Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

INSTRUCTIONS

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

GRANTEE SUBMISSION REQUIRED FIELDS

Title of article, paper, or other content

Assessing Readability by Filling Cloze Items with Transformers

All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.

Last Name, First Name	Academic/Organizational Affiliation	ORCID ID
Olney, Andrew M.	University of Memphis	0000-0003-4204-6667

Publication/Completion Date—(if *In Press,* enter year accepted or completed) 2022

Check type of content being submitted and complete one of the following in the box below:

- ☐ If article: Name of journal, volume, and issue number if available
- ☑ If paper: Name of conference, date of conference, and place of conference
- ☐ If book chapter: Title of book, page range, publisher name and location
- ☐ If book: Publisher name and location
- ☐ If dissertation: Name of institution, type of degree, and department granting degree

22nd International Conference on Artificial Intelligence in Education, July 27-31 2022, Durham UK

DOI or URL to published work (if available) https://doi.org/10.1007/978-3-031-11644-5_25

Acknowledgement of Funding— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

"This work was supported by U.S. Department of Education **[Office name]** Institute of Education Scie through **[Grant number]** R305A1904 to **Institution]** the University o. The opinions expressed are those of the authors and do not represent views of the **[Office name]** Institute of Education Scie or the U.S. Department of Education.

Assessing Readability by Filling Cloze Items with Transformers

Andrew M. Olney[1 [0000−0003−4204−6667]]

University of Memphis, Memphis TN 38152, USA
aolney@memphis.edu

Abstract. Cloze items are a foundational approach to assessing readability. However, they require human data collection, thus making them impractical in automated metrics. The present study revisits the idea of assessing readability with cloze items and compares human cloze scores and readability judgments with predictions made by T5, a popular deep learning architecture, on three corpora. Across all corpora, T5 predictions significantly correlated with human cloze scores and readability judgments, and in predictive models, they could be used interchangeably with average word length, a common readability predictor. For two corpora, combining T5 and Flesch reading ease predictors improved model fit for human cloze scores and readability judgments.

Keywords: readability · assessment · cloze · transformers

1 Introduction

Cloze items, also known as fill-in-the-blank items, are widely used in education for assessment and some types of instruction (e.g. vocabulary instruction). However, cloze items also have a long history as a measure of readability, i.e. of text difficulty [16]. The standard approach to assessing readability with cloze items is called nth deletion, where every nth word in a text is deleted and replaced with a blank of fixed size. The task of the reader is to use their knowledge and context cues across the entire text to fill in the blanks.

It has long been known that a higher number of correct completions on nth deletion cloze tests is a strong indicator of higher readability (low difficulty) and aligns with well-known readability metrics like Flesch reading ease and Dale–Chall readability, aptitude tests, and standard comprehension questions [16, 17, 2, 4]. Unlike comprehension test questions, which are difficult to create and confound the measurement of text difficulty with question difficulty, cloze items can be generated directly from text and have less measurement error.

Despite their effectiveness as a readability measure, cloze items are not a practical in most cases because they require human-subjects data collection. For this reason, practical readability metrics have been developed to have high correlation to measures like cloze, but otherwise use easily calculated characteristics of the text to determine a readability score. Common examples of such metrics are Flesch reading ease [10], which uses average sentence length in words (ASL)

and average word length in syllables (AWL), Dale-Chall readability, which uses the proportion of difficult words (defined by a word list) and ASL [7, 5], and the Lexile measure, which uses word frequency and ASL [15].

These metrics are quite simple but also quite effective at assessing readability on a large scale - not because they are causally related to readability but rather because they are so strongly correlated with factors that influence readability. The seeming paradox that the simplest measures would be the best predictors of readability was addressed by Bormuth, who described it as a trade-off between face validity and predictive validity [3]: many linguistic variables correlate with readability, so a metric with face validity would include many linguistic variables; however, the measurement error associated with these variables means that a metric with fewer variables has better predictive power when applied to unseen texts. Thus, while there has been continued interest in creating better readability metrics, especially in the modern era (see [6] for a review), these simple metrics are a challenging baseline. For example, Martinc et al. found that their deep neural language models were not able to outperform an ASL baseline ($r = .906$) on the Newsela corpus in an unsupervised setting [12].

