Extending the MNREAD sentence corpus: Computer-generated sentences for measuring visual performance in reading

Mansfield, J. S.[a], Atilgan, N.[b], Lewis, A. M.[a], Legge, G. E[b]

[a] *Department of Psychology, SUNY College at Plattsburgh, Plattsburgh, New York, USA*
[b] *Department of Psychology, University of Minnesota, Minneapolis, Minnesota, USA*

Abstract

The MNREAD chart consists of standardized sentences printed at 19 sizes in 0.1 logMAR steps. There are 95 sentences distributed across the five English versions of the chart. However, there is a demand for a much larger number of sentences: for clinical research requiring repeated measures, and for new vision tests that use multiple trials at each print size. This paper describes a new sentence generator that has produced over 9 million sentences that fit the MNREAD constraints, and demonstrates that reading performance with these new sentences is comparable to that obtained with the original MNREAD sentences. We measured reading performance with the original MNREAD sentences, two sets of our new sentences, and sentences with shuffled word order. Reading-speed versus print-size curves were obtained for each sentence set from 12 readers with normal vision at two levels of blur (intended to simulate acuity loss in low vision) and with unblurred text. We found no significant differences between the new and original sentences in reading acuity and critical print size across all levels of blur. Maximum reading speed was 7% slower with the new sentences than with the original sentences. Shuffled sentences yielded slower maximum reading speeds and larger reading acuities than the other sentences. Overall, measures of reading performance with the new sentences are similar to those obtained with the original MNREAD sentences. Our sentence generator substantially expands the reading materials for clinical research on reading vision using the MNREAD test, and opens up new possibilities for measuring how text parameters affect reading.

Keywords: MNREAD, reading acuity, critical print size, maximum reading speed, blur

1. Introduction

The MNREAD Acuity Chart is a continuous-text reading-acuity chart designed for assessing how reading performance depends on print size (Mansfield et al., 1993; Mansfield and Legge, 2007). Each chart consists of sentences printed in a series of 19 sizes in 0.1 logMAR increments. The MNREAD sentences are intended to satisfy two requirements: they need to be realistic so that they demand the same perceptual and cognitive processes that are required for normal everyday reading, and they each need to be matched for readability and legibility so that they will yield consistent and reliable measures of reading performance from trial to trial. To meet these goals, the sentences are constrained to use a restricted vocabulary, to have the same length (60 characters), and to have a tightly constrained physical layout when printed (complete specifications are described in Section 2).

There has been considerable demand for a large number of standardized sentences for testing vision. The recent advent of computer- and tablet-based tests of reading acuity (e.g. Calabrèse et al., 2018; Xu and Bradley, 2015) has made it feasible for vision tests to include a large number of sentences. Further, research studies often require a large number of sentences so that repeat measures can be obtained without reusing sentences in order to minimize learning and repeated testing effects. It has proven challenging to create sentences that meet the MNREAD constraints. One difficulty is that, in addition to using a limited vocabulary, word choice is further restricted both by the number of letters in the word and by the width of the word when it is printed. The width of a word is only loosely linked to the number of letters it contains (e.g., 'common' is almost twice the width of 'little' even though they both have six letters). Composing MNREAD sentences has required using a computer program to keep track of the line width and letter count while the user adds words to the sentence. Even with this computer assistance, relatively few sentences have been created. There are only 95 sentences distributed across the five English versions of the MNREAD Acuity Chart.

The need for expanded testing materials has been addressed by other researchers. Xu and Bradley (2015) created a computer-based continuous-text acuity chart that contains 422 sentences similar to MNREAD sentences. Crossland et al. (2008) and Perrin et al. (2015) have designed sentence generators that produce thousands of short sentences that are objectively true or false (e.g., "some dogs are animals"). These sentences allow reading accuracy to be verified simply by requiring the reader to

make a true/false response to each sentence rather than reading the sentence aloud. Rassia and Pezaris (2018) have created a large corpus of sentences by parsing sentences from texts downloaded from Project Gutenberg (http://www.gutenberg.org), and selecting those that conformed with the MNREAD criteria. They additionally applied transforms to sentences that were close to the target length of 60 characters by replacing words with shorter or longer alternates that did not alter the grammatical structure of the sentence (e.g., changing 'she' to 'he'), to produce a corpus of 1600 sentences.

We have built a computer algorithm that has produced more than 9 million MNREAD sentences. This paper describes our sentence generator, and demonstrates that reading performance with these computer-generated sentences is comparable to that obtained with the original MNREAD sentences.

2. MNREAD sentences

2.1. Language constraints

The MNREAD sentences are simple declarative sentences; they use a lexicon of the 3000 most-common words in 3rd grade reading materials (Zeno et al., 1995). The sentences do not contain any proper nouns, and only the initial letter of each sentence is capitalized. The sentences have no punctuation.

During development of the original MNREAD charts, candidate sentences were tested for readability to eliminate any that were read too quickly or too slowly. The sentences used for each chart were selected to minimize repetitions of concrete nouns, and ordered to avoid semantic content running across adjacent sentences.

2.2. Length constraints

Each MNREAD sentence consists of 60 characters, including a single space between each word and an implied period at the end. The number of words in each sentence is allowed to vary (so long as the sentence has 60 characters). The original MNREAD sentences contain from 10 to 15 words (average 12.27 words per sentence). The use of 60-character sentences is convenient for scoring reading performance because each sentence is equivalent to 10 six-character standard-length words (Carver, 1976).

