International Academic Conference on Teaching, Learning and Education
T L E C O N F

Amsterdam, Netherlands
23- 25 June 2023

6th International Academic Conference
on Teaching, Learning and Education

Comparing The Readability Level of Biology Exams Texts in Malta

Marthese Azzopardi[1], and Carmel Azzopardi[2]
Department of Biology, University of Malta Junior College, Msida MSD1252, Malta

Abstract

We used a quantitative approach to evaluate the readability level of comprehension test passages in high-stakes, advanced-level biology exams in Malta. We studied 37 passages from a public institution and 54 from national exams from the years 1996 to 2022. Using an online computational tool called Coh-Metrix we calculated three readability formulas, Flesch Reading Ease, Flesch-Kincaid Grade Level, and Coh-Metrix L2 Reading Index. All three indices indicated that all texts were "difficult" and understood by English-speaking undergraduates; however, for Maltese students, English is a second language (L2). We computed ANOVA to analyse differences in readability indices in the two high-stakes exams. The results revealed no significant difference in the readability of the texts, implying that the institution exposes students to the same level of difficulty as the national average. The Pearson correlation coefficient revealed that both exams maintained the same degree of readability throughout the study period. We found that the average word count and sentence length were significantly shorter in the public institution exams. Lexical diversity is an important indicator of how complex and difficult a text is to read. It measures how many different words occur in the text. The two indices of lexical diversity used showed that the texts of the institution were less complicated and difficult to read. Thus, in a way, paper setters at the institution attempt to match difficulty to the students' level, keeping in mind that they have been learning the subject for one year. This study benefits paper-setters by allowing them to evaluate and match the readability of the comprehension test passages with 17- to 18-year-old L2 students.

Keywords: Coh-Metrix; corpus; L2 Reading Index; lexical diversity; high-stakes exams

1. Introduction

Teachers are faced with the challenge of supplying students with reading material that is at the appropriate level of difficulty. If the reading material is too easy, students are not challenged. On the other hand, if students are presented with material that is too difficult, they fail to make progress and may become so frustrated that they simply give up. To make the match, the readability level of materials must be known.

Readability measures how easy it is to read a text. It refers to how comprehensive the written content is. Readability is determined by the complexity of sentence structure and the vocabulary used (Bailin & Grafstein, 2016). It is tightly related to reading comprehension, retention, reading speed, and persistence (DuBay, 2004). Readability formulas have been used since the 1920s by educators and researchers to predict the difficulty level of a selected text (Flesch, 1948; Kincaid et al., 1975; DuBay, 2004). Giving a historical review of readability is not the intention of this study. DuBay (2004) gives an extensive review of research that spans 100 years. The readability formulas use variables that are known to be among the first causes of reading difficulties, such as the average length of sentences, the number of new words contained, and the grammatical complexity of the language used in a passage. While there are many other features of language they do not measure, the readability formula scores correlate .90 and above with comprehension as measured by reading tests (DuBay, 2004).

Otlhomile et al. (2018) argue that text readability is an important criterion that guides teachers in the selection of texts and assigning them to different groups in class. Likewise, examiners must choose lengthy passages intended for high-stakes exams while keeping in mind the proficiency of the students in English. To the knowledge of the authors, no studies have been done to estimate the readability of biology texts in high-stakes exams in Malta. This study tries to fill this gap by discovering whether the difficulty of a text matches the students' education grade level. It also attempts to compare the readability of comprehension test passages in a public institution exam taken after learning biology for one year and those in the national exam that is taken after a two-year course.

In this study, a computational tool called Coh-Metrix, developed by Arthur C. Graesser and Danielle McNamara in 2004, was used. It "analyses texts on multiple measures of language and discourse that are aligned with multilevel theoretical frameworks of comprehension" (Graesser et al., 2014). The Coh-Metrix website can be accessed for free at www.cohmetrix.com. Coh-Metrix gives a total of 108 indices, classified into eleven categories. Table 1 shows the seven indices belonging to three categories that were computed in this study.

