Assessing Readability Formulas in the Wild

Scott Crossley[1[0000-0002-5148-0273]] Stephen Skalicky[2[0000-0001-8329-4531]] Cynthia Berger[3[0000-0001-7876-3997]] Ali Heidari[1[0000-0003-2123-7616]]

[1] Georgia State University
[2] Victoria University of Wellington
[3] Duolingo
scrossley@gsu.edu
scskalicky@gmail.com
cindymayberger@gmail.com
heidari.ali2011@gmail.com

Abstract. Recent advances have facilitated major improvements in developing intelligent and purpose-oriented readability formulas to predict the overall difficulty of a text in terms of text comprehension and processing. Such readability formulas are mediating technologies that help match appropriate reading texts with students, thus enabling the development of smart learning environments that adapt learning resources to learner skills. Newer readability formulas include linguistic features that are more predictive of human judgments of text readability than traditional readability formulas, such as Flesch-Kincaid Grade Level. However, in many cases these formulas have not been tested beyond their ability to predict reading scores. The purpose of this study is to examine the validity of newer readability models along with more traditional readability formulas using behavioral data and text comprehension scores. The results indicate that readability models employing linguistic features more theoretically related to text processing and comprehension outperform readability models that do not employee similar features. The findings support the long-term growth of readability formulas that are continuously improved to increase the wellbeing of learners.

Keywords: Text readability, Linguistics, Natural language processing

1 Introduction

In the United States, students perform below average on standardized reading tests regardless of grade level (U. S. Department of Education, 2020). This performance minimizes opportunities for future student success and lowers wellbeing within communities ranging from the social to the academic (Powell, 2009). The primary reason for low reading success rates is the inherent difficulty of developing reading skills (National Assessment of Educational Progress, 2011). This difficulty can be mediated using a number number of strategies to help learners develop stronger reading skills (McNamara, Levinstein, & Boonthum, 2004). Chief among these is the careful

selection of texts to ensure that readers have challenging texts that are comprehensible (Wolfe et al., 1998).

A common technique to match readers with appropriate texts has been the use of readability formulas meant to access text difficulty, which in return ensures a well-balanced and personalized experience for learners. Since the 1940s, over 200 readability models have been created, indicating a long-term vision to help students succeed through better text matching (Benjamin, 2012). Traditional readability formulas developed through the 1940s to the 1980s like the Flesch-Kincaid Grade Level (Kincaid, Fishburne, Rogers, & Chissom, 1975) are based on surface level textual features that can be hand counted. These features include the number of letters (or syllables) per word to approximate lexical sophistication and the number of words in a sentence to approximate syntactic complexity.

Newer formulas developed after the advent of desktop computing are more intelligent and purpose oriented, allowing for automatic assessment of text difficulty using deeper linguistic features often inspired by theories of reading. These features allow newer readability formulas to better assess elements of reading including decoding (i.e., word recognition). Traditional readability formulas, which rely on the number of letters or syllables per word to measure decoding do not tap directly into the linguistic components of readability (Crossley et al., 2008). Newer formulas, however, can compute the frequency of words; lexical properties of words including concreteness, imageability, and age of acquisition; psycholinguistic norms including word naming and lexical decision times; and phonological neighborhood effects (among many others) that are more strongly related to decoding. Similar examples exist in how newer and traditional formulas calculate syntactic complexity. Chiefly, traditional readability formulas examine sentence length while newer readability formulas measure phrasal and clausal complexity. Additionally, newer formulas calculate features related to text cohesion and semanticity, which are not measured by traditional formulas.

Even though traditional readability formulas are not very smart, they have been widely adopted by publishers, researchers, primary and secondary schools, universities, the military, and testing agencies where they are used to select reading materials for a variety of learners (DuBay, 2004). The purpose of this study is to compare newly developed and more intelligent readability formulas to traditional readability formulas using behavioral data (i.e., text processing data) and reading comprehension scores. The goal is to better understand how these formulas perform in comparison to one another and how they perform when co-varied with individual difference variables (e.g., reading skills, reading confidence, number of books read), study design (e.g., order of texts read), and demographic information (e.g., age and gender).

2 Method

2.1 Participants

Sixty undergraduate English-speaking students from a southeastern public research university participated in a reading experiment. The participants were recruited from undergraduate linguistics courses. Demographic information for the participants was

collected using a self-reported online survey. Complete data were collected from 54 participants and included in the present study (43 female, 10 male, and one participant who declined to choose). The participants' average age was 23 (SD = 7.01, min = 18, max = 51). Participants self-identified as either monolingual English speakers or bilinguals (seven participants) who spoke additional languages including German, Spanish, Vietnamese, and Portuguese. All the participants had normal to corrected vision. Since the study was advertised as an ordinary reading study, all the participants were naïve about the purpose of the study. Participants received $20 Amazon gift cards in compensation for participating in the experiment.

