Behav Res (2015) 47:1055–1064
DOI 10.3758/s13428-014-0521-8

A new sentence generator providing material for maximum reading speed measurement

Jean-Luc Perrin · Damien Paillé · Thierry Baccino

Published online: 16 October 2014

Abstract A new method is proposed to generate text material for assessing maximum reading speed of adult readers. The described procedure allows one to generate a vast number of equivalent short sentences. These sentences can be displayed for different durations in order to determine the reader's maximum speed using a psychophysical threshold algorithm. Each sentence is built so that it is either true or false according to common knowledge. The actual reading is verified by asking the reader to determine the truth value of each sentence. We based our design on the generator described by Crossland et al. and upgraded it. The new generator handles concepts distributed in an ontology, which allows an easy determination of the sentences' truth value and control of lexical and psycholinguistic parameters. In this way many equivalent sentence can be generated and displayed to perform the measurement. Maximum reading speed scores obtained with pseudo-randomly chosen sentences from the generator were strongly correlated with maximum reading speed scores obtained with traditional MNREAD sentences (r = .836). Furthermore, the large number of sentences that can be generated makes it possible to perform repeated measurements, since the possibility of a reader learning individual sentences is eliminated. Researchers interested in within-reader performance variability could use the proposed method for this purpose.

J.-L. Perrin · T. Baccino
CHArt/LUTIN (EA 4004), Paris 8 University, Paris, France

J.-L. Perrin (✉) · D. Paillé
World R&D Optics Department, Vision Science Department, Essilor International, 13 rue Moreau, Paris 75012, France
e-mail: perrinjl@essilor.fr

Keywords Reading · Speed · Measurement · Sentence · Generator · Psychophysics

Introduction

The present paper focuses on a method to measure maximum reading speed for adult readers. Researchers of different fields study the activity of reading at various levels: from perceptual aspects, such as letter and word recognition (Pelli, Burns, Farell, & Moore-Page, 2006; Sereno & Rayner, 2003), to higher cognitive processes, like comprehension (Snow, 2002). One way to address the problem of reading measurement is to evaluate reading speed, which reflects the efficiency of information gathering and processing. This efficiency varies with many factors. For example, reading speeds vary between readers: they are measured in educational science and in developmental psychology in relation to the level of education and the level of cognitive development (Landerl & Wimmer, 2008). Reading speeds can also vary within readers. As an example, vision scientists, optometrists, and ophthalmologists may be interested in assessing visual function over time in patients affected by low vision pathologies, or measuring the result of some therapeutic intervention (Ahn, Legge, & Luebker, 1995; Seiple, Szlyk, McMahon, Pulido, & Fishman, 2005; Trauzettel-Klosinski, Dietz, & the IReST Study Group, 2012). In addition, reading speed data can be collected to observe changes due to variation in the text itself or in its presentation format. Psychologists and psycholinguists have shown that the number of words read in one minute varies with textual aspects such as lexical frequency (Kliegl, Grabner, Rolfs, & Engbert, 2004) and text difficulty (Just, Carpenter, Keller, Eddy, & Thulborn, 1996). How a text is displayed (Dyson, 2004) as well as its support (Dillon, 1992) are other factors which affect reading speed. The wide use of reading speed as a metric for experimental research

calls for the development of efficient tools to obtain a valid and reliable measurement.

The main objective of this paper is to provide a means to evaluate reading using maximum reading speed measurements. Ultimately this should offer reliable measurements to quantify the effect of an experimentally manipulated factor (e.g., comparison of the level of several readers, effect of pathology on reading speed over time, differences in reading performance between two interfaces…). To be successful, the method must address four points: (1) the method must allow the study of reading alone. Thus it is important to exclude other processes such as phonological production and address silent reading; (2) the reading material must also reflect the readers' ecological experiences. Thus it should be composed of complete and structured sentences. For evaluating processing speed with highly controlled material, one could use a corpus of isolated words. This solution is however associated with single word identification, which is different from real-life reading (Latham & Whitaker, 1996); (3) the reading of any one sentence from the material should be equivalent to any other. Complex comprehension processes must also be avoided because they may alter the reading pattern depending on the difficulty (Staub, 2010) and previous knowledge of the reader (Kendeou & van den Broek, 2007); and (4) the material should contain enough texts to allow repeated measurements while preventing the reader from learning a text and thus reading faster.

Experiencing the same need, Crossland, Legge, and Dakin (2008) developed an automatic sentence generator to provide thousands of sentences intended for reading speed assessment. Their algorithm randomly selects a quantifier, an object, and a trait to form a grammatically valid sentence. Some examples of generated sentences are the following: "No chimps have feathers", "Some comedians are unemployed"… The process of generating sentences can be repeated to obtain many texts. These sentences can be used like psychophysical stimuli: sentences can be displayed for different durations and the ability to read the whole sentence at each rate can be quantified. Crossland et al. (2008) used a method of constant stimuli, displaying sentences at durations ranging between 0.016 and 0.5 seconds per word. The percentage of correctly read sentences was plotted as a function of the duration, and a curve was fitted to these data. A threshold of performance is determined (e.g., 80 %) and the score that corresponded to the intercept between the psychometric curve and the threshold value was calculated to obtain the maximum reading speed. Effective reading was assessed through a comprehension task. Each sentence was either true or false, depending on its elements. The reader simply stated whether the sentence he read was true or false. This allows silent reading to be evaluated without oral recitation. As the content of each sentence was rather easy to understand, a wrong answer should have denoted the inability to read the sentence at a given display rate.