Modern deep learning methods, notably Transformers [19] offer a potential alternative to traditional readability metrics. As described above, the traditional approach is to use linguistic features to predict cloze item performance, which is a ground-truth measure of readability. In contrast, Transformers can be used to predict cloze difficulty directly because this is how they are trained in the first place - to predict masked tokens in their input. Deep learning methods based on masked language modeling have proven to be extremely effective in a variety of natural language processing (NLP) tasks [8, 13], so presumably, they would function well for a task aligned with their pre-training objective. The idea of using Transformers to directly measure cloze difficulty was first investigated by Benzahra & Yvon, unfortunately without much success [1]. They used GPT-2, an autoregressive Transformer, to predict cloze completions on two corpora with experts-labeled grade levels and achieved overall correlations of .05 and .13, respectively. However, we argue that GPT-2 is the wrong model to use for this task because it is autoregressive and only allows leftward context to be used to predict the next word or words. In contrast, the nth deletion cloze task allows the use of both left and right context across the entire document. Therefore, additional study of Transformers to directly predict cloze difficulty is warranted.

The present investigation examines the application of Transformers to measuring both cloze difficulty and grade-level readability. Our primary research question is whether Transformer cloze scores correspond with these measures and standard readability metrics. The remainder of the paper is organized around three different studies with different corpora. The first corpus, the Bormuth passages [4], allow direct comparison to cloze item difficulty calculated from human subjects experiments, in addition to comparison to other relevant measures like comprehension tests. The second two corpora, the OneStopEnglish corpus (OSE) [18] and the Newsela corpus [20], allow comparison to expert-defined grade lev-

<extra_id_0> of the Big Cats, <extra_id_1> well as the lesser <extra_id_2>, have wonderful eyes. They <extra_id_3> see clearly even on <extra_id_4> dark night. This is <extra_id_5> of the way they <extra_id_6> made. There is a <extra_id_7> of window in each <extra_id_8>. This window is called <extra_id_9> pupil. It is black <extra_id_10> is placed in the <extra_id_11> of the colored part <extra_id_12> the eye. The pupil <extra_id_13> light come in and <extra_id_14> a kind of mirror <extra_id_15> the back of each <extra_id_16>. These mirrors reflect everything <extra_id_17> is in front of <extra_id_18> eyes. Right away a <extra_id_19> nerve carries these reflected <extra_id_20> to the brain. Then <extra_id_21> brain sends a quick <extra_id_22> to all parts of <extra_id_23> body. This signal may <extra_id_24> to attack, hide, be <extra_id_25>

Fig. 1. A chunk representing one-half of a 250-word text submitted to T5. Each cloze word has been replaced by a highlighted T5 sentinel token, which is ordered sequentially using nth deletion, $n = 5$.

els for each text. Across all corpora, additional comparisons will be made to standard readability metrics like Flesch reading ease.

2 Approach

All of the following studies use a Transformer called T5 [13] to measure both cloze difficulty and readability. T5 is a suitable model because it attends to both left and right contexts and because it is trained on a denoising objective that closely matches the cloze task. To match the method of Bormuth [4], only the first 250 words of each text are subjected to nth deletion ($n = 5$). Five clozed versions of each text are created by using different offsets for nth deletion, e.g. starting at words 1, 2, 3, 4, and 5, after which subsequent words have been deleted by a previous version. As a result, every word in the text is subjected to cloze in exactly one offset version. During development, it was discovered that the T5 model used[1] produces degenerate responses to cloze items after the 27th item[2]. Therefore, each text was split into two chunks, each representing 125 words and 25 cloze items, given the $n = 5$ nth deletion strategy, and the two chunks were submitted to T5 separately for each of the 5 offsets noted above The need to break the text into chunks for T5 is a notable departure from Bormuth's method because it creates less context for T5 to complete the task than what is afforded to humans, making the task more difficult. Otherwise, this task is broadly consistent with T5's unsupervised denoising training objective of predicting the randomly deleted 15% of tokens vs. predicting nth-deleted tokens, $n = 5$, or 20% of tokens. An example of a chunk input to T5 is shown in Figure 1.

Several approaches to generating cloze predictions were explored during initial investigations, with the goal of generating multiple predictions for each cloze item. Multiple predictions are desirable because they allow partial credit for

[1] `https://huggingface.co/t5-large`
[2] `https://github.com/huggingface/transformers/issues/8842`

lower-ranked predictions using metrics like reciprocal rank, where a correct prediction at rank N receives a score of $1/N$. Multiple predictions are also desirable because they potentially reflect a distribution of predictions across human subjects, rather than a single prediction. Our investigations suggested that greedy beam search had desirable properties of being highly repeatable but the disadvantage of not producing much diversity when multiple predictions were requested, regardless of the number of beams and diversity penalties applied.