2.2 Materials

Questionnaire. An online questionnaire was used to collect reader literacy and demographic information including reading habits, reading confidence, reading enjoyment, exposure to TV programs, and demographic information including age, gender, knowledge of second languages, and vision quality.

Reading Comprehension Scores. Reading comprehension ability for the participants was assessed using the Gates-MacGinitie (4th ed.) reading comprehension test (form S) level 10/12 (MacGinitie & MacGinitie, Cooter & Curry, 1989). The reading comprehension test included 48 multiple-choice questions that measured comprehension of short passages. Each passage was followed by two to six questions which measured reading comprehension competence involved in both surface and deeper level comprehension processes such as inference, text recall, and main ideas. The test involved standard instructions, two practice questions, and the comprehension items. Participants were allowed 25 minutes to complete the test.

Corpus. Twelve texts from a previous study assessing the development of new text readability formulas (Crossley et al., 2019) were used in the experiment. The texts included six texts from Simple English Wikipedia and six texts from regular Wikipedia. Simplified and regular Wikipedia texts were selected to provide variation in the difficulty of the texts. Two simplified and two regular texts were selected from each of three topic domains: history, technology, and science. The average length of the texts was 159 words (SD = 27.6). Three multiple choice comprehension questions were developed for each text. Each set of comprehension questions contained a question related to making inferences from the content, one related to text recall, and one related to the main idea of the text.

Readability Formulas. We selected three traditional readability measures: Flesch Reading Ease (Flesch, 1948), Flesch Kincaid Grade Level (Kincaid et al., 1975), and the FOG Index (Gunning, 1952) to assess reading speed and comprehension. All traditional readability formulas assess syntactic complexity (through sentence length) and lexical sophistication (through word length).

We selected three NLP inspired readability formulas for comparison. First, we selected the New Dale-Chall readability formula (Chall & Dale, 1995), which includes both traditional and newer measures of text difficulty. The formula measures syntactic complexity using a traditional approach (sentence length), but it measures lexical sophistication using a list of the 3,000 most frequent words in English. We also selected two newer readability formulas: Crowdsourced Algorithm of REading Comprehension (CAREC) and the Crowdsourced Algorithm of REading Speed (CARES, Crossley et al., 2019). CAREC measures 13 language variables related to lexical sophistication, n-gram features, text cohesion, and sentiment. CARES predicts judgments of reading speed using NLP features related to number and types of words, sophisticated words, syntactic complexity, and variation in paragraph size. Crossley et al. reported that both CAREC and CARES outperformed traditional readability formulas in predicting judgments of reading comprehension and speed.

All calculated readability formulas are available in the Automatic Readability Tool for English (ARTE; Choi & Crossley, in press). ARTE provides free and easy access to a wide range of readability formulas and automatically calculates different readability formulas for batches of texts (i.e., thousands of texts can be run at a time) to produce readability scores for individual texts in an accessible spreadsheet output.

2.3 Experimental Design

The experiment was designed using Eye-gaze Edge experiment builder software, *Nyan 2.0*. All texts and reading comprehension questions were typed in double-spaced, Times New Roman font (font size 14) in block text format with landscape text orientation. Nyan 2.0 software was used to randomize texts and reading comprehension questions for each text. In total, the experiment included 13 texts (one practice trial and twelve original texts) and 39 questions (three practice trial questions and thirty-six original text questions).

2.4 Procedure

Participants were first assigned a unique participant code and then provided informed consent. After consent, they were guided to a computer to complete the online demographic information survey and the Gates-MacGinitie (GMG) reading comprehension test. Participants were then led to another testing booth where they began the reading experiment portion of the study. This portion included a brief introduction of the experimental procedure to participants. The procedure involved initial presentation of one practice trial text and three multiple-choice questions to familiarize the participants with the nature of the experiment. After finishing the practice trial, participants read the twelve unique texts in random order and answered the three multiple-choice questions for each text at their own pace (neither text reading nor the multiple-choice comprehension test items were time limited). Questions were presented immediately after each text and participants could not return to the passages once they started answering questions. The participants used the keyboard number pad (keys 1-4) to record their responses to comprehension questions, and then pressed space

bar to move forward between passages and corresponding questions. The collection of demographic information, reading comprehension ability, and reading data was done in a single session, which lasted approximately 60 minutes.