2.3. Layout constraints

The MNREAD sentences are printed using the Times-Roman font on three lines of left-right justified text. The width of each line of text is $20\bar{w} - w_{\text{space}}$, where w_{space} is the width of a 'space' and $\bar{w}$ is the average character width calculated according to $\bar{w} = \Sigma w_c f_c$, where w_c is the width of each character determined from the Times-Roman font metrics, and f_c is the relative frequency of each character (calculated from 11,000 60-character sentences created by our text generators, see Appendix.) Note that in the original specification Mansfield and Legge (2007), the average character width was determined using word frequencies reported by Kučera and Francis (1967), and this average was multiplied by 19 to give a line width of 17.5 x-heights. Our new specification gives a slightly shorter line width of 17.3 x-heights.

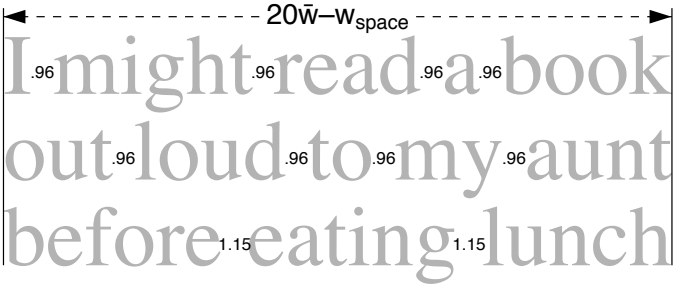

Figure 1: Typesetting MNREAD sentences. Sentences are printed with left and right justification on three lines of text that snugly fit into a box that is 17.3 x-heights wide (i.e., the average width of 20 Times-Roman characters minus the width of a space). Line-to-line spacing is 1em. The width of the between-word spaces (shown as multiples of the normal space width) must be within 0.80 and 1.25.

When the MNREAD sentences are typeset, the spaces between the words on each line are adjusted in order to achieve left-right justification so that the sentence snugly fits into the sentence bounding box (see Figure 1). However, each space may be narrowed to no less than 80% or widened to no more than 125% of w_{space}. In this way we avoid excessively loose or tight spacing between the words on each line which could otherwise impact the legibility of the sentence. Note that in the original specification (Mansfield and Legge, 2007), a line of text was considered acceptable provided the line length was within half an average character's width of the target length, and the required width adjustment was distributed among the spaces on the line. However, in generating sentences for this study we noted that our original specification occasionally resulted in the between-word space being shrunk to less than 20% of its original width so that adjacent words seemed to run into each other. Our modified specification avoids this problem by taking into account the number of spaces over which the width adjustment is distributed.

3. Computer-generated MNREAD sentences

We have built a computer algorithm for composing MNREAD sentences. The algorithm works in two stages: a) a text generator creates candidate sentences; b) the candidate sentences are then filtered to select only those that fit the MNREAD length and layout constraints.

3.1. Sentence templates

Our generator uses sentence templates to create sentences much in the same way that stories are created in the British parlor game *Consequences*. In the game, players choose words or phrases that fit into each placeholder of a template (e.g., "`person`$_1$ met `person`$_2$ at `location`

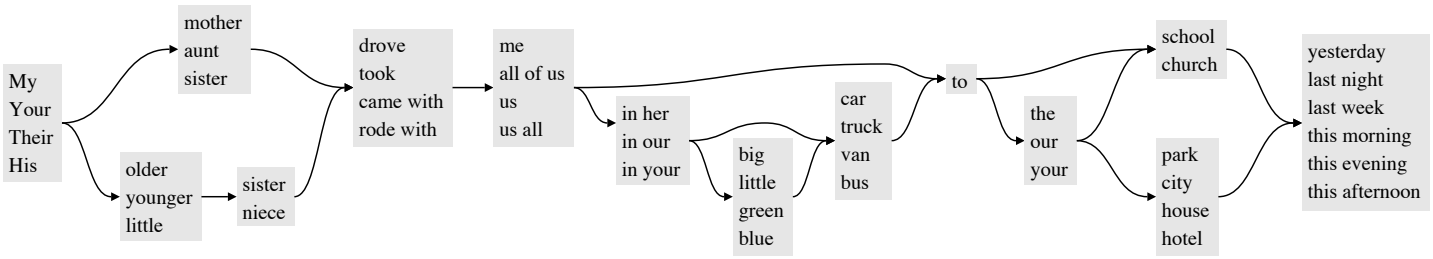

Figure 2: A simplified sentence template. The generator builds a sentence by selecting a word at random from the options in each box, and moving on to the next box in sequence. Where the path branches, it chooses a path randomly. This generator produces sentences like: *My younger sister rode with us in your little van to school yesterday, Their mother drove me in our car to your hotel last week, …* etc.

…") In our sentence generator, each template consists of a sequence of placeholders, each containing a list of possible words that fit into the sentence at that point. Figure 2 shows a simplified template. Sentences are created by randomly selecting a word from the available options at each placeholder. The path through the sentence template can branch to allow a single template to create sentences with different grammatical structures. When a branch is reached during sentence generation, the route taken is chosen randomly.

We have created eleven templates with sentence structures similar to the original MNREAD sentences. Each template is populated with words from the MNREAD lexicon with the appropriate part of speech and meaning that fit into the placeholders in the sentence. Each placeholder requires a large variety of word options, with differing word lengths and word widths in order to increase the possible number of sentences that could subsequently fit the MNREAD length and layout constraints. During development of a template, the distribution of sentence lengths it yields is verified aiming for the target length of 60 characters. If the sentences are too short, extra placeholders can be added to the template to add adjectives to nouns, adverbs to verbs, etc., or to add optional clauses to the start or end of the sentence (e.g., "I thought that …", "I asked my mother if …", "…yesterday evening", "…last week", "…every morning"). The end result is a template that produces sentences with considerable variability in word selection, that on average contain 60 characters.