Therefore, for the top prediction, we used greedy beam search with one beam, and for the remaining predictions, we used sampling with both top-K [9] and top-p [11] approaches. Because sampling is stochastic, the sampling results are not highly repeatable, but because the cloze metrics are assessed per text, we consider these as being repeated 250 times, once for each word in the text. To avoid repetitions and multi-word predictions, which are impossible given the task, duplicate predictions were removed from lower ranks, and predictions that contained internal whitespace (as a separator between words) were excluded.

Two accuracy metrics were calculated for each cloze item using these predictions. Correct at rank 1 was defined by an exact match between the top prediction and the original word, normalized for case and leading/trailing whitespace. Correct at any rank was defined by a similar exact match on a prediction of any rank, weighted by reciprocal rank. In addition to the T5 cloze metrics, Flesch and Dale-Chall readability metrics were calculated for each text[3].

3 Study 1: Bormuth passages

3.1 Data

The Bormuth passages were used in a major study of readability that incorporated cloze items (nth deletion; $n = 5$), reading rate, and pre/post comprehension questions [4]. To create these passages, Bormuth ranked 330 passages used in another study [3] by cloze difficulty, divided the difficulty range into 8 points, and selected the 4 passages closest to those 8 points, such that no more than 4 came from the same subject matter category and each text was at least 250 words in length. Thus the 32 passages represent 8 difficulty levels spanning from first grade to college. Each passage and corresponding measures were extracted from the appendix [4] and manually checked for errors; passages were additionally submitted to Grammarly to catch any errors. Grammarly revealed that several passages (3213, 5226, 6441, 7151, and 8552) contained spelling mistakes. In order to prevent T5 from correctly predicting a word but not matching the original misspelled form, all spelling errors were corrected. Additionally, passage 6545 had no corresponding entry in the data tables and passage 6535 listed in the data tables had no corresponding text; these were assumed to refer to the same passage. Finally, Lexile scores were calculated using the Lexile Text Analyzer[4]. Because the Analyzer only allows 50 texts to be processed per month, Lexiles were only computed for this dataset.

[3] `https://github.com/cdimascio/py-readability-metrics`
[4] https://hub.lexile.com/analyzer

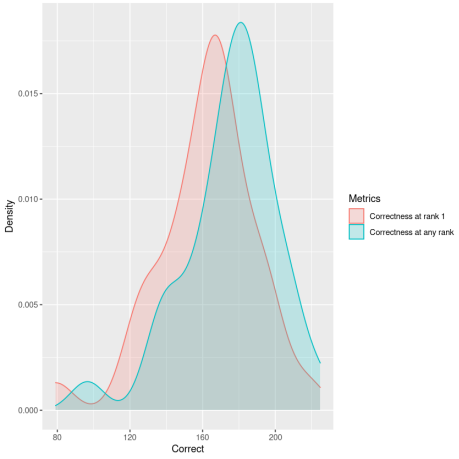

Fig. 2. Density plot for correctness at rank 1 and correctness at any rank.

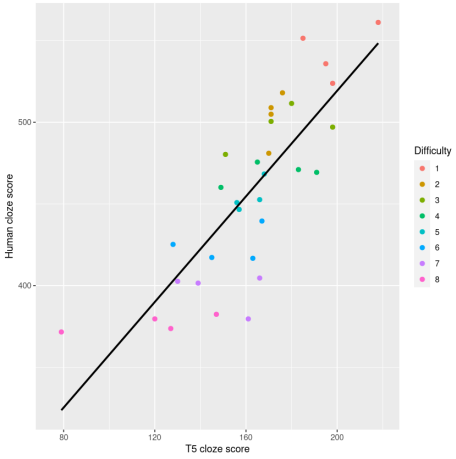

Fig. 3. Scatterplot of T5 cloze correctness and human correctness. Regression lines show smoothed (blue) and linear (red) fits.