Table 1: Coh-Metrix 3.0 indices selected for this study

Aspect	Analysed Variable	Coh-Metrix Index	Label
Descriptive	Basic counts	Number of words	DESWC
		Average sentence length	DESSL
Lexical Diversity	Variety of unique words	Measure of textual lexical diversity	MTLD
		Probabilities of word occurrence	VOCD
Readability	Method of assessing text difficulty	Flesch Reading Ease	FRE
		Flesch-Kincaid Grade Level	FKGL
		Second-language readability score	L2

Source: www.cohmetrix.com

1.1 Readability Formulas

Readability formulas have survived 100 years of intensive application, investigation, and controversy, irrespective of their limitations. The calculation of readability indices has become easier with the availability of automated tools. In this study, the readability of texts was calculated using three different readability formulas. The Flesch Reading Ease Score (hereafter, FRE) (Flesch, 1948) and Flesch-Kincaid Grade Level (hereafter, FKGL) (Kincaid et al., 1975) are a unidimensional representation of text difficulty. They were developed for first-language (L1) texts and are the traditional readability measures that focus on the number of words and sentences. These readability indices are robust predictors of sentence-level understanding (McNamara et al., 2014). The FRE rates texts from 0 to 100, where the higher the score, the easier the text is (Table 2). The formula (McNamara et al., 2014) is provided below:

FRE = 206.835 – (1.015 x number of words/number of sentences) – (84.600 x number of syllables/number of words)

Table 2: Flesch Reading Ease (FRE) and interpretation

FRE	Readability Level	Education Level
0-29	Very confusing	College Graduates
30-49	Difficult	College
50-59	Fairly Difficult	High School Senior
60-69	Standard	13 to 15 year-olds
70-79	Fairly Easy	12 year-olds
80-89	Easy	11 year-olds
90-100	Very Easy	10 year-olds

Source: http://heidicohen.com/linkedin-publishing/

FKGL converts FRE to a grade-school level, where the higher the number, the harder it is to read the text (Table 3). The formula (McNamara et al., 2014) is provided below:

RDFKL = (0.39 x number of words/number of sentences) + (11.8 x number of syllables/number of words) – 15.59

Table 3: Flesch-Kincaid Grade Level (FKGL) and interpretation

FKGL Score	School Level	Comprehension
5.0-5.9	5th Grade	Very easy to read
6.0-6.9	6th Grade	Easy to read
7.0-7.9	7th Grade	Fairly easy to read
8.0-8.9	8th & 9th Grade	Conversational English
10.0-12.9	10th, 11th & 12th Grade	Fairly difficult to read
13.0-15.9	College	Difficult to read
16.0-17.9	College Graduate	Very difficult to read
18+	Professional	Extremely difficult to read

Source: https://www.textcompare.org/readability/flesch-kincaid-grade-level

The third index calculated, Coh-Metrix L2 Readability Index (hereafter, L2), is a multidimensional representation of text difficulty. It was developed for second language (L2) learners by McNamara, Graesser, McCarthy, and Cai (2014). It can be computed by the Coh-Metrix online tool, which can be freely accessed at http://cohmetrix.memphis.edu/cohmetrixhome/. It incorporates three factors: CELEX Word Frequency (logarithm mean for content words), Sentence Syntax Similarity (sentence-to-sentence adjacent mean), and Content Word Overlap (proportional adjacent sentences unweighted). This index using the mentioned three variables significantly differs from traditional readability formulas, which mostly consider just formal aspects of a text, like the average length of a word or the average length of a sentence, without any respect to the content of a text (Kiselnikov et al., 2020). Table 4 shows the L2 mean and standard deviation norms that correspond to the U.S. education grade level (McNamara et al., 2014). The higher the score, the easier the text is to read. Its formula is:

L2 = – 45.032 + (52.230 x Content Word Overlap Value) + (61.306 x Sentence Syntax Similarity Value) + (22.205 x CELEX Frequency Value)

These three readability formulas were chosen from the numerous ones available as they are the most frequently used and best known in readability analysis. One reason for choosing the L2 Index is that it is the latest and most comprehensive readability tool. The number of researchers using Coh-Metrix along with other classic readability formulae is on the increase. A quick search on Google Scholar (30 December 2022) revealed a dramatic rise between 2003 and 2004 (24 results) and between 2021 and 2022 (1,140 results).

1.1.1 Lexical Diversity (LD)

LD provides information regarding the variety of unique word types that appear in the text and the overall number of words. It indicates the complexity and difficulty of a text. Since the literature shows that LD indices are affected by text length, care was taken in the selection of the index. The Measure of Textual Lexical Diversity (MTLD) was computed as it is least affected by text length (Zenker & Kyle, 2021). McCarthy & Jarvis (2010) advise researchers to consider using MTLD, VOCD, and Maas rather than any single index. VOCD is a lexical diversity measurement instrument based on the probability of new vocabulary being introduced into longer and longer samples of text and has been developed as an attempt to overcome text length. Hence, VOCD was calculated as well. Table 4 shows the mean MTLD, VOCD, and standard deviation norms that correspond to the grade level (McNamara et al., 2014).