3.2. Selecting sentences that fit the MNREAD length and layout specifications

Our sentence templates can generate millions of unique sentences. But these sentences do not necessarily fit the MNREAD length and layout constraints. The raw output from our templates contains sentences ranging in length from 18 to 120 characters — typically only one in every 30 sentences has exactly the required 60 characters, and even these 60-character sentences do not automatically fit the MNREAD layout constraints. Thus, the raw output of the sentence templates is filtered to select only the sentences that fit the MNREAD length and layout constraints. The filter, using the width metrics of the Times-Roman font, attempts to fit the 60-character sentences onto three lines of text according to the MNREAD layout constraints. Only 1 in every 8,000 generated sentences fits all these constraints (the remaining non-compliant sentences are discarded).

The sentence generator is written in the Perl programming language, using the `Inline::Spew` module (Schwartz, 1999, 2003).

The generator creates sentences using one template at a time. Initially sentences are output quickly, but the output gradually slows as the generator exhausts the possible combinations. Generally, the generator was stopped once the output had fallen below one sentence per minute. In this way the generator has produced over 9 million sentences. Sample sentences from each of the eleven templates are shown in Figure 3.

4. Reading performance with the computer-generated MNREAD sentences

Our new sentences, by design, match the original MNREAD sentences in vocabulary, length, and layout. The distribution of the number of words per sentence is also very similar for the new versus original sentences: the median for each is 12 words (95% CI [10, 14]). But do the new and original sentences yield similar measures of reading performance? We have measured reading speed as a function of print size using the computer-generated sentences to obtain three parameters of reading performance: *reading acuity* – the smallest print that can just be read, *maximum reading speed* – reading speed when performance is not limited by print size, and *critical print size* – the smallest print that can be read at the maximum reading speed.

We compared the measures obtained with the computer-generated sentences to those obtained using the original MNREAD sentences and to those obtained with shuffled sentences (sentences with randomized word order). These comparisons will allow us to calibrate the readability of the new sentences: ideally, reading performance with the computer-generated sentences will be similar to that with the original MNREAD sentences, and substantially better than that obtained with the shuffled sentences (which have low syntactic and semantic content). We also compared measures obtained with two random selections of

1.	It was raining when his father took us to the swimming pool	My cousin took me in our blue truck to the forest last week	His aunt came with me in her small bus to church last night	Our father drove us in our big car to the woods last summer	Their mother drove us in our red bus to the beach yesterday
2.	The student always wants to read when it is raining outside	Their friend always loves to read at the lake on sunny days	The general always loves to read at the college on vacation	The woman always does not like to eat dinner at the airport	His teacher loves to play catch at school on sunny weekends
3.	I will sing songs to your aunt when we are eating breakfast	You usually like to sort out my clothes before going to bed	I will often need to read the newspaper to your grandfather	They should read a book to your sisters while eating dinner	I could write letters to my mother when we are eating lunch
4.	They want to read a book for our cousin while eating supper	The doctor told me a funny story about animals in the night	The policeman told me not to walk near the trees for a week	He noticed that this man carried the cup from the mountains	I was worried when this young boy told me a story at dinner
5.	I know it is a rather long walk from the tower to the woods	It was almost seven kilometers from the lake to your school	You think there are four farms between the city and the sea	We knew it was an easy train ride from the town to the lake	I knew that it was a long drive from the caves to my school
6.	His cousin said that we have to sort out your bed right now	Our kids asked you all to look at funny books about history	Her daddy said that you must sail to the ships by next week	My family said that both of you have to study serious books	Their granny asked all of us to march to the basement today
7.	My mother ordered the lamb and potato pie for a late supper	His nurse ordered a dish of chicken and bean soup for lunch	Our grandma asked for a plate of turkey and corn for dinner	The teacher wanted a plate of lamb and potato pie for lunch	His father needed a plate of salmon and bean pie for supper
8.	Your kids each had a cup of cold water with their breakfast	Her baby asked for a glass of water and a bowl of ice cream	Your student asked me for a cold glass of juice with dinner	Our neighbors each shared a cold bottle of milk and cookies	Her daddy asked us for a cup of cold tea and some ice cream
9.	Your granny hoped all the soldiers saw the frog on the beds	The policeman told me my children put the seats in the cart	The witch said that my sons wanted the hatch in the kitchen	All the women told us that my aunt saw the boys on the hill	The prince asked if my brothers hid the box on the elevator
10.	Rather than making breakfast she could bicycle to the party	After dreaming you all must sing in his home if it is windy	Besides fishing she will like to walk to the bathroom today	In place of dancing you all might crawl to the home tonight	Instead of watching movies they should march to the capital
11.	The chief has a pet puppy with a white spot on its left paw	The worker keeps a cat with a pink spot underneath its nose	My uncle has a pet bunny with a white patch on its left ear	Our cousin keeps a fish with a tiny red spot over its mouth	My brother keeps a mouse with a black mark on its left side

Figure 3: Example sentences from each of the 11 sentence templates

computer-generated sentences, in order to assess the homogeneity of the computer-generated sentences.

These data were obtained for reading performance with two levels of image blur to simulate reading with different levels of acuity loss, and with unblurred sentences.