3.2 Results

The primary questions of interest in this study are whether the T5 cloze scores correspond with the human cloze scores, as well as how these scores comparatively relate to other measures of readability. To address the first question, we examined the differences between the correctness at rank 1 metric and the correctness at any rank metric in order to determine which was the most appropriate measure for the following analyses. As shown in Figure 2, the correctness at rank 1 ($M = 162.22, SD = 27.24$) and correctness at any rank ($M = 175.33, SD = 25.97$) are approximately equivalent in distribution, except correctness at any rank is slightly more lenient and therefore right-shifted. Because the difference seemed relatively negligible and correctness at rank 1 has better repeatability, we only report results for correctness at rank 1, which we will refer to as T5 cloze scores.

The Spearman rank order correlation between the T5 cloze scores and human cloze scores was significant, $r(30) = .86$, $p < .001$. A scatterplot between the two scores is shown in Figure 3. The relationship is approximately linear, with visible separation of the eight difficulty levels along with visible overlap. Two ANOVA analyses were conducted to examine the discriminability of human and T5 cloze scores according to these levels. The human cloze score ANOVA was significant, $F(7, 24) = 93.19$, $p < .001$. Pairwise tests using Tukey's HSD revealed that every difficulty level was significantly different from the other, $p < .05$, except for levels 2 and 3, levels 4 and 5, and levels 7 and 8, i.e. there are effectively 5 levels of difficulty rather than 8 according to this measure. The T5 cloze score ANOVA was also significant, $F(7, 24) = 7.50$, $p < .001$. Pairwise tests using Tukey's HSD revealed that level 1 was significantly different from levels 6, 7,

Table 1. Rank order correlations between readability measures for Bormuth passages.

	Cloze	Pre	Post	Read	ASL	AWL
Pre	.87					
Post	.95	.92				
Reading	.88	.84	.87			
ASL	-.88	-.72	-.80	-.69		
AWL	-.84	-.77	-.86	-.85	.67	
T5	.86	.72	.72	.72	-.80	-.62

Note: All correlations significant, $p < .001$.

and 8; additionally levels 2, 3, 4, and 5 were significantly different from level 8, all $p < .05$. These results indicate that although the correlation between T5 cloze scores and human cloze scores is strong, the discriminability of T5 cloze scores with respect to the assigned difficulty levels is less than that of the human cloze scores. One possible reason for this is that the human cloze scores were based on students from grades 3 to 12, so the scores for difficult passages were drawn down by students from lower grades. In contrast, T5 is a single model with a single ability level.

Correlations with additional readability measures are shown in Table 1. The highest correlations were between the human measures in the upper left. The T5 cloze scores and the classic readability components, average sentence length in words (ASL) and average word length in syllables (AWL), have similar correlations to each of the human components, with the exception of post-test score. Since post-test score represents human performance after reading the text, a low correlation might be expected, but it is notable that AWL and ASL have a stronger correlation with post-test than pre-test, while T5 cloze scores have the same correlation with both. Surprisingly, the T5 cloze scores are more strongly correlated with ASL than AWL, suggesting that T5 is using linguistic information at the sentence level more than at the word level as it makes cloze predictions.

The results in Table1 suggest that the T5 cloze scores could be combined with ASL, AWL, or both to create a model in the style of classic readability metrics like Flesch reading ease. To investigate this possibility and compare to the standard Flesch model, four models were constructed using combinations of these predictors. The models and their fits are reported in Table 2. The best-fitting model used all predictors, giving it a .06 improvement in fit over the Flesch reading ease model. However, this comparison is somewhat unfair as the Flesch model has only two predictors. The remaining models are two predictor contrasts to the Flesch model. The T5+ASL model has a fit .01 below the Flesch model, and the T5+AWL has a fit .03 below the Flesch model. Altogether, these models indicate that the T5 cloze scores potentially have some additive benefit to ASL and AWL and can be used almost interchangeably for this task.

Table 2. Linear models predicting human cloze scores for Bormuth passages.

	Coefficients			
Model	T5	ASL	AWL	R^2
All	.62	-4.56	-95.31	.94
Flesch		-6.09	-116.87	.88
T5+ASL	.79	-6.58		.87
T5+AWL	.95		-149.03	.85

Note: For all coefficients, $p < .001$.