Table 4: The mean, and standard deviation norms for Coh-Metrix L2 Readability Index (L2), Measure of Textual Lexical Diversity (MTLD), and Probabilities of word occurrence (VOCD) based on science texts and the corresponding grade level

Grade Level	L2		MTLD		VOCD	
	Mean	SD	Mean	SD	Mean	SD
K-1	32.462	7.265	42.003	14.061	58.811	17.974
2-3	28.470	5.953	47.259	13.335	63.544	16.570
4-5	25.014	5.553	50.508	14.491	67.023	16.656
6-8	20.866	5.245	60.219	20.326	73.977	22.106
9-10	18.776	5.122	64.567	22.525	74.525	20.588
11 +	15.066	5.066	67.781	23.384	76.040	21.928

Source: McNamara et al., 2014 p. 266;268

2. Research Aims and Questions

This study aims to find the readability level of the comprehension test passages used in two high-stakes advanced biology exams in Malta. Another aim is to compare the readability level and complexity of the passages in the two high-stakes exams to establish whether students at a public institution are prepared for the national exams. Finally, the study may establish whether the level of difficulty obtained by the traditional readability formulas differs from that of L2.

This investigation aspires to answer the following three research questions for each exam type:

1. What are the readability level and lexical diversity of the comprehension test passages for the May, September, and combined sessions?
2. Is there a change in the seven indices from the years 1996 to 2022?
3. What is the correlation between the readability indices?

3. Methodology

3.1 The Corpus

The corpus was comprised of 91 comprehension test passages that appeared in the Biology advanced-level exams between 1996 and 2022. The source of the passages was as follows: 54 from Malta's national exams (27 May sessions and 27 September sessions) and 37 from those of a public institution (20 May sessions and 17 September sessions).

3.1.1 Data Analysis

In this investigation, Coh-Metrix 3.0 was used to get the values for the seven indices selected. The SPSS (Statistical Package for the Social Sciences) version 24 was used to compute the Kolmogorov-Smirnov test to check for normal distribution. As the data was found to be normally distributed, ANOVA was computed to analyse differences between groups. Pearson correlation coefficient formula was used to find if the 7 indices varied over the years 1996 to 2022, and to compute the strength and direction of a linear relationship between:

1) Flesch-Kincaid Grade Level and Flesch Reading Ease;

2) Flesch-Kincaid Grade Level and Coh-Metrix L2 Readability Index;

3) Flesch Reading Ease and Coh-Metrix L2 Readability Index.

A *p* value of .05 was considered to indicate statistical significance.

4. Results

4.1 Descriptive statistics

Results from Table 5 show that the mean number of words (DESWC) was less (370.595 ± 84.467) for the public institution exams (n = 37) compared to the national (n = 54) (455.741 ± 108.653). The same applied to the mean number of words per sentence (DESSL:18.648 ± 3.833) for the institution and national ones (DESSL: 21.147 ± 3.403). Table 6 shows the ANOVA results conducted using the seven Coh-Metrix measures as the dependent variables and the exams type as the independent variable. Results showed that DESWC and DESSL were significantly different at $p = .05$.

Results from Tables 5 and 6 show that the mean number of words (DESWC) for the May session texts was significantly lower ($p = .003$) (372.1 ± 68.245) for the public institution exams (n = 20) compared to the national (n = 27) (464.481 ± 120.07). This was also the case for the mean number of words per sentence, DESSL:18.227 ± 2.757 for the institution and national ones, DESSL: 21.537 ± 3.498) ($p = .001$).

The mean number of words (DESWC) for the September session texts was significantly lower (p = .015) (368.824 ± 102.548) for the public institution exams (n = 17) compared to the national (n = 27) (447.000 ± 97.414). However, the mean number of words per sentence (DESSL: 19.142 ± 4.852) for the institution and the national ones (DESSL: 20.757 ± 3.25) was not significantly different (p = .197).