Their work was an important advancement in the ability to perform repeated measurements of reading speed. This is a key point for within subject experimental design, or for the screening of visual function over time. However, their method can be enhanced further. Crossland et al. (2008) made some readers determine the truth value of generated sentences. Some wrong answers were observed, even in a condition with unconstrained reading time (between 2 % and 17 % of wrong determination for each reader). These wrong answers cannot be due to the display duration. Therefore they must be due to the readers' difficulties in inferring the truth values which are determined by the generator. Such difficulties could be the result of semantics or of the automatic attribution of truth values. Both of these aspects are dealt with in the following sections (see respectively "Choice of the objects" and "Truth value handling"). When adopting a psychophysical approach, this phenomenon can bias the estimate of maximum reading speed: a psychometric curve should ideally reach the 100 % proportion of correct response shortly after having reached the perceptual threshold. Wrong determination of the truth value of sentences when they are fully read can thus distort the curve and bias the threshold estimation.

The aim of our work was to improve this assessment method by refining the generator. The main improvement of our new sentence generator is the increase of sentences' homogeneity. This is made possible by a stronger control on psycholinguistic variables. Since lexical frequency has been shown to modulate reading speed, and valence and concreteness values can alter speed and comprehension (Egidi & Gerrig, 2009; Sadoski, 2001), we assumed that controlling these variables could improve reliability. The second improvement is a novel method to determine the truth values of the sentences. This method is used in order to allow the reader to determine unambiguously the truth values and thus to avoid wrong answers when the sentences are correctly read. An additional benefit of the method is that it facilitates the development of the corpus of sentences. Finally, it is necessary to validate the generated sentences as reliable material for reading speed assessment on a large number of adult readers. The upgraded generator produces French sentences for practical reasons, though the principle could be easily applied to any language which is syntactically close to French.

Operating principles of the sentence generator

As in Crossland et al.'s method (2008), the purpose of the new generator is to generate sentences composed of a quantifier, an object, and a property. Each sentence has to be true or false. The reader must be able to effortlessly determine the truth or falsehood of the sentence since we are only interested in reading capability, and not comprehension aspects. If the reader succeeded in reading the whole sentence, he should

be able to give the correct answer; otherwise, he will give a random answer. In addition, the sentences must be lexically and comprehensively equivalent, so that the ability to respond correctly is only affected by the display duration of the sentence.

Handling of truth values

Each generated sentence must have a truth value ("True" or "False"), according to general knowledge. The truth value could be directly assigned to every sentence by hand. It would nevertheless be extremely long and repetitive, since the aim of the generator is to provide a large number of sentences. We propose a method which can determine the truth values with little effort.

Crossland et al. (2008) proposed to add a "trait" following the quantifier and the object. The trait is a two-word description. The first word is a verb, the second one can be an adjective, or another word which gives sense to the phrase when following the verb ("is democrat", "design buildings"…). Our version of the generator is different. A corpus constituted of several common nouns is considered. All the nouns in this pool were structured in an ontology (see Fig. 1 for an overview example), using the Protégé 4.1 software (Stanford Center for Biomedical Informatics Research, Stanford, CA, USA). If an object is generally considered to be part of a class represented by another noun, it is specified in the ontology (e.g., "dog" is part of the "animal" class). The structure of the ontology will allow the attribution of truth values. Thus, the classification must be universally accepted.

We refer to "classes" for concepts which include at least one other concept; "entities" represent concepts which are included in a class but which, themselves, do not include any other concept. The main difference between class and entity is detailed in the next subsection. To build sentences, we consider couples of elements within the pool. Each couple must contain at least one entity. These couples will serve to build parts of sentences as follows: "*Element 1* is an *Element 2*". The couples have a "proto-truth value", which is not limited to a boolean choice (i.e., true or false). Some couples are always true (denoted "T"): "a dog is an animal", whatever the "dog". Others are always false ("F"): "a dog is a bird". But element couples can also be true sometimes and false in other instances ("T/F"): "an animal is a dog", which is true for some "animals", but not for all. We use the position of the two elements in the ontology to automatically determine the proto-truth value of the couple. If the first element is a descendant of the second one, the couple is always true. In contrast, if it is the second element which is the descendant, the couple is T/F. Lastly, if the two elements are neither an ascendant nor a descendant of each other, the couple is always false.

Due to the tree structure of the ontology, an entity can have only one direct ascendant. This structure cannot handle couples made up of words with several uses (a word which denotes one concept, but this concept could be included in several classes). We must specify in the ontology the relationships which are forbidden. The two elements will then simply not be given in the same sentence (see Fig. 2 for an example).

By pairing each element of the ontology with all other allowed elements, we can generate a lot of couples and their proto-truth values. We need to ensure that the final sentences will have a binary truth value (T or F, and nothing else). A quantifier must be added before the couple to disambiguate the proto-truth value. We use the same quantifiers as Crossland et al. (2008): "No", "Some," and "All". We propose to determine the truth values, depending on the quantifier and the proto-truth value of the couple. With three quantifiers, and three proto-truth values, there are nine possible combinations. Yet we do not consider every combination for two reasons.