Data for seven participants was unusable. In one case, a participant mentioned to the researcher that they had a reading processing disorder, and, in another case, a participant was observed using their cell phone while reading the texts. Additionally, data for three participants was lost due to software errors. Finally, analyses of accuracy revealed one participant to have below 5% total accuracy on the comprehension questions and another that spent less than 10 seconds reading each text, indicating a lack of engagement with the study task. Thus, the final dataset included data from 53 participants.

2.5 Statistical Analysis

Because the readability formulas all purport to measure similar constructs, they were assessed for multicollinearity using non-parametric correlations and variance inflation factors (VIF). Multicollinearity was defined as any two variables correlated at a higher absolute value than .7 or with a higher VIF value than 2.5 in the context of other predictors. In order to systematically compare effects for the different measures used in this study, all numerical predictors were standardized into z-scores. Additionally, we did not include text domain or simplification level as categorical factors in our models because there were only 12 texts in total, and this small number of texts resulted in multcollinearity between the readability formulas and levels of these variables, making it difficult to associate variance with the categorical label or differences in the readability formulas predictors.

We used Linear Mixed Effects (LME) models to test which effects exerted significant influences on reading times and comprehension scores. We built our models in R (R Core Team, 2017) using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015). For the reading time model, we entered reading time (in seconds) as the dependent variable. We then entered the following predictor variables as fixed effects: presentation order of the texts (to control for reading fatigue over time and labeled as trial order), participants' scores from the GMG reading comprehension test, age, English L1 status (yes/no), and participant survey responses for reading and television behavior, and a single readability formula (e.g., CAREC, Dale-Chall, or Flesch-Kincaid Grade Level). Subjects were entered as a random effect. We then hand-pruned the models to keep only those predictors that were significant. After hand-pruning, we also checked for evidence of suppression effects as manifested through mismatched correlation and regression coefficients. We followed the same general procedure to test participant comprehension accuracy using overall comprehension scores for the three questions in each text. We used the *MuMIn* package in order to obtain the marginal and conditional R^2 values for our models (Nakagawa & Schielzeth, 2013).

3 Results

3.1 Correlations

We conducted correlations between of dependent variables (comprehension and reading speed) and our readability formulas and individual difference measures. Initial correlation analyses showed that Flesch Reading Ease, FOG, and Flesch-Kincaid Grade Level were highly multicollinear. We thus removed Flesch Reading Ease and FOG from the subsequent analyses and kept Flesch-Kincaid Grade Level because it showed the highest correlation with comprehension scores and reading times. Correlations between readability formulas and comprehension scores indicated weak correlations for CAREC, Flesch-Kincaid, and Dale Chall. Correlations between readability formulas and reading speed indicated weak correlation for CARES (see Table 1).

Correlations between individual differences and comprehension scores indicated a weak correlation with GMG. Correlations between individual differences and reading speed indicated weak to moderate correlations with education background, amount of time reading, enjoyment of reading, confidence in reading, and GMG (see Table 2).

Table 1. Correlations between comprehension scores/reading times and readability formulas

Variables	1	2	3	4	5	6
1. Comprehension	1	-0.04	0.043	-0.167	-0.135	-0.242
2. Reading speed	-0.04	1	0.236	-0.087	-0.045	-0.067
3. CARES	0.043	0.236	1	-0.252	-0.354	-0.5
4. CAREC	-0.167	-0.087	-0.252	1	0.297	0.659
5. Flesch Kincaid	-0.135	-0.045	-0.354	0.297	1	0.503
6. Dale Chall	-0.242	-0.067	-0.5	0.659	0.503	1

Table 2. Correlations between comprehension scores/reading times and individual differences

Variables	1	2	3	4	5	6	7	8
1. Comprehension	1	-0.04	-0.005	-0.008	-0.006	-0.005	0.049	0.192
2. Reading speed	-0.04	1	-0.112	-0.138	0.087	-0.29	-0.221	-0.328
3. Education	-0.005	-0.112	1	0.211	0.004	0.279	0.08	-0.053
4. Amount read	-0.008	-0.138	0.211	1	0.141	0.609	0.243	0.103
5. Amount TV	-0.006	0.087	0.004	0.141	1	0.153	-0.031	-0.022
6. Enjoy read	-0.005	-0.29	0.279	0.609	0.153	1	0.303	0.208
7. Confidence read	0.049	-0.221	0.08	0.243	-0.031	0.303	1	0.293
8. GMG	0.192	-0.328	-0.053	0.103	-0.022	0.208	0.293	1

3.2 Comprehension Models

We conducted three linear mixed effects models to predict comprehension scores with each model featuring a different readability formula and all models including individual difference features. For each model, only the readability formula and the GMG scores were predictive. Model summaries including the variables kept, the *t* values for those variables, and the overall variance explained by each model (i.e., r2) are reported in Table 3. The strongest model was reported for the Dale-Chall readability formula

model, which explained ~10% of the variance. The CAREC model explains ~7% of the variance while the Flesch-Kincaid Grade Level model explained ~6% of the variance.