Table 5: Descriptive statistics, lexical diversity, and readability level for institution and national exams for the May, September, and combined sessions

Exams		**DESWC**	**DESSL**	**MTLD**	**VOCD**	**FRE**	**FKGL**	**L2**
Institution All Sessions (n = 37)	Min	166	13.095	49.781	57.577	16.470	8.444	1.235
	Max	550	28.684	159.113	163.092	57.850	18.088	18.914
	Mean	370.595	18.648	86.281	93.26	41.515	12.101	9.475
	SD	84.467	3.833	29.205	27.551	11.409	2.343	4.242
Institution May Sessions (n = 20)	Min	222	13.875	49.781	60.894	20.730	8.792	3.645
	Max	523	23.389	131.926	145.66	57.740	15.851	17.948
	Mean	372.1	18.227	81.338	89.28	42.932	11.799	9.790
	SD	68.245	2.757	22.225	23.006	10.620	1.918	3.745
Institution September Sessions (n = 17)	Min	166	13.095	49.906	57.577	16.470	8.444	1.235
	Max	550	28.684	159.113	163.092	57.850	18.088	18.914
	Mean	368.824	19.142	92.096	97.943	39.849	12.457	9.104
	SD	102.548	4.852	35.581	32.191	12.388	2.781	4.855
National All Sessions (n = 54)	Min	312	11.79	0.67	0.51	19.960	8.100	1.610
	Max	836	30	145.11	158.41	61.780	17.930	16.470
	Mean	455.741	21.147	102.626	104.179	41.611	12.708	8.647
	SD	108.653	3.403	23.632	22.815	9.808	1.955	3.412
National May Sessions (n = 27)	Min	312	15.586	67.964	69.716	19.960	9.42	2.20
	Max	836	30	145.112	158.413	61.780	17.93	15.98
	Mean	464.481	21.537	103.175	106.013	41.704	12.793	9.068
	SD	120.07	3.498	21.086	20.807	10.378	2.113	3.452
National September Sessions (n = 27)	Min	330	11.79	0.61	0.51	23.93	8.10	1.61
	Max	681	27.31	144.67	133.250	57.99	15.74	16.47
	Mean	447.000	20.757	102.077	102.346	41.519	12.623	8.226
	SD	97.414	3.325	26.329	24.923	9.400	1.188	3.384

Table 6: ANOVA results for the seven indices computed for the institution and national exams

Index	All sessions				May Session				September Session			
	Sum of Squares	df	F	Sig.	Sum of Squares	df	F	Sig.	Sum of Squares	df	F	Sig.
DESWC	159178.381	1	16.052	.000*	98054.098	1	9.523	.003*	63754.689	1	6.453	.015*
DESSL	137.117	1	10.681	.002*	125.832	1	12.243	.001*	27.190	1	1.720	.197
MTLD	5865.334	1	8.656	.004*	5478.660	1	11.771	.001*	1038.958	1	1.140	.292
VOCD	2617.809	1	4.243	.042*	3216.990	1	6.792	.012*	202.234	1	.260	.613
FRE	.201	1	.002	.966	17.346	1	.158	.693	29.095	1	.257	.615
FKGL	8.088	1	1.799	.183	11.340	1	2.743	.105	.291	1	.058	.810
L2	15.049	1	1.059	.306	5.988	1	.468	.498	8.047	1	.501	.483

**p < .05*

Results in Table 7 provide the answer to the second question of this study: Is there a change in the seven indices from the years 1996 to 2022? A Pearson correlation coefficient was performed to evaluate the relationship between the year and the seven indices calculated for the institution and national exams. Results indicate that the seven indices have no significant relationship with the year.

Table 7: Pearson Correlation Coefficient (r) between the year and the seven indices over the period 1999-2022

Type		DESWC	DESSL	MTLD	VOCD	FRE	FKGL	L2
Institution (n = 37)	Pearson Correlation	.374*	.321	.344*	.380*	-.239	.292	-.046
	Sig. (2-tailed)	.023	.053	.037	.020	.155	.079	.785
National (n = 54)	Pearson Correlation	-.097	-.237	-.127	-.073	-.052	-.067	.158
	Sig. (2-tailed)	.483	.085	.360	.600	.711	.633	.253

**p < .05*

4.1.1 Lexical Diversity

The mean MTLD and VOCD shown in Table 5 were higher in the national exams compared to those at the institution, irrespective of whether the May, September or combined sessions were considered. These differences were significant at p = .05 when tested with ANOVA. Passages were more lexically diverse in national exams. All the mean values computed for MTLD and VOCD are equivalent for undergraduates (Table 4).