(1) There is no absolute equivalence between logical and natural languages. An example is a combination of the quantifier "some" (translated to ∃ in logical symbolism) and a couple with a true proto-truth value ("a dog is an animal"). The sentence "Some dogs are animals" is true according to first order logic. ∃x (Dog(x) ∧ Animal(x)) is a true proposition because there is at least one thing in the world which is at the same time a "dog" and an "animal". But in natural language, this sentence could implicitly mean that some "dogs" are not "animals". Such sentences can, at best, complicate comprehension, and, at worst, make the reader answer incorrectly, even if he

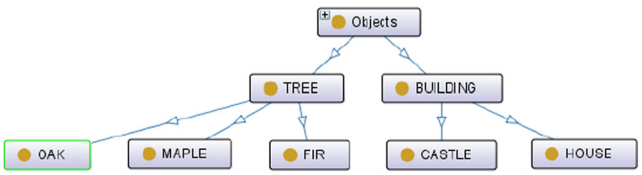

Fig. 1 Example of a part of the semantic tree. Concepts at the bottom correspond to the entities, higher ones are classes. The top concept "Objects" includes every element of the ontology. Example words are given in English for illustration purposes. This visualization and the following one were obtained with the Protégé 4.1 software

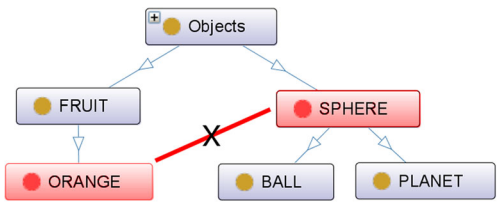

Fig. 2 Example of a concept with two uses in the ontology. The word "orange" is considered in the ontology as being a "fruit". The sentence "an orange is a fruit" is true. But in common knowledge, "an orange is a sphere" is also true. As the ontology handles only one use for each word, we specify that the generator does not create sentences with these two words

or she was able to read the text. To avoid this issue, sentences with this proto-truth value and this quantifier must not be generated.

(2) Some combinations which could be translated from logical form to natural language were also suppressed, in order to balance the final truth values. Thanks to these suppressions, the probability for the sentence to be true is always 50 %, whatever the quantifier or the proto-truth value of the couple (see Table 1 for all possible combinations). Thereby one cannot increase their chances of guessing the truth value of sentence having only seen a part of it.

An advantage of the automatic attribution of truth values to the sentences through the ontology is that it makes the generation process easier and faster: all possible couples are selected; the appropriate quantifiers are attributed, leading to non-ambiguous sentences in terms of truthfulness. This is important in order to prevent the readers from having difficulties in determining the truth value of the sentence. Thus it should limit the occurrence of wrong answers when the sentence is fully read. Any change in the ontology is directly taken into account in the generation of the sentences.

Choice of objects

The previous subsection explained how to combine the words in the pool; here we describe the selection process to fill it with common nouns. As the lexical frequency of words has an important impact on reading speed (Kliegl et al., 2004) and the sentences have to be equivalent, the selected words must have a high lexical frequency. The mean number of occurrences of the words in our pool per million words was 144.9, which corresponds to a mean of the log10 values of 1.80 ($SD = 0.56$), according to the "Lexique 3" database (New, Pallier, Ferrand, & Matos, 2001). Every word belongs to the 7,000 most frequent words in the French language. Compound and uncountable nouns were also excluded.

Table 1 Truth table of generated sentences depending on the element couples and the selected quantifier

	Proto-truth value of the couple			
	T	T/F	F	p(True)
Quantifier				
"All"	T	F	(f)	0.5
"Some"	x	T	F	0.5
"No"	F	(f)	T	0.5
p(True)	0.5	0.5	0.5	

Note. The truth values between brackets represent combinations removed to balance the probability of sentences being true. The value represented by an "x" was not used as it is ambiguous

In the ontology, "entities" are words which correspond to concepts with at least one ascendant and no descendant. These words must be easy to understand. If not, the readers could misunderstand the whole sentence and then wrongly determine its truth value, even if the sentence is fully read. Abstract words are also to be avoided, so that the reader does not have to analyze and interpret ambiguous concepts. Certain variables exist in psycholinguistics that allow the control of such properties. Only words with a high concreteness value and a neutral or positive emotional valence were selected, based on the measurements of Bonin et al. (2003).

For the classes—i.e., words which have at least one descendant—the lexical frequency was the only controlled value. As these words must encompass more concepts, they can be less specific (e.g., the concept "Animal"). A class can eventually become an entity as well, on condition that it respects the concreteness and valence values chosen for the entities and that it has at least one ascendant. The pool was composed of 65 words in total.

Sentence generation and reading

To generate the sentences, all possible couples are selected taking into account the aforementioned rules. The proto-truth value for each couple is obtained as explained earlier. Then all the appropriate quantifiers are chosen; an application of the table of truth (Table 1) gives the final truth value of the sentence. The sentence is comprised of a quantifier, a common noun, the verb "to be," and a second common noun. The last step is to add an appropriate article to the second noun, and to apply grammar and conjugation rules to obtain a well-constructed sentence, with an allocated truth value. From the pool of 65 words previously described, and given the forbidden relationships specified in the ontology, we could generate almost a thousand sentences. A moderate amount of concepts in the ontology can thus generate a large amount of sentences, which limits the risk of a sentence being displayed several times. A sentence would repeat every thousand times, and this frequency could be lowered still by including some more words in the ontology.

In addition to the words with several uses previously described, some words can have several meanings (one word denotes several concepts): i.e., homonyms and polysemous words. Those for which the different meanings are commonly used should be avoided (e.g., "ring", which denotes both the sound and the object). But many words have one major meaning and one or several minor meanings; it would be pointless to try to exclude them all. As an example, in our system the sentence "All banks are institutions" is true. However, the word bank can denote the financial institution, but also—in physical geography—a slope bordering water. This cannot be directly handled by the ontology. We advise

test administrators to instruct the reader to consider the first and most common sense of words during the evaluation.