5. Method

5.1. Participants

Data were collected from 14 native English speaking undergraduate students, aged between 18 and 30 years, recruited from the University of Minnesota Psychology Department's Research Experience Program participant pool. They gave written consent to participate in accordance with the policies of the University of Minnesota IRB.

5.2. Materials

5.2.1. Sentence sets

Four sets of sentences were used: the original MNREAD sentences, shuffled sentences, and two sets of computer-generated sentences. Each sentence set consisted of 102 sentences, sufficient for 6 'charts' with 17 sentences at different sizes. (Unlike the physical MNREAD charts, the sentences on these charts were displayed one at a time on a computer display.)

Original MNREAD sentences. The original set of MNREAD sentences only contains 95 sentences, so 7 additional sentences (selected from a pool of sentences written during the development of the original MNREAD charts) were added — these extra sentences were placed at the smallest print sizes on each chart.

Shuffled sentences. These were also selected from the pool of additional sentences generated during the development of the original charts. The word order for these sentences was randomized with the constraint that the shuffled sentences fit the MNREAD layout constraints (with the new initial letter in uppercase).

Computer-generated sentences. Two sets of sentences were selected from our corpus of computer-generated sentences. We noted that some of these differed from each other by only one or two words. To avoid using sentences that were similar to each other, we selected sentences by first tagging the five lowest-frequency words (based on their frequency in 3rd-grade text) in each sentence in the corpus, and then

randomly selecting 12 charts of 17 sentences that had no tagged words in common. These 12 charts were randomly assigned to two sets of 6 charts.

5.2.2. Blur

The text images were filtered using low-pass filters designed to mimic different levels of acuity loss (Lei et al., 2016). We tested three blur conditions: no blur, mild blur – using a filter that mimics an acuity limit of 0.6 logMAR (Snellen equivalent of 20/80), and severe blur – using a filter that mimics an acuity limit of 1.2 logMAR (Snellen equivalent 20/320).

5.2.3. Print size

For the no-blur condition, the print sizes ranged from 1.3 logMAR to –0.3 logMAR in 0.1 logMAR steps (x-height angular size = 1.66° to 0.0418°). The viewing distance was 40cm for the 6 largest sizes and was increased to 100cm for the remaining sizes. For the mild and severe blur conditions the print sizes ranged from 1.6 logMAR to 0.0 logMAR (angular size = 3.31° to 0.0833°). The viewing distance was 40cm for all print sizes. Note that no participant was able to read text smaller than 0.3 logMAR with mild blur or smaller than 1.0 logMAR with severe blur.

5.2.4. Display

The sentences were displayed with black text (0.42 cd/m^2) on a white background (432 cd/m^2) on a LCD computer monitor (27"Apple Cinema Display, 2560x1440 pixels, pixel density: 109 ppi, displayed at a frame rate of 60Hz).

5.3. Procedure

Each participant was tested twice in all 12 conditions (4 sentence sets at 3 levels of blur), the order of conditions being chosen randomly. The order of the sentences on each chart was the same for all participants, but the charts were randomly assigned to the blur conditions, so that a chart could be tested with no blur for one participant and with severe blur for another participant. For each condition, reading speed was measured starting at the largest print size, progressing to smaller sizes until no words could be read in a sentence. At each print size, the sentence was displayed and the participant read the sentence aloud. The time interval to read the sentence was recorded from the moment that the sentence was displayed until the participant pressed a computer key to indicate that he or she had uttered the last word in the sentence (or had given up attempting to read the sentence). The experimenter kept track of any reading errors. The display of the sentences, and the recording of reading time were controlled using Psychophysics Toolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007).

6. Results

Reading speed in words per minute (wpm) for each sentence was calculated as follows:

$$\text{reading speed (wpm)} = 60 \times \frac{10 - \text{errors}}{\text{time in seconds}}$$

These data were used to produce plots of reading speed as a function of print size (see Figure 4). Estimates for the maximum reading speed and critical print size for each condition were obtained by fitting the reading-speed versus print-size data with curves of the form:

$$y = \text{MRS} + \frac{\alpha}{2}\left[\, x - \text{CPS} - \sqrt{(x - \text{CPS})^2 - \lambda}\,\right]$$

where y is $\log_{10}$ reading speed (in wpm), MRS is the maximum reading speed, α is the slope of the rising portion of the curve (and was set to 6.0), x is logMAR print size, CPS is the critical print size, and λ controls the sharpness of the roll-over (and was set to 0.001). This curve is convenient for modeling MNREAD data: it provides a good fit to the data and the model's parameters correspond directly to the measures we are interested in, maximum reading speed and critical print size (Cudeck and Harring, 2010). Reading acuity was calculated for each chart as the smallest print size at which any words were read, and this was adjusted by +0.01 logMAR for each reading error. We took the average reading acuity for the two repeated measures for each condition.

Separate mixed-effects models (with participant as a random effect) were then calculated to determine how maximum reading speed, critical print size, and reading acuity were affected by sentence set and level of blur.