4 Study 2: OneStopEnglish

4.1 Data

The OneStopEnglish corpus (OSE) is a balanced corpus consisting of 189 texts topics, each in three versions of difficulty, for a total of 567 texts [18]. Texts were collected from onestopenglish.com, a site for English language learners, and consisted of news stories that had been simplified by teachers for news-based lessons. The three levels of difficulty are thus aligned with pedagogical goals in ESL. Each difficulty level has a reported Flesch-Kinkaid Grade Level, 6.4 for Beginner, 8.2 for Intermediate, and 9.5 for Advanced. Unlike the Bormuth passages, OSE has no human-derived readability measures, so its primary utility for readability research stems from its expert-labeled difficulty levels. OSE is a popular corpus for readability research and was used in several studies mentioned in Section 1 [1, 12].

4.2 Results

The primary research question for this study is the alignment of T5 cloze scores with expert difficulty and other measures of readability. To keep the results comparable with the last study, difficulty in these results is reverse scaled as ease. As in the previous study, we checked the distributions of the correct at rank 1 and the correct at any rank metrics. The distributions were comparable to Figure 2 relative to each other, but the distributions of correct at rank 1 ($M = 156.82, SD = 11.28$) and correct at any rank ($M = 171.32, SD = 10.51$), were markedly narrower than in the last study, likely reflecting the smaller range of difficulty in OSE compared to the Bormuth passages. We again chose correct at rank 1 as our T5 cloze score metric in the following analyses.

The rank-order correlation between the T5 cloze scores and the three levels of ease was significant, $r(565) = .19$, $p < .001$, but notably smaller than in the last study. We additionally calculated Kendall's tau-b to compare to the previous work that used Transformers to predict cloze scores on this corpus, Benzahra and Yvon [1]. Our $\tau_b = .15$ for OSE versus their $\tau_b = .05$, a threefold improvement but a modest score nonetheless.

Table 3. Rank order correlations between readability measures for the OneStopEnglish corpus.

	Ease	ASL	AWL
ASL	-.58*		
AWL	-.37*	.29*	
T5	.19*	.02	-.07

*Note: * $p < .001$.*

Table 4. Linear models predicting reading ease for the OneStopEnglish corpus.

	Coefficients			
Model	T5	ASL	AWL	R^2
All	.01	-.09	-1.91	.41
Flesch		-.09	-2.05	.38
T5+ASL	.01	-.10		.37
T5+AWL	.01		-3.19	.16

Note: For all coefficients, $p < .001$.

Rank-order correlations with readability measures shown in Table 3 provide further insight into this low overall correlation between T5 cloze scores and levels of ease. In contrast to the previous study, the T5 cloze scores are not significantly correlated with either ASL or AWL. Additionally, the correlation between ASL and AWL is less than half what it was in the previous study. The cause of these changes in correlation is not clear, and possible explanations include the limited range of ease, the different genres (news text vs. informational text), and the mode of construction (artificially created vs. naturally occurring).

An ANOVA analysis was conducted to examine the discriminability of T5 cloze scores according to the levels of ease. The ANOVA was significant, $F(2, 564) = 14.47$, $p < .001$. Pairwise tests using Tukey's HSD revealed that Elementary texts ($M = 160.29, SD = 9.59$) were significantly easier than Intermediate texts ($M = 154.58, SD = 11.35$), $p < .001$, as well as Advanced texts ($M = 155.59, SD = 11.98$), $p < .001$. Intermediate and Advanced texts were not significantly different from each other, $p = .643$.

Although the correlations in Table 3 are lower than the previous study, each of the metrics is significantly correlated with the level of ease. Therefore additional regression models matching those in Table 2 were created, and the results are presented in Table 4. The rank order of model fit matches the previous study. The model containing all predictors had the best fit, followed by Flesch reading ease. The T5+ASL model has a fit .01 below the Flesch model, and the T5+AWL was markedly worse at .22 below the Flesch model. As before, models improve with the T5 predictor, and the T5 predictor is almost interchangeable with AWL. However, on OSE, the T5 predictor is not as interchangeable with ASL, as shown by the poor fit of the final model.

5 Study 3: Newsela

5.1 Data

The Newsela corpus[5] was introduced by Xu and colleagues [20] as a resource for text simplification research, but it has also been used for readability research [12]. Like OSE, the Newsela corpus contains multiple versions of the same text topic at different difficulty levels, and the text topics are drawn from the news. However, Newsela is different from OSE in a number of ways. Newsela has a greater number of versions for each text topic (typically 5) and spans a greater range of difficulty, grade 2 to grade 12. However, the distribution of grade-level text in the corpus is not balanced, and the number of texts at each grade level ranges from 2 to 2096. Newsela is designed to match English language learning needs of native, rather than ELL speakers. Finally, Newsela's grade levels are approximately aligned with Lexile, which increases its usefulness for readability research. All 9565 English texts of Newsela were used, consisting of 1911 text topics.