An example of each quantifier and each proto-truth (p-t) value is given in the following sentences with the associated final truth (ft) value (direct translation from French generated sentences): “A train is not an animal” (p-t: F, ft: T) , “Some trains are animals” (p-t: F, ft: F), “Some flowers are roses” (p-t: T/F, ft: T), “All flowers are roses” (p-t: T/F, ft: F), “An orange is not a fruit” (p-t: T, ft: F), “All oranges are fruits” (p-t: T, ft: T).

Validation of the generated sentences in a reading speed evaluation

The generator is able to produce a large number of sentences. The reading of these sentences must be confirmed as being a valid assessment of maximum reading speed. The method of answering with the true or false task must also be confirmed as effectively equivalent to an oral recitation of the sentence silently read.

Protocol

Participants

To be eligible, participants must have a good binocular visual acuity when reading on a computer screen of at least Log MAR = 0, wearing their single vision optical correction if needed. Wearing bifocal or progressive lenses was not allowed in order to prevent alteration in visual performance related to the position of the word in the lenses. Subjects’ acuities were measured with an Optoprox (Essilor International, Créteil, France) at the distance of the test (44 cm from the screen). Subjects were native French speakers and between 18 and 45 years of age. The limit of 45 years was set to avoid as much as possible including persons with presbyopia, and thus the wearing of progressive or bifocal lenses. Any potential subject with visual pathology or pathology likely to interfere with the study was excluded. Subjects for this validation were directly recruited by the experimenters. All participants signed an informed consent form. Forty-five participants took part in the experiment. They were 23 women and 22 men. Mean age was 32 years (*SD*: 7), ranging from 22 to 45.

Apparatus

Sentences were displayed on an E2311H screen (Dell, Round Rock, TX, USA) with a resolution of 1280*800. Letters were displayed in black on a white background, corresponding to a contrast of 98.5 % (Michelson) in ambient light. The background luminance was set to 88.6 cd/m^2, as measured with a Cal-SPOT 401 (The Cooke Corporation, Romulus, USA). The subjects were positioned on a head-and-chin rest so that a viewing distance of 44 cm was maintained. The Arial font was used and the font size was 18 px. With this set up, a lowercase letter “o” had a diameter of 16.6 min of visual angle (which corresponded to approximately 0.5 Log MAR). The displaying software was written in Python 2.7.6 with the module PyGame 1.9.1.

Reading tests

The test validation consisted of ensuring that the maximum reading speed score obtained by reading sentences from the generator was predictably related to a score obtained using sentences from a test recognized in the literature. We chose the same reference test as Crossland et al. (2008) when validating their own generator, i.e., the MNREAD (Ahn et al., 1995; Legge, Ross, Luebker, & LaMay, 1989). This test serves to evaluate oral reading speed mainly, but can also be used for silent reading. We used a French version of the MNREAD (Senécal, Gresset, & Overbury, 2006) since the sentences of our system generated to cater for the French participants. All subjects read the MNREAD sentences and the sentences generated by our algorithm. The latter was performed twice: once with an oral recitation of the reading and once giving a “True or False” answer. The order of the three tests was pseudo-randomized for each subject. Each measurement was made binocularly.

Procedure and measurements

For each of the three tests, several sentences were displayed for a limited period of time. Each sentence display was preceded by a fixation cross at the location of the first letter. When the sentence disappeared, it was followed by a mask. Both the cross and the mask were displayed for one second. To verify effective reading, the subject then had to orally recite the last four words of the sentence of the MNREAD; the readers could also recite the whole sentence if they preferred, but only the effective reading of the four last words was taken into account. For the generated sentences, the subject had to recite the whole sentence or indicate whether the content was true or false. If they recited all requested words with no error (oral tests), or if they gave the right truth value (true or false test), the response was considered to be correct. The subjects could take as much time as they required to give their answer after the display. For the MNREAD, each sentence was randomly selected among all the sentences of this test and was displayed only once per reader. Each generated sentence of the test was pseudo-randomly selected among all the generated sentences. A weighting factor was applied so that each

proto-truth value (False, True/False, True) had the same probability of being selected as the others.

The variable used for modulating the stimulus level was the display period per word in milliseconds. As an example, if the sentence was five words long, and the variable was set to 200 ms, the whole sentence was displayed for 1 s. The test score is the minimum duration of display per word which allows the subject to read the whole sentence. This measurement can easily be interpreted in the classical words per minute metric, and thus as a maximum reading speed score.

A staircase method has been used to obtain the minimum display duration score: stochastic approximation (Robbins & Monro, 1951; Treutwein, 1995). For the two tests requiring oral recitation, the performance level to converge on was set to 50 %: we expect the reader to be unable to recite the sentence when the display duration was not long enough to allow the reading. For the true or false test, as the reader still has a 50 % probability to give the correct answer by chance even if the display duration is too short, we set the threshold to 75 %. The other parameters used for the staircase were computed from the experimental data of ten subjects who took the tests using the constant stimulus method: several staircases with different sets of parameters were simulated by using the observed data. The parameters of the staircases which attained the closest scores in comparison to those obtained with the constant stimulus method were selected. All the selected parameters for the three tests are given in Table 2. The minimum display duration per word score was computed from the mean display durations of the last six reversals of the staircase.

Each test was preceded by an instruction and a familiarization phase. The reader was informed that a cross would appear, on which he/she would have to fixate, a sentence would then appear, and it should be read as quickly as possible. It was explained that the display duration would vary from one sentence to another and that it was normal to not be able to read the whole sentence. For the MNREAD, the reader was asked once the sentence had disappeared to recite the last four words he/she could read or the whole sentence, but was notified that only the last four words would be taken into account. For the generated sentences with oral recitation, the reader was asked to recite the whole sentence. For the other modality, the reader was asked if the sentence was true or false, he/she had to give an answer even if the sentence was not read. Before every test two sentences were displayed, one for a long duration and the other for an extremely short duration. For the True or False modality, two supplementary sentences were displayed: one for which the truth value was false, and one with a homonym so that we could explain that only the major meaning of the word should be taken into account.