6.1. Maximum reading speed

We found main effects of sentence set ($F_{3,143}$=172.2, $p < 0.001$) and blur ($F_{2,143} = 32.5$, $p < 0.001$) on maximum reading speed, but no interaction between sentence set and blur ($F_{6,143} = 0.82$, $p = 0.56$). Figure 5A shows the average maximum reading speed for each condition, along with pairwise comparisons (±95% CI) between each sentence set, and between each level of blur. These data show that the maximum reading speed for the original sentences is 50% faster than for the shuffled sentences ($t_{149} = 20.64$, $p < 0.001$), and 7% faster than for the two sets of generated sentences ($t_{149} = 3.32$, $p < 0.01$). The two sets of generated sentences yield a maximum reading speed that is 40% faster than for the shuffled sentences ($t_{149} = 17.1$, $p < 0.001$). The maximum reading speeds for the two sets of generated sentences are not significantly different from each other ($t_{149} = 0.25$, $p = 0.802$).

Maximum reading speeds for the no blur condition are not significantly different from those for mild blur ($t_{149} = 0.86$, $p = 0.394$), but are 13% faster than for severe blur ($t_{149} = 7.40$, $p < 0.001$). The MRS with mild blur is 12% faster than for severe blur ($t_{149} = 6.55$, $p < 0.001$).

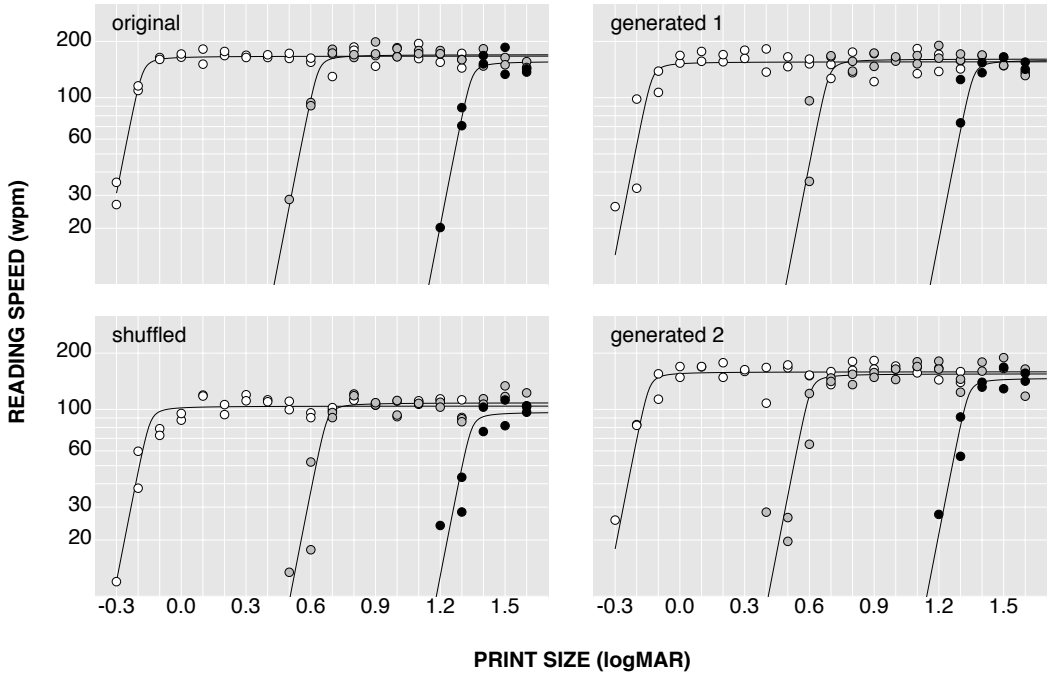

Figure 4: Reading-speed versus print-size data from one participant. Each plot shows data from one sentence set for: no blur (white symbols), mild blur (gray symbols), and severe blur (black symbols), along with their best-fitting curves.

6.2. Critical print size

We found a main effect of blur ($F_{2,143} = 15364$, $p < 0.001$) on critical print size, but no effect of sentence set ($F_{3,143} = 1$, $p = 0.39$), and no interaction between sentence set and blur ($F_{6,143} = 0.50$, $p = 0.81$). Figure 5B shows the average critical print size for each condition, along with pairwise comparisons (±95% CI) between each sentence set, and between each level of blur. Differences in critical print size for the different sentence sets are all less than 0.016 logMAR and are not statistically significant. The critical print size for mild blur is 0.75 logMAR larger than for no blur ($t_{149} = 92$, $p < 0.001$), and the critical print size for severe blur is 0.69 logMAR larger than for mild blur ($t_{149} = 85$, $p < 0.001$).

6.3. Reading acuity

We found a main effect of sentence set ($F_{3,143} = 11.4$, $p < 0.001$) and blur ($F_{2,143} = 19528$, $p < 0.001$) on reading acuity, but no interaction between sentence set and blur ($F_{6,143} = 0.40$, $p = 0.89$). Figure 5C shows the average reading acuities for each condition, along with pairwise comparisons (±95% CI) between each sentence set, and between each level of blur. These data show that reading acuity for the original sentences is 0.036 logMAR smaller than for the shuffled sentences ($t_{149} = 4.25$, $p < 0.001$) and that reading acuity for the generated sentences is on average 0.041 smaller than for the shuffled sentences ($t_{149} = 4.54$, $p < 0.001$). Differences between the reading acuity for the original sentences and the generated sentences, and between the two sets of generated sentences, are less than 0.01 logMAR and are not statistically significant.

Reading acuity for mild blur is 0.75 logMAR larger than for no blur ($t_{149} = 103$, $p < 0.001$), and reading acuity for severe blur is 0.70 logMAR larger than for mild blur ($t_{149} = 96$, $p < 0.001$).

