Make It Big! The Effect of Font Size and Line Spacing on Online Readability

Luz Rello
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
luzrello@cs.cmu.edu

Martin Pielot
Telefonica Research
Placa Ernest Lluch i Marti, 5
Barcelona, 08019, Spain
martin.pielot@telefonica.com

Mari-Carmen Marcos
DigiDoc & Web Research Group
Universitat Pompeu Fabra
Barcelona, Spain
mcarmen.marcos@upf.edu

ABSTRACT
We report from an eye-tracking experiment with 104 participants who performed reading tasks on the most popular text-heavy website of the Web: Wikipedia. Using a hybrid-measures design, we compared objective and subjective readability and comprehension of the articles for font sizes ranging from 10 to 26 points, and line spacings ranging from 0.8 to 1.8 (font: *Arial*). Our findings provide evidence that readability, measured via mean fixation duration, increased significantly with font size. Further, comprehension questions had significantly more correct responses for font sizes 18 and 26. For line spacing, we found marginal effects, suggesting that the two tested extremes (0.8 and 1.8) impair readability. These findings provide evidence that text-heavy websites should use fonts of size 18 or larger and use default line spacing when the goal is to make a web page easy to read and comprehend. Our results significantly differ from previous recommendations, presumably, because this is the first work to cover font sizes beyond 14 points.

Author Keywords