Table 3. LME Models to Predict Comprehension Scores

Model	Variable 1	Variable 2	t value (var 1)	t value (var 2)	Model r2
1	CAREC	GMG	-4.28**	4.926**	0.065
2	FKGL	GMG	-3.422**	4.888**	0.055
3	Dale-Chall	GMG	-6.296**	5.009**	0.095

3.3 Reading Speed Models

We conducted three linear mixed effects models to predict reading speed with each model featuring a different readability formula and all models including individual difference scores. Model summaries including the variables kept, the *t* values for those variables, and the overall variance explained by each model (i.e., r2) are reported in Table 4. The strongest model was reported for the CARES readability formula model, which explained ~19% of the variance and included GMG scores and trial order for the texts. The Dale-Chall model explained ~14% of the variance and also included trial order of the texts and GMG scores. Flesch-Kincaid Grade Level was not a significant predictor. The model using only trial order of the texts and GMG scores explained ~13% of the variance.

Table 4. LME Models to Predict Comprehension Scores

Model	Variable 1	Variable 2	Variable 3	t value (var 1)	t value (var 2)	t value (var 3)	Model r2
1	CARES	Trial order	GMG	8.903**	-5.248**	-3.497**	0.185
2	Trial order	GMG	NA	-5.857**	-3.487**	NA	0.134
3	Dale-Chall	Trial order	GMG	-2.239*	-5.737**	-3.489**	0.137

4 Discussion

This study examined if readability formulas were predictive of text reading times and comprehension scores stemming from a behavioral reading study. The results provide evidence that a model with the new Dale-Chall Readability formula explained the most variance in text comprehension scores while a model including the CARES formula explained the most variance in reading times when other factors related to text readability (e.g., reading proficiency), individual differences, and experimental design (i.e., trial order) were also considered. These findings suggest that newer formulas that better tap into the reading construct such as the New Dale-Chall and CARES formulas are likely the best predictors of text comprehension and processing speed, respectively, for the small corpus of text analyzed in this study. These formulas, which are intelligent improvements over previous formulas, can be used as technological mediators to better match learners with texts to keep learners on track to better achieve reading goals

With reference to comprehension scores, the model including the New Dale-Chall readability formula was the strongest predictor of comprehension scores. The negative coefficient indicated that texts with higher Dale-Chall scores were more difficult to comprehend. Beyond readability formulas, Gates-MacGinitie (GMG) Reading Test scores were also significant predictors. As would be predicted, the reading score coefficient indicates that students with higher reading scores had higher comprehension accuracy. In total, the New Dale-Chall and GMG explained ~10% of the variance. The models including GMG scores and either CAREC or Flesch-Kincaid Grade Level explained ~7% and ~6% respectively.

In terms of reading time, the model that included CARES explained the most variance in reading time. Other significant predictors in this model included trial order and GMG Test scores. The trial order results indicate that participants began to read texts more quickly as the experiment moved forward. The reading test scores indicate that more proficient readers took less time to read the texts. None of the other fixed factors were significant predictors of reading times. The model with CARES explained ~19% of the variance. A model including GMG scores, trial order, and the New Dale Chall formula explained around ~14% of the variance. Flesch-Kincaid Grade Level, when moderated by GMG scores and trial order, was not included as a significant factor in an individual model

These findings make important contributions to our understanding of the reliability and validity of various readability formulas in terms of their prediction success for text processing speed and comprehension for adult readers. We find that the New Dale-Chall formula explains the most variance for the data examined here. The New Dale-Chall formula contains an updated version of frequent words and the implementation of it in ARTE includes an expanded list of morphologically related words. Word frequency more strongly taps into lexical sophistication and should have stronger overlap with decoding than traditional lexical measures based on number of letters per word.. In contrast, the formula measures syntactic complexity solely as a function of average sentence length, striking a balance between new and old approaches toward measuring comprehension. Surprisingly, CAREC performed weaker than Dale_chall even though recent studies have shown improvements for CAREC when assessing text readability over other formulas (Crossley et al., 2019; Crossley et al., 2021). This may be a function of the manner in which CAREC was trained (using crowd-sourced judgments), the manner in which comprehension was operationalized in this study (i.e., multiple choice comprehension questions), the population studied, or the small number of texts analyzed.