5.2 Results

The primary research question for this study is again the alignment of T5 cloze scores with expert difficulty and other measures of readability, and whether this alignment will be more consistent with the first or second study. To keep the results comparable, grade level is reverse scaled as ease. Distributions of the correct at rank 1 and the correct at any rank metrics were similar to the distributions in Section 4. The distribution of correct at rank 1 ($M = 159.00, SD = 12.89$) and correct at any rank ($M = 173.35, SD = 12.08$), were comparably narrow as the OSE distributions, suggesting that the narrowness of the distributions is not attributable to a restricted range of difficulty. To stay consistent with the other studies, correct at rank 1 was again chosen as our T5 cloze score metric in the following analyses.

The rank-order correlation between the T5 cloze scores and the 11 levels of ease was significant, $r(9563) = .33$, $p < .001$, was in between the correlations found in the previous studies. An ANOVA conducted to examine the discriminability of T5 cloze scores according to the levels of ease was significant, $F(10, 9554) = 126.91$, $p < .001$. Pairwise tests using Tukey's HSD revealed that texts from grade 2 were significantly easier than texts from grades 5-10 and 12; texts from grade 3 were significantly easier than texts from grades 4-10 and 12; texts from grade 4 were significantly easier than texts from grades 5-10 and 12; texts from grade 5 were significantly easier than texts from grades 6-10 and 12; texts from grade 6 were significantly easier than texts from grades 8, 10, and 12; texts from grade 7 were significantly easier than texts from grades 8, 10, and 12; texts from grade 8 were significantly easier than texts from grade 12; and texts from grade 9 were significantly easier than texts from grade 12, all

[5] `https://newsela.com/data/`

Table 5. Rank order correlations between readability measures for the Newsela corpus.

	Ease	ASL	AWL
ASL	-.95		
AWL	-.63	.62	
T5	.33	-.29	-.16

Note: For all r, $p < .001$.

Table 6. Linear models predicting reading ease for the Newsela corpus.

	Coefficients			
Model	T5	ASL	AWL	R^2
All	.02	-.56	-1.50	.83
Flesch		-.58	-1.49	.83
T5+ASL	.02	-.58		.83
T5+AWL	.05		-17.27	.40

Note: For all coefficients, $p < .001$.

$p < .05$. Nonsignificant comparisons involving grades 10 and 11 are perhaps best explained by the small number of texts assigned to these levels, 22 total. Altogether, the ANOVA results indicate that T5 cloze scores afford a fair level of discriminability for Newsela grade levels.

Correlations with readability measures are shown in Table 5. The strength of the correlations again falls in between those of the previous studies. ASL and AWL are correlated comparably to the first study, but ASL is much more strongly correlated with ease than in the second study. Although the cause of these differences in correlation remains uncertain, it seems that genre can be ruled out as a cause, given that the corpora from the second and third studies are news corpora. These correlations provide additional evidence for another possible cause, which is the larger range of ease. A larger range of ease is a common characteristic between the first study and the third study and so may explain the similarities in correlation.

Regression models matching those used in the previous studies were created and results are presented in Table 6. The fits of the models follow a different pattern from the previous studies, with the first three models achieving the same fit. For the first time, the T5+ASL model has a fit equivalent to Flesch, providing additional evidence that T5 cloze scores are almost interchangeable with AWL. As in study 2, the poor fit of the T5+AWL indicates that T5 is not interchangeable with ASL.

6 Discussion

The focus of this work was to examine the use of T5 for predicting cloze item difficulty, a standard for readability, along with its analogous grade-level readability. A consistent pattern of results emerged across the three studies. In each case, T5 cloze scores significantly correlated with the outcome measures of interest, human cloze difficulty or expert-assigned grade level. Additionally, T5 cloze scores typically improved prediction of the outcome measures of interest when combined with the Flesch reading ease components of average sentence length (ASL) and average word length (AWL). In all studies, T5 cloze scores could be

substituted for AWL in linear models and provide a fit almost as good, or as good, as Flesch reading ease.

