spache	0.0665	(0.0555)			5,211
Spache (old)	0.0697	(0.0549)			5,211

Note. Table replicates Table B.1 but uses count data from `Textatistic` to calculate readability scores.

References

Bormuth, John R. (1969). *Development of readability analysis.* Tech. rep. ED029166. Department of Health, Education, and Welfare.

Dale, Edgar (1931). “A comparison of two word lists”. *Educational Research Bulletin* 10 (18), pp. 484–489.

Granberg, Mark, Joakim Jansson, and Yifan Yang (2024). “A comment on “Publishing while female” by Hengel (2022)”. Mimeo.

Hengel, Erin (2022). “Publishing while female. Are women held to higher standards? Evidence from peer review.” *The Economic Journal* 132 (648), pp. 2951–2991.

Kincaid, J. Peter *et al.* (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel.* Tech. rep. 8-75. Institute for Simulation and Training.

McLaughlin, G. Harry (1969). “SMOG grading—a new readability formula”. *Journal of Reading* 12 (8), pp. 639–646.

Powers, R. D., W. A. Sumner, and B. E. Kearl (1958). “A recalculation of four adult readability formulas”. *Journal of Educational Psychology* 49 (2).