4.1.1. Readability Values

Irrespective of the exam type and session, the comprehension test passages were classified as "difficult" by all three readability indices. Thus, they are understood by 17- and 18-year-old L1 students. The mean FRE readability values shown in Table 5 range between 39.849 and 42.932; the FKGL readability values range from 11.799 to 12.793, and the L2 values range from 8.226 to 9.475. The readability index values were not significantly different (Table 6).

The Pearson correlation coefficient (*r*) was computed to answer the third question imposed in this study: What is the correlation between the readability indices? Results in Table 8 show that there was a significant negative relationship between FRE and FKGL, $r = -.954$ (n = 37), $p = .01$ for the institution, and $r = -.931$ (n = 54), $p = .01$ for the national texts. The negative relationship is because for FRE, the higher the value, the easier the text is, but for FKGL, the inverse applies. The Pearson correlation coefficient between L2 and each of the traditional indices is weaker when compared to that obtained for the traditional indices, FRE and FKGL (Table 8).

Table 8: Pearson Correlation Coefficient (r) between the readability indices (Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL), and Coh-Metrix L2 Readability Index (L2)) for all comprehension test passages

	Pearson Correlation Coefficient (*r*)	
Readability Indices	**Institution (n = 37)**	**National (n = 54)**
FRE and FKGL	-.954**	-.931**
FRE and L2	.588**	.307*
FKGL and L2	-.610**	-.335*

$p = .01$ **
$p = .05$ *

Discussion and Conclusions

The main aim of this study was to determine the readability of comprehension text passages in two high-stakes biology exams in Malta. All three readability indices indicated that the passages used were "difficult" and understood by college L1 students. The relatively high level of difficulty in readability is understandable when identifying the sources of the comprehension test passages. Texts were extracted from very technical books such as "*Facets of Maltese Prehistory"* in the case of the national exams in 2020 or scientific journals such as *Marine Biology*, *Nature Science News, Scientific American,* and *The Royal Society.* Plavén-Sigray et al. (2017) investigated 709,577 abstracts published between 1881 and 2015 from 123 biomedical and life sciences journals and demonstrated a steady decrease in the readability of science. A similar result was obtained more recently by Graf-Vlachy (2022), who sampled 28,345 abstracts from 17 leading journals over 30 years. This trend is

partially explained by an increase in scientific jargon. Given this, a recommendation that stems from the present study is that examiners should check whether a passage is appropriate linguistically for L2 students and, as Sung et al. (2015) explain if it is geared to their performance. Lee & Lee (2020) argue that automatic readability assessment allows for the fast selection of texts for L2 students by quantifying the level of difficulty that a reader experiences in comprehending a certain text. The next recommendation is to simplify the text to make it more suitable for L2 students. When the passages in the exams were compared to their online sources, it was noted that simplification was attempted at times by omitting sentences rather than sentence restructuring. Goodman & Freeman (1993) argue that the process of simplification will increase the reader's ability to understand and interact with a text.

The developers of Coh-Metrix, Graesser & McNamara (2014), suggest that the text analysed by the online tool, should be more than 100 words but less than 1000. Texts that are less than 100 words might be troublesome for a variety of reasons since the texts are unlikely to have completely established their range of cohesion values. As the texts investigated in this study ranged from 166 to 681 words, it may be concluded that the values computed by Coh-Metrix are reliable.

Crossley et al. (2012) indicated that lexical diversity and word frequency are important factors influencing the text difficulty experienced by L2 learners. Studies by Koizumi (2012) confirmed that MTLD has enough sensitivity to distinguish texts of different degrees of lexical diversity. Thus, a significant result obtained with MTLD values in this study is a clear indication that the passages at the institution are less lexically diverse and hence less complex than those at the national level when considering the combined and May sessions. This is greatly appropriate, as students sitting for the institution exams have been learning biology for one year, compared to two years for the national exams. McCarthy & Jarvis (2007) state that recently, VOCD has become the virtual tool of the lexical diversity trade. In this investigation, VOCD gave the same results as MTLD.