Data analysis

Once the three scores of each subjects were obtained, we focused on the comparison between the scores of the MNREAD and of the generated sentences with oral recitation to verify that the reading of the generated sentences provides a measure which is related to the reading performance observed with the classical test. Then we examined the agreement between the two modalities (oral and True/False) of our test. The aim of the new generator is to obtain sentences for which the truth value can be unambiguously determined. Thus the score obtained with the oral recitation should equal the one obtained with the "True or False" modality.

A common means to determine the agreement between two measurement methods is the Bland-Altman plot (Bland & Altman, 1986): one should first obtain a data set of the same objects measured using both methods. For each pair of measurements, the signed difference between the two scores is plotted against their average value. Nevertheless this technique is criticized because it can present a bias which is not due to differences in the tested methods but due to the Bland-Altman technique itself: if the standard deviations of each tested method are not strictly equal, the plotting of the data will tend to add an artificial negative correlation between average value and score difference (for a comprehensive explanation on the origin of this bias and on how to avoid it, refer to the article and simulation spreadsheet of Hopkins, 2004). Using a correlational approach prevents the emergence of this artificial bias, which is why we adopted this method for

Table 2 Parameters for computing stimulus level at trial n

Test	Parameters			
	Number of trials N	Level at first trial [ms] X_1	Initial step size [ms] c	Probability value Φ
MNREAD	20	200	200	0.5
Generated sentences (oral)	35	500	500	0.5
Generated sentences (true or false)	45	250	500	0.75

Note. The stimulus level at the n^{th} trial is given by: $X_n = X_{n-1} - ((c / n) * (Z_{n-1} - \Phi))$

Z_n, the answer of the subject at the n^{th} trial (1 = correct answer, 0 = wrong answer)

the two comparisons. The approach is composed of several steps: first we performed regressions to verify the relationships of the two pairs of scores: the correlation coefficients *r* were computed with the associated Student's tests to verify the significance of these coefficients. Once the relationships were found, we compared the parameters of the regression equations with the identity function to verify the absence of bias between the measurements. Finally, we analyzed the residuals by verifying their normality with Lilliefors tests and visually controlled the absence of heteroscedasticity on the regression graphs: if the variance of the residuals does not show an increase at the extremities of the data plot, this means that the error is globally the same for each subject.

Results

Two volunteers had to be removed from the analysis as their results did not allow the staircase to reach a stable level in the MNREAD test. Two others were removed since they admitted to having used a strategy to improve their reading speed during the "True or False" trial. This strategy is discussed in the last section. Ultimately, 21 women and 20 men participated in the validation experiment. Each staircase produced at least six reversals and thus allowed computation of a minimum display duration threshold.

The measured minimum display durations for each test and for each subject are given in Fig. 3. The mean of these durations for the MNREAD test was 115 ms (*SD*: 32.8). The equivalent in reading speed is 521.7 wpm. This reading speed is quite high compared to reading speed for traditional texts. We can explain this faster reading by the fact that the texts are extremely short (three lines of 20 characters) and easy to understand. The average minimum display durations for the generated sentences tests were shorter: 60 ms for the test with oral response and 58 ms for the test with "True or False" return (*SD*s were 25.2 and 26.8 respectively).

There is a large difference between the MNREAD and our tests scores. This can be due to two reasons. In our tests, the sentences are displayed on only one line, compared to three for the MNREAD. The two saccades to the beginning of the next line in the MNREAD are time consuming. Moreover, the generated sentences always follow the same pattern. The reader can rapidly get used to this presentation, and predict where the important words will be situated. This is in contrast to the syntactical pattern changes from one sentence to another in the MNREAD.

To validate our test, we performed two agreement assessments: (1) comparison between the oral recitation scores on the MNREAD and the generated sentence test, and (2) comparison between the two modalities of the generated sentence scores (oral and True or False). The correlation graphs corresponding to each assessment are plotted in Fig. 4. For assessment (1), the correlation coefficient is quite acceptable: r = .836, $t(39) = 9.504$, $p < .001$, 95 % CI [0.658, 1.014], *SEE* = 14.03. We could make the same observation for assessment (2): $r = .839$, $t(39) = 9.645$, $p < .001$, 95 % CI [0.663, 1.015], *SEE* = 14.77. As the correlation coefficients are important for the two pairs of data (each $r > .80$), we can state that the measures are strongly related to each other.

Regression (1) leads to the following equation: $score_{GenSentences\_Oral} = A_0 + A_1 * score_{MNREAD}$ with A_0 = -14.04, 95 % CI = [-30.37, 2.30] and A_1 = 0.64, 95 % CI = [0.51, 0.78]. The 95 % CI for the slope of the regression line implies that this gradient is significantly different from 1. The regression function is thus different from the identity function: the MNREAD scores are different from the generated sentences scores. This confirms the previously observed difference, indicating that there is a disparity between the scoring of the MNREAD and the generated sentences test.