7. Discussion

We have created over 9 million sentences that match the vocabulary, length, and layout properties of the original MNREAD sentences. Our reading-speed measurements demonstrate that the new sentences are also very similar to the original MNREAD sentences in terms of reading performance. Estimates of reading acuity and of critical print size obtained with the new sentences match those obtained with the original sentences. The similarity between the sentence sets for these measures extends over a more than 10-fold range of acuity. Estimates of maximum reading speed with the new sentences are 7% slower than with the original MNREAD sentences, indicating that the computer-generated sentences are slightly more difficult to read than the original sentences. This reading speed deficit is likely due to how the sentences are constructed, where the large choice of options for the placeholders in the template results in word selections that have low predictability given the context of the rest of the sentence. But the 7% difference in maximum reading speed between the new and original sentences is slight compared to the 50% difference we measure between original and shuffled sentences.

Comparing reading performance with the two sets of computer-generated sentences shows that differences in

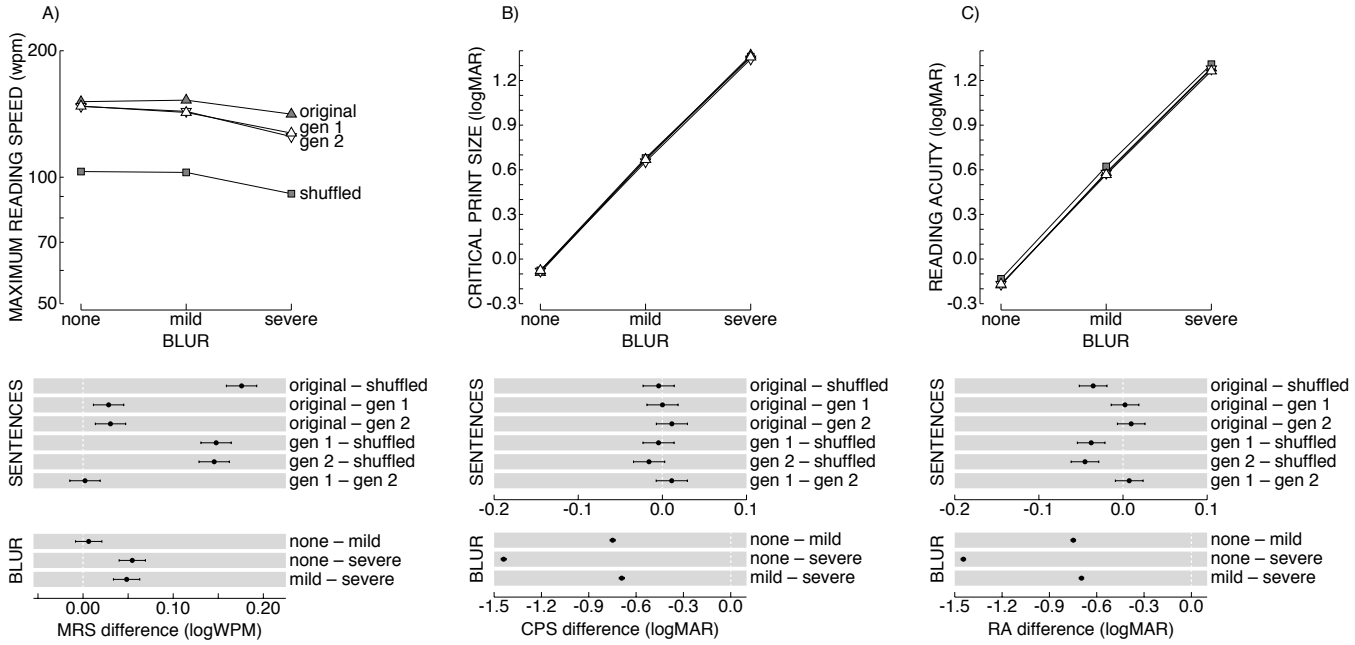

Figure 5: A) Top: Maximum reading speed as a function of blur for: original sentences (filled triangles), two sets of generated sentences (white triangles), and shuffled sentences (filled squares). Bottom: Differences (and 95% confidence intervals) in maximum reading speed for each pair of sentence sets and each pair of blur levels. Confidence intervals that overlap zero indicate the difference is not statistically significant. B) Top: Critical print size as a function of blur for the different sentence sets. Bottom: Pairwise differences in critical print size. C) Top: Reading acuity as a function of blur for the different sentence sets. Bottom: Pairwise differences in reading acuity.

maximum reading speed, critical print size, and reading acuity are less than 0.01 log unit. This indicates a high degree of consistency in reading performance obtained with different samples of generated sentences.

Our data show 50% faster maximum reading speeds for the original sentences than for the shuffled sentences. This difference is similar to the context advantage reported in previous studies for reading static text (e.g. Bullimore and Bailey, 1995; Lueck et al., 2000; O'Brien et al., 2000; Sass et al., 2006) and drifting text (Fine and Peli, 1996), but considerably smaller than the context advantage reported for RSVP reading (e.g. Latham and Whitaker, 1996; Chung et al., 1998). A separate finding in this study is that reading acuity measured with the shuffled sentences is about 0.04 logMAR larger than when measured with the original or computer generated sentences. This difference has not been noted in prior studies although it is evident in the data presented by Lueck et al. (2000) whose participants were unable to read random words at sizes smaller than -0.2 logMAR but were able to read meaningful sentences at that size. This seems reasonable in that shuffled sentences contain fewer syntax and context cues that could help a reader identify hard-to-see words.