According to the Coh-Metrix creators, Crossley et al. (2008), L2, which is based on the lexis and syntax of a text, is more accurate than the standard readability formula, which is only based on word and sentence length. Crossley, Allen, and McNamara (2011) suggest that L2 be used rather than the traditional readability formulas, suggesting that the variables used in this index are more closely aligned with the intuitive text processing employed by authors when simplifying texts. Kiselnikov et al., 2020 reported a strong negative Pearson correlation (r = -0.889) between FRE and FKGL when analysing the readability of the textbook "English for Engineers" used in Russian technical universities but less strong between each of the traditional indices and L2. The same result was obtained in this study and may be due to different factors incorporated in the traditional indices compared to those in L2. In fact, although all three indices determined that texts were "difficult", some of them could be

classified as "very difficult" by L2. For example, the most difficult passage (institution, September 2002; source: *Biological Control of Pests and Weeds*, *M. J.* Samways) scored a value of 1.24 by the readability index L2 when the norm for L1 college students is 15.066 ± 5.066.

The main conclusion is that the comprehension test passages used in the institution exams have the same readability as those in national ones but are significantly shorter, the sentence length is shorter, and they are less complex.

Finally, some suggestions for future studies that would make up for the limitations in this study include the following: First, since in this study only three categories out of the 11 given by Coh-Metrix were used, other aspects of discourse, such as reference cohesion and semantic cohesion, could be included. In this study, the difficulty level of texts was assessed quantitatively, and it would be interesting to combine a qualitative method by asking students and teachers to judge the comprehension test passages.

References

Bailin, A. and Grafstein, A. (2016). *Readability: Text and context*. Palgrave Macmillan London.

Crossley, S.A., Allen, D.B. and McNamara, D.S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, *23*(1), pp. 84-101. Available: https://ejournals.um.edu.mt/login?url=https://www-proquest-com.ejournals.um.edu.mt/scholarly-journals/text-readability-intuitive-simplification/docview/881471674/se-2

Crossley, S.A., Allen, D. and McNamara, D.S. (2012). Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, *16*(1), pp. 89-108. Available: doi:https://doi-org.ejournals.um.edu.mt/10.1177/1362168811423456

DuBay, W.H. (2004). *The Principles of Readability*. Impact Information, Costa Mesa, CA.

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

Goodman, K. and Freeman, D. (1993). What's simple in simplified language. In: M.L. Tickoo (Ed.), Simplification: Theory and application (pp. 69–76). Singapore: SEAMEO Regional Language Center.

Graf-Vlachy, L. (2022). Is the readability of abstracts decreasing in management research? *Rev Manag Sci*, vol. 16, pp. 1063-1084. https://doi.org/10.1007/s11846-021-00468-7

Kincaid, J.P., Fishburne, R.P., Rogers, R.L. and Chissom, B.S. (1975). *Derivation of new readability formulas (Automated Readability Index. Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel.* Available: http://www.dtic.mil/dtic/tr/fulltext/u2/a006655.pdf

Kiselnikov, A., Vakhitova, D. and Kazymova, T. (2020). Coh-metrix readability formulas for an academic text analysis. *IOP Conf. Ser.: Mater. Sci. Eng.* 890 012207 doi 10.1088/1757-899X/890/1/012207

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary learning and Instruction*, 1(1), pp. 60-69.

Lee, B.W. and Lee, J. (2020). LXPER index 2.0: Improving text readability assessment model for L2 English students in Korea. In *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pp. 20–24, Suzhou, China. Association for Computational Linguistics.

McCarthy, P.M. and Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing, 24*(4), pp. 459-488. doi:https://doi.org/10.1177/0265532207080767

McCarthy, P.M. and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), pp. 381-392. doi:10.3758/BRM.42.2.381

McNamara, D.S., Graesser, A.C., McCarthy, P.M. and Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix* (Cambridge: Cambridge University Press).

Plavén-Sigray, P., Matheson, G.J., Schiffler, B.C. and Thompson, W.H. (2017). The readability of scientific texts is decreasing over time. *eLife*, 6 Available: https://www.readcube.com/articles/10. 7554/eLife.27725

Sung, Y.T., Dyson, S. B., Chen, Y.C., Lin, W.C. and Chang, K.E. (2015). Leveling L2 Texts Through Readability: Combining Multilevel Linguistic Features with the CEFR. *The Modern Language Journal*, *99* (2), pp. 371–391. Available: http://www.jstor.org/stable/43650032

Zenker, F. and Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices, *Assessing Writing*, Vol. 47, 100505, ISSN 1075-2935, https://doi.org/10.1016/j.asw.2020.100505. (https://www.sciencedirect.com/science/article/pii/S1075293520300660)