Regression (2) gives: $score_{GenSentences\_TorF} = A_0 + A_1 * score_{GenSentences\_Oral}$, with A_0 = 4.53, 95 % IC = [-7.60, 16.68] and A_1 = 0.89, 95 % CI = [0.71, 1.08]. Given the confidence intervals of A_1 and A_0, the slope of the regression line is not statistically different from 1 and its intercept is not statistically different from 0. Thus the hypothesis that the regression equation equals the identity function cannot be rejected (i.e., $score_{GenSentences\_TorF} = score_{GenSentences\_Oral}$ cannot be rejected). This argues in favor of a similar measurement between the two testing modalities.

The hypotheses of normality were not rejected for the residuals of both regressions as revealed by the Lilliefors tests: in (1) $d = .092$, $p > .20$ and in (2) $d = .105$, $p > .20$. Visual observation of the two plots did not reveal any heteroscedasticity problem.

Discussion/conclusion

For the comparisons between MNREAD and generated sentences, we obtained strong correlation coefficients, normality and homoscedasticity of the residuals. This implies that the generated sentences test can be considered as a good predictor of the score obtained with a classical test. Nevertheless, there is a shift between the scores obtained with the two tests. The readers were able to perform more quickly on the generated sentences than on the MNREAD. As previously mentioned, this can be explained by a difference in the number of lines displayed (three for the MNREAD versus one for the generated sentences), by the repetitive pattern of the generated sentences, and by the use of frequent words only. The reading speed is thus greater in our test, but still linearly related to a classical measurement. This implies that the maximum reading speed obtained with the new test only requires taking into account the linear difference to be equivalent to an already validated reading speed. The fastest readers we

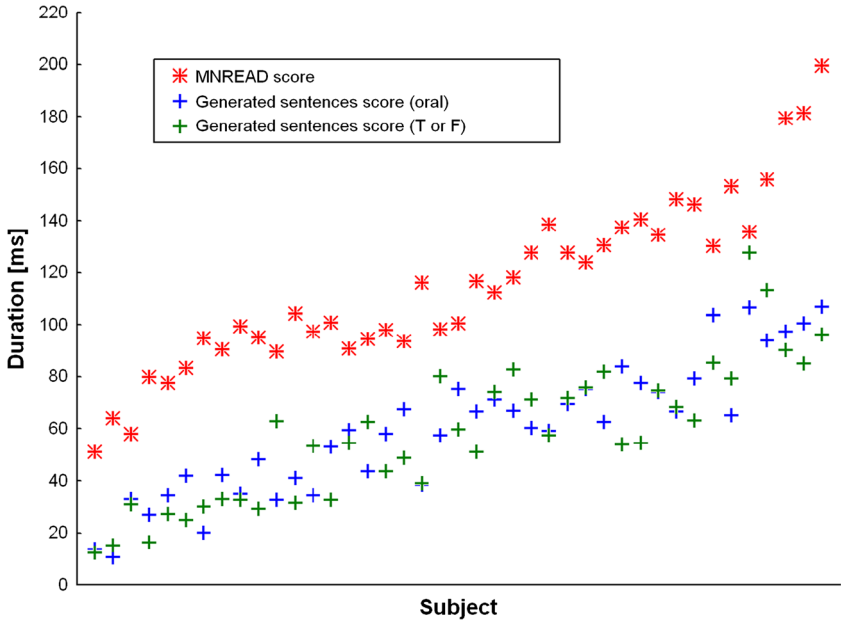

Fig. 3 Estimated minimum display duration per word permitting effective reading. Participants are classified with respect to their mean minimum display duration at the three tests

observed obtained extremely fast reading speed scores: the sentences could still be read, even if displayed for less than 100 ms, which should not allow them to accomplish eye movements. They were able to read the sentence with only one fixation, in the manner of the Rapid Serial Visual Presentation (RSVP) paradigm (Forster, 1970). These readers should have a perceptual span large enough to cover almost the whole sentence, allowing them to integrate the sentence with no eye movements. Arguing in favor of this explanation, previous research had already shown that fast readers have a larger perceptual span (Rayner, Slattery, & Bélanger, 2010). Despite this unusual behavior, where we could have expected a break in the linear relationship, the fast readers' scores on the MNREAD remained associated to the ones obtained with generated sentences in the same manner as for "normal" readers. This shows that the test measurements reflect capacities which are linked to those used during normal reading (i.e., including eye movements).

Furthermore, one may question the equivalence between the values obtained through our test with real-life reading speed, given the fast reading speed scores. Real-life reading speed is not a constant value for each reader; its variability is due to changes in the reader's strategy (Carver, 1990; Lemaire, Guérin-Dugué, Baccino, Chanceaux, & Pasqualotti, 2011) and to visual and semantic aspects of the reading material. These are the aspects we have tried to control in the generated sentences to obtain highly homogeneous material. The reading strategy used by the readers was also controlled by asking them to read as fast as they could and to read the sentences so that they could recite or comprehend them. Thereby the new test is able to measure the maximum reading speed, which is not the usual speed one would observe in natural conditions. This specific measurement allows avoiding within-reader variability in order to obtain a stable value, and to perform repeated measurements. The effect of a change in the visual presentation of the sentences and the evolution of the reader's performance over time can thus be studied while removing other factors.