The effects of blur in our study are largely consistent with our expectations: the acuity size and the critical print size increase roughly in proportion to the level of blur added to the stimuli. However, our data show a significant decrease in maximum reading speed with severe blur. Some of this decrease could be an artifact of the curve-fitting procedure – typically, the participants read fewer sentences in the severe blur condition, so that only three or four measurements were obtained at print sizes larger than the critical print size (the sizes over which the maximum reading speed is defined) which might lead the curve fit to underestimate the maximum reading speed. However, this does not account for the entire reduction in our maximum reading speed estimates because the effect is still there (albeit to a lesser extent) if we restrict the data at the other levels of blur to just the six smallest print sizes that were read by each participant. Of course, the slower maximum reading speed for severe blur is also consistent with the reduction in reading speeds that has been reported for character sizes larger than 2 degrees (e.g. Legge et al., 1985).

Overall, these data indicate that our computer-generated sentences give reliable measures of reading performance that match, or are very similar to, those obtained with the original MNREAD sentences.

7.1. Limitations

A potential problem in the use of our generated sentences is that, while no two sentences are identical, many of them differ from each other by only a few words. It would be undesirable to have sequences of similar sentences in most typical testing situations. To explore this issue, we have quantified the extent to which the MNREAD sentences are similar to each other. We calculated sentence similarity for pairs of sentences, s_1 and s_2, as the number of words in s_1 that are also in s_2. We calculated sentence similarity for all pairwise comparisons of the 95 original

MNREAD sentences and for all pairwise comparisons of a random sample of 1100 computer-generated sentences (100 from each sentence template). As anticipated, the generated sentences have higher similarity scores than the original sentences: the mean sentence similarity is 1.76 words [95% CI (0, 6)] for the generated sentences and just 1.54 words [95% CI (0, 4)] for the original sentences. This shows that attention needs to be paid to sentence similarity when choosing sentences from the extended corpus for use in a study. It is straightforward to select sets of sentences from the extended corpus purposely to reduce the similarity between sentences (as described in section 5.2.1).

Another limitation of the generated sentences is that, due to the way they are constructed, some of the sentences contain semantically unpredictable words that may impact reading performance (they are read 7% slower than the original sentences). During the development of the original MNREAD charts, reading times for candidate sentences were obtained in a pilot study so that any that produced unusually long or short reading times could be discarded. A similar process is infeasible for our corpus of over 9 million computer-generated sentences, leaving open the possibility that there will be more variability in the reading speeds obtained with the sentences. Indeed, participants in our study occasionally reported that some sentences seemed easier to read than others. We recommend that researchers take this into account when using these sentences for their studies. For example, the generated sentences could be screened for number of words, number of syllables, or for word frequency, in order to avoid sentences that might produce atypical reading speeds.[1]

Currently our sentence generator only has sentence templates to create sentences in English. There has been considerable interest in MNREAD charts for other languages and versions have been developed in Japanese, Italian, Portuguese, French, Spanish, Turkish, and Greek, but we currently do not have sentence generators for these languages.

7.2. Applications of the sentence generator

In addition to increasing the number of sentences available for MNREAD testing, our sentence generator allows us to create sentences that examine variations of the MNREAD constraints. For example, the sentence layout algorithm, which uses the font metrics for Times-Roman, can be modified to create MNREAD sentences for any font. This could be used to create standardized to assess readability versus print-size for new fonts or for applications that require specific fonts (e.g., road signs, military applications, etc.) Further, sentences can be created that simultaneously fit the MNREAD constraints for multiple fonts. Xiong et al. (2018) used our computer-generated sentences to compare reading performance using two new fonts designed for patients with central vision loss to reading performance using Times, Helvetica, and Courier. The generated sentences allowed for the same sentences to be used for all five fonts while also equating for sentence length and sentence layout.

Another possibility is to select sentences from the corpus that have a specific number of words, or number of syllables, or that have other properties of interest. For example, Mansfield et al. (2018) used the new sentences to show that the critical print size was linked to letter recognition by measuring reading performance for sentences that differed in the number of easy-to-recognize letters that they contained.

The sentence layout parameters can also be modified to create sentences with different line lengths (i.e., rather than having the 60 characters on 3 lines of left-right justified text, we can generate sentences that format onto 1, 2, 3, 4, or 5 lines.) This manipulation is useful for testing reading performance in specific situations where text layout is constrained by the reader's device, tablet, smart phone, desktop display, etc. (Atilgan et al., 2017).

In summary, we have created a sentence generator that has yielded a large set of MNREAD sentences. This has substantially expanded the reading materials for clinical vision research using the MNREAD test, and opens up new possibilities for measuring factors that affect how reading depends on text parameters. These sentences are available for download from <insert URL here>, and we hope they will be useful to other researchers.

Appendix A. Appendix

References

Atilgan, N., Mansfield, J. S., Legge, G. E., 6 2017. Impact of line length on reading performance for normal vision and simulated acuity reduction. Vision 2017 Conference, The Hague, Netherlands.

Brainard, D. H., 1997. The psychophysics toolbox. Spatial vision 10, 433–436.

Bullimore, M. A., Bailey, I. L., 1995. Reading and eye movements in age-related maculopathy. Optometry and vision science 72 (2), 125–138.

Calabrèse, A., To, L., He, Y., Berkholtz, E., Rafian, P., Legge, G. E., 2018. Comparing performance on the mnread ipad application with the mnread acuity chart. Journal of vision 18 (1), 8–8.

Carver, R. P., 1976. Word length, prose difficulty, and reading rate. Journal of Reading Behavior 8 (2), 193–203.