We find that a model built on top of CARES was the strongest predictor of reading speed. Considering that CARES is the only readability formula specifically normed for processing speed, the results seem intuitive. Examinations of correlations between the other readability formulas and reading speed indicated few associations, providing evidence that most traditional formulas do not measure features of texts which influenced processing for this data set.

5 Limitations

There are a number of limitations to the approach used in this study and we discuss the most salient below with the goal of guiding future research. Most importantly, this study only examined 12 texts, which is not a large enough sample size to generalize about the strength of readability formulas beyond the scope of this study. Sample size is a continuous problem with reading comprehension studies because collecting readability criteria across large text samples is time consuming. However, the use of larger samples would help to extend these findings to a broader population. A larger sample size would allow for greater inclusion of a number of fixed factors as well, since autocorrelation was a problem with only 12 texts. For instance, domain and simplification levels were not included in the models because of autocorrelation. Post-hoc examinations of text reading times indicated that all texts regardless of domain were read at about the same speed, although simplified texts took a bit longer to read. In a larger corpus, these differences may become significant, potentially because of the extra length of simplified texts attributable to text elaboration. Additionally, it is likely that there are effects based on text topics. Sampling a large number of texts would also allow researchers to include topic as a random effect.

Another concern is that comprehension for this study was operationalized as multiple-choice comprehension questions, which may not be the most effective way to measure comprehension. Other approaches to measure comprehension include cloze-tests and textual recall, summarization, and inferential activities. Lastly, we did not assess text domain familiarity in this study, which may have influence on text processing and comprehension.study examined if readability formulas were predictive of text reading times and comprehension.

6 Conclusion

This study tested the prediction rates for classic and newer readability formulas for text comprehension and reading speed. We find that a model including the New Dale-Chall formula was the strongest predictor of comprehension (along with GMG scores) and that a model including newer readability formula related to text processing was the strongest predictor of reading speed (along with GMG scores and trial order). While the reading speed model explained about 19% of the variance, the reading comprehension model explained only around 10% of the variance. Overall, this study provides some validation for the use of smarter readability formulas which incorporate NLP inspired measures to predict text reading speed and comprehension. These newer readability formulas tap into the reading construct more intuitively and their increased accuracy over traditional readability formulas can act as technological mediators to increase reading success and the development of the reading process, thus creating the foundations for a smart personalized learning environment which presents learning resources adequate for each learner.

References

1. Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.
2. Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24*, 63–88.
3. Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.
4. Choi, J., & Crossley, S.A. (in press). *Automated Readability Web App for English.* Twenty-second IEEE International Conference on Advanced Learning Technologies *(ICALT 2022)*, Bucharest, Romania.
5. Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42*, 475-493.
6. Crossley, S. A., Heintz, A., Choi, J., Batchelor, J., Karimi, M., & Malatinszky, A. (in press). A large-scaled corpus for assessing text readability. *Behavior Research Methods.*
7. Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: new methods and new models. *Journal of Research in Reading, 42*(3-4), 541-561.
8. Davison, A., & Kantor, R. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly, 17*(2), 187–209.
9. DuBay, W. H. (2004). *The Principles of Readability.* Costa Mesa, CA: Impact Information
10. Fry, E. (1977). Fry's readability graph: Clarifications, validity, and extension to level 17. *Journal of reading*, *21*(3), 242-252.
11. Gunning, R. (1952). *The technique of clear writing*. New York, NY: McGraw-Hill.
12. Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability Formulas: (Automated readability index, fog count and Flesch Reading Ease Formula) for Navy enlisted personnel*. (No. RBR-8-75). Naval Technical Training Command, Millington, TN: Research Branch.
13. Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13).
14. MacGinitie, W. H., MacGinitie, R. K., Cooter, R. B., & Curry, S. (1989). Assessment: Gates-Macginitie Reading Tests. *The Reading Teacher, 43*(3), 256-258.
15. McNamara, D. S., Levinstein, I. B. & Boonthum, C. ((2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers 36* (2), 222-233. https://doi.org/10.3758/BF03195567
16. Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, *4*(2), 133-142.
17. National Assessment of Educational Progress. (2011). The Nation's Report Card: Writing 2011.
18. Newbold, N., & Gillam, L. (2010). The linguistics of readability: the next step for word processing. *In Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids* (pp. 65-72). Association for Computational Linguistics.
19. Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 186-195). Association for Computational Linguistics.
20. Powell, P. R. (2009). Retention and writing instruction: Implications for access and pedagogy. *College Composition and Communication, 60*, 664-682.

21. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2020). *The nation's report card*. Washington, DC: National Center for Education Statistics.
22. Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes, 25*(2-3), 309-336.