The scores obtained by the two modalities of response for the generated sentences are also well correlated. Moreover, their relationship is close to equality. This is an encouraging result in favor of the validity of the True or False modality. One of the aims of our work was to provide sentences for which the truth value could be unambiguously resolved by the reader. If ambiguous sentences had been displayed to the readers, they would have given wrong answers, even if the display duration had been long enough for effective reading. Thus the staircase would have attained a longer minimum display duration threshold, and we would have observed a tendency for slower reading speed scores in the True or False modality. As this is not the case, we can state that this aim is accomplished. The two modalities can therefore be used in the same manner to measure a maximum reading speed. The True or False modality is easier to set up, because it allows the reader to take the test without needing an observer to verify an oral response. Alternatively, if a subject does not feel comfortable with the True or False modality, the observer can suggest taking the oral test. The new method for generating sentences can offer substantial material for reading speed assessment. We have proposed a method that allows a straightforward production of the sentences: the only effort required of the generator developer is to select suitable common nouns and structure them in an ontology. The successive steps to produce sentences are automated. One of the major improvements of our method is that the selection of manual nouns allows the creator to control lexical and psycholinguistic variables. Controlling these parameters is crucial to avoid

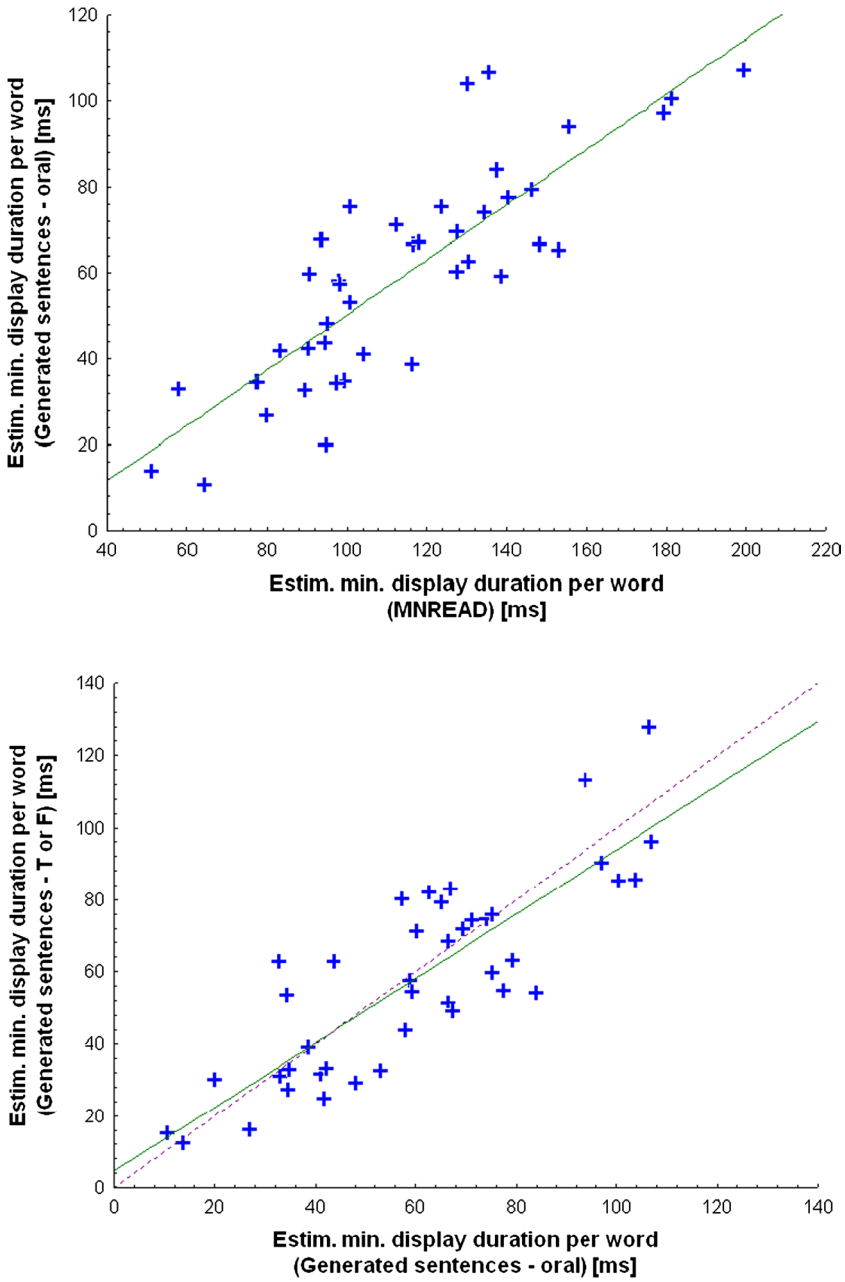

Fig. 4 Regression graphs for the comparisons between the two tests and the two modalities. Green lines represent regression lines. In the second graph, the purple dotted line represents the identity function

variability due to the text itself. This control, coupled with a near-natural language formalism, facilitates the reader's accurate answering.

In the validation phase, each sentence was displayed in full, not word-by-word as in the RSVP paradigm, where each word of the sentence can be displayed for an extremely short period of time. Traditional computers' display speed is limited by the refresh rate of the screens, and thus cannot display every word when its frequency of appearance is close to the frequency of the screen. Our sentences could very well be displayed with an RSVP, but on a high temporal resolution display system.

We had to exclude a few subjects as their results in the staircase never reached stability. This indicates a change in their level of performance. This could be explained by a modulation in their attention level, but it could also be due to a change in strategies used for reading. Some subjects explained having used a strategy which was not to fixate the cross preceding the sentence, but to fixate a few centimeters further. This strategy allows the visual span to cover more space on the text. To prevent the readers from using this strategy, one could displace the fixation cross further on the line or record eye movement and start the display of the sentence only if the reader is effectively fixating this cross. Nevertheless, the majority of the subjects' data show encouraging results.