However, there were some notable differences across the studies. The most striking difference is that T5 cloze scores were much more strongly correlated with human cloze scores (study 1) than with expert-assigned grade levels (studies 2 and 3). It seems unlikely that this difference can be explained by differences in genre or patterns of correlation across the studies, since studies 1 and 3 have similar patterns of correlation between the outcome measures, ASL, and AWL, but studies 2 and 3 shared the same genre, news. Neither can the differences be explained by the range of difficulties in the texts, since both studies 1 and 3 have approximately the same range of grades. Rather, the results across the studies suggest that T5 cloze scores are more aligned with human cloze scores than with expert-assigned grade levels, which is somewhat surprising because human cloze scores and expert-assigned grade levels themselves should be highly correlated [3, 4]. Clearly, further research on this question is needed, focusing on naturalistic informational texts to replicate the strong findings found in study 1.

The question remains as to whether T5 cloze difficulty has the potential to improve readability measures that have been in place for many decades. After all, in our studies, T5 cloze scores at best replaced a component of Flesch reading ease. The primary reason that T5 might be useful going forward is that it encodes substantial knowledge about the world, and it makes cloze predictions using that knowledge. For example, T5 has been used for closed book trivia question answering without explicitly teaching it the knowledge involved [14]. This capability is analogous to a human reader bringing to bear background knowledge in order to understand a text, and it is something that isn't captured by word- or sentence-length metrics. Exactly how to manifest this capability in a readability model such that it consistently outperforms established metrics is a matter for future research.

Acknowledgments This material is based upon work supported by the National Science Foundation under Grants 1918751 and 1934745 and by the Institute of Education Sciences under Grant R305A190448. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Institute of Education Sciences.

References

1. Benzahra, M., Yvon, F.: Measuring text readability with machine comprehension: a pilot study. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 412–422. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-4443
2. Bormuth, J.R.: Cloze test readability: Criterion reference scores. Journal of Educational Measurement **5**(3), 189–196 (1968)
3. Bormuth, J.R.: Development of readability analysis. Tech. Rep. ED 029 166, University of Chicago (1969), `https://eric.ed.gov/?id=ED029166`

4. Bormuth, J.R.: Development of standards of readability: Toward a rational criterion of passage performance. Final report. Tech. Rep. ED 054 233, University of Chicago (1971), https://eric.ed.gov/?id=ED054233
5. Chall, J., Dale, E.: Readability Revisited: The New Dale-Chall Readability Formula. Brookline Books (1995)
6. Collins-Thompson, K.: Computational assessment of text readability: A survey of current and future research. ITL - International Journal of Applied Linguistics **165**(2), 97–135 (2014)
7. Dale, E., Chall, J.S.: A formula for predicting readability. Educational Research Bulletin **27**(1), 11–28 (1948), http://www.jstor.org/stable/1473169
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019)
9. Fan, A., Lewis, M., Dauphin, Y.N.: Hierarchical neural story generation. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. pp. 889–898. Association for Computational Linguistics (2018)
10. Flesch, R.: A new readability yardstick. Journal of Applied Psychology **32**(3), 221 – 233 (1948)
11. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. In: 8th International Conference on Learning Representations. OpenReview.net (2020), https://openreview.net/forum?id=rygGQyrFvH
12. Martinc, M., Pollak, S., Robnik-Šikonja, M.: Supervised and Unsupervised Neural Approaches to Text Readability. Computational Linguistics **47**(1), 141–179 (2021)
13. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020)
14. Roberts, A., Raffel, C., Shazeer, N.: How much knowledge can you pack into the parameters of a language model? In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 5418–5426. Association for Computational Linguistics, Online (Nov 2020)
15. Stenner, A.J., Sanford-Moore, E.E., Burdick, D.: The Lexile Framework for Reading technical report. Tech. rep., MetaMetrics, Inc (2007)
16. Taylor, W.L.: "Cloze procedure": A new tool for measuring readability. Journalism Quarterly **30**(4), 415–433 (1953)
17. Taylor, W.L.: "Cloze" readability scores as indices of individual differences in comprehension and aptitude. Journal of Applied Psychology **41**(1), 19–27 (1957)
18. Vajjala, S., Lučić, I.: OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 297–304. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Proceedings of the Thirty-first Annual Conference on Neural Information Processing Systems. pp. 5998–6008 (2017)
20. Xu, W., Callison-Burch, C., Napoles, C.: Problems in current text simplification research: New data can help. Transactions of the Association for Computational Linguistics **3**, 283–297 (2015)