Chung, S. T., Mansfield, J. S., Legge, G. E., 1998. Psychophysics of reading. xviii. the effect of print size on reading speed in normal peripheral vision. Vision research 38 (19), 2949–2962.

Crossland, M. D., Legge, G. E., Dakin, S. C., 2008. The development of an automated sentence generator for the assessment of reading speed. Behavioral and Brain Functions 4 (1), 14.

Cudeck, R., Harring, J. R., 2010. Developing a random coefficient model for nonlinear repeated measures data. In: Chow, S. M. Ferrer, E., Hsieh, F. (Eds.), The Notre Dame series on quantitative methodology. Statistical methods for modeling human dynamics: An interdisciplinary dialogue. Routledge/Taylor & Francis Group, New York, NY, US, pp. 289–318.

[1] We thank an anonymous reviewer for this suggestion.

Table A.1: Relative frequency of characters calculated from 11,000 60-character sentences generated by the sentence templates. Letters not shown in this table (i.e., many uppercase letters) do not occur in the generated sentences.

character	frequency	character	frequency	character	frequency
A	0.000847	a	0.063750	n	0.045500
B	0.000406	b	0.011152	o	0.062092
H	0.001652	c	0.019777	p	0.015286
I	0.002164	d	0.030624	q	0.000179
M	0.001274	e	0.095391	r	0.052750
O	0.001215	f	0.016415	s	0.052041
R	0.000195	g	0.014529	t	0.075289
S	0.000306	h	0.051495	u	0.027235
T	0.006711	i	0.048806	v	0.004288
W	0.000102	j	0.000711	w	0.020733
Y	0.001795	k	0.012127	x	0.000273
		l	0.033785	y	0.016798
space	0.194365	m	0.017689	z	0.000252

Fine, E. M., Peli, E., 1996. The role of context in reading with central field loss. Optometry and vision science: official publication of the American Academy of Optometry 73 (8), 533–539.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., Broussard, C., et al., 2007. What's new in psychtoolbox-3. Perception 36 (14), 1.

Kučera, H., Francis, W. N., 1967. Computational analysis of present-day American English. Dartmouth Publishing Group.

Latham, K., Whitaker, D., 1996. A comparison of word recognition and reading performance in foveal and peripheral vision. Vision Research 36 (17), 2665–2674.

Legge, G., Pelli, D. G., Rubin, G. S., Schleske, M. M., 1985. Psychophysics of reading. i. normal vision. Vision Research 25, 239–252.

Lei, Q., Kersten, D., Thompson, W., Legge, G. E., 2016. Simulating reduced acuity in low vision: Validation of two models. Investigative Ophthalmology & Visual Science 57 (12), 634–634.

Lueck, A. H., Bailey, I. L., Greer, R., Dornbusch, H., 2000. Magnification needs of students with low vision. In: Arditi, A., Horowitz, A., Lang, M. A., Rosenthal, B.and Seidmans, K. (Eds.), Vision rehabilitation in the 21st century. Swets & Zeitlinger, Downington, PA, US, pp. 311–313.

Mansfield, J. S., Ahn, S. J., Legge, G. E., Luebker, A., 1993. A new reading-acuity chart for normal and low vision. Ophthalmic and Visual Optics/Noninvasive Assessment of the Visual System Technical Digest 3, 232–235.

Mansfield, J. S., Legge, G. E., 2007. The mnread acuity chart. In: The Psychophysics of Reading in Normal and Low Vision. Lawrence Erlbaum Associates, Mahwah, NJ, Ch. 5, pp. 167–191.

Mansfield, J. S., West, T., Dean, Z., Sep 2018. Is the critical print size for reading linked to letter recognition? Journal of Vision 18 (10), 1163–1163.
URL `http://dx.doi.org/10.1167/18.10.1163`

O'Brien, B. A., Mansfield, J. S., Legge, G. E., 2000. The effect of contrast on reading speed in dyslexia. Vision research 40 (14), 1921–1935.

Pelli, D. G., 1997. The videotoolbox software for visual psychophysics: Transforming numbers into movies. Spatial vision 10 (4), 437–442.

Perrin, J.-L., Paillé, D., Baccino, T., 2015. A new sentence generator providing material for maximum reading speed measurement. Behavior research methods 47 (4), 1055–1064.

Rassia, K. E. K., Pezaris, J. S., 2018. Improvement in reading performance through training with simulated thalamic visual prostheses. Scientific Reports 8 (1), 16310.
URL `https://doi.org/10.1038/s41598-018-31435-0`

Sass, S. M., Legge, G. E., Lee, H.-w., 2006. Low-vision reading speed: Influences of linguistic inference and aging. Optometry and Vision Science 83 (3), 166–177.

Schwartz, R., 1999. Writing nonsense with perl.
URL `http://www.linux-mag.com/id/309/`

Schwartz, R., 2003. Inline::spew (perl module computer source code).
URL `http://search.cpan.org/ merlyn/Inline-Spew-0.02`

Xiong, Y., Lorsung, E., Mansfield, J. S., Bigelow, C., Legge, G. E., 2018. Fonts designed for macular degeneration: Impact on reading. Investigative Ophthalmology & Visual Science 59 (9), 2562–2562.

Xu, R., Bradley, A., 2015. Iuread: a new computer-based reading test. Ophthalmic and Physiological Optics 35 (5), 500–513.

Zeno, S., Ivens, S. H., Millard, R. T., Duvvuri, R., 1995. The educator's word frequency guide. Touchstone Applied Science Associates.