We proposed this method to assess maximum reading speed, assuming a psychophysical approach. Such an

approach allows for multiple repetitions of the measurement. The method is dedicated to evaluate the differences in reading performance related to visual factors. Nevertheless, reading is not limited to a visual processing. The purpose of reading is to extract knowledge included in the text, which implies multiple cognitive processes which occur based on the visual information obtained. By suggesting easily understandable sentences which follow the same pattern, and are composed of highly frequent words denoting concrete concepts, we have minimized the effect of these processes on the resulting measure. Reading evaluation research can thus not be thorough without assessing these processes: for example, the new test is not sensitive to vocabulary acquisition as the words are only ones that are commonly used. The role of comprehension is also extremely limited in our test since the comprehension task depended on common prior knowledge. We hope that the reading research community will continue developing reliable tools to study reading at all levels, and that our approach can inspire future tools designed for this purpose.

Acknowledgments This research was supported by Essilor International. The authors would like to thank Catherine Agathos, Delphine Tranvouez, and William Seiple for the proofreading and Jean Lieber for the advice about the use of ontology.

References

Ahn, S. J., Legge, G. E., & Luebker, A. (1995). Printed cards for measuring low-vision reading speed. *Vision Research, 35*(13), 1939–1944. doi:10.1016/0042-6989(94)00294-V

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307–310. doi:10.1016/S0140-6736(86)90837-8

Bonin, P., Méot, A., Aubert, L., Malardier, N., Niedenthal, P., & Capelle-Toczek, M. C. (2003). Normes de concrétude, de valeur d'imagerie, de fréquence subjective et de valence émotionnelle pour 866 mots. *L'Année Psychologique,* 655–694. doi:10.3406/psy.2003.29658

Carver, R. P. (1990). *Reading rate: A review of research and theory.* Academic Press.

Crossland, M., Legge, G., & Dakin, S. (2008). The development of an automated sentence generator for the assessment of reading speed. *Behavioral and Brain Functions, 4*(1), 14. doi:10.1186/1744-9081-4-14

Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics, 35*(10), 1297–1326. doi:10.1080/00140139208967394

Dyson, M. C. (2004). How physical text layout affects reading from screen. *Behaviour & Information Technology, 23*(6), 377–393. doi:10.1080/01449290410001715714

Egidi, G., & Gerrig, R. J. (2009). How valence affects language processing: Negativity bias and mood congruence in narrative comprehension. *Memory & Cognition, 37*(5), 547–555. doi:10.3758/MC.37.5.547

Forster, K. (1970). Visual perception of rapidly presented word sequences of varying complexity. *Perception & Psychophysics, 8*(4), 215–221. doi:10.3758/BF03210208

Hopkins, W. G. (2004). Bias in Bland-Altman but not regression validity analyses. *Sportscience, 8*(4).

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. *Science, 274*(5284), 114–116. doi:10.1126/science.274.5284.114

Kendeou, P., & van den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Memory & Cognition, 35*(7), 1567–1577. doi:10.3758/BF03193491

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16*(1-2), 262–284. doi:10.1080/09541440340000213

Landerl, K., & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology, 100*(1), 150. doi:10.1037/0022-0663.100.1.150

Latham, K., & Whitaker, D. (1996). A comparison of word recognition and reading performance in foveal and peripheral vision. *Vision Research, 36*(17), 2665–2674. doi:10.1016/0042-6989(96)00022-3

Legge, G. E., Ross, J. A., Luebker, A., & LaMay, J. M. (1989). Psychophysics of reading. VIII. The Minnesota Low-Vision Reading Test. *OptomVis Sci, 66*(12), 843–853.

Lemaire, B., Guérin-Dugué, A., Baccino, T., Chanceaux, M., & Pasqualotti, L. (2011). A cognitive computational model of eye movements investigating visual strategies on textual material. *cogSci 2011 Proceedings,* 1146–1151.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE. *L'Année Psychologique,* 447–462. doi:10.3406/psy.2001.1341

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision Research, 46*(28), 4646–4674. doi:10.1016/j.visres.2006.04.023

Rayner, K., Slattery, T. J., & Bélanger, N. N. (2010). Eye movements, the perceptual span, and reading speed. *Psychonomic Bulletin & Review, 17*(6), 834–839. doi:10.3758/PBR.17.6.834

Robbins, H., & Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics, 22*(3). doi:10.1214/aoms/1177729586

Sadoski, M. (2001). Resolving the effects of concreteness on interest, comprehension, and learning important ideas from text. *Educational Psychology Review, 13*(3), 263–281. doi:10.1023/A:1016675822931

Seiple, W., Szlyk, J. P., McMahon, T., Pulido, J., & Fishman, G. A. (2005). Eye-movement training for reading in patients with age-related macular degeneration. *Investigative Ophthalmology & Visual Science, 46*(8), 2886–2896. doi:10.1167/iovs.04-1296

Senécal, M. J., Gresset, J., & Overbury, O. (2006). *Minnesota Low-Vision Reading Test, version française: Échelle d'acuité visuelle MNREAD*. Longueuil: Institut Nazareth & Louis-Braille.

Sereno, S. C., & Rayner, K. (2003). Measuring word recognition in reading: Eye movements and event-related potentials. *Trends in Cognitive Sciences, 7*(11), 489–493. doi:10.1016/j.tics.2003.09.010

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Rand Corporation.

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition, 116*(1), 71–86. doi:10.1016/j.cognition.2010.04.002

Trauzettel-Klosinski, S., Dietz, K., & the IReST Study Group. (2012). Standardized assessment of reading performance: The New International Reading Speed Texts IReST. *Investigative Ophthalmology & Visual Science, 53*(9), 5452–5461. doi:10.1167/iovs.11-8284

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research, 35*(17), 2503–2522. doi:10.1016/0042-6989(95)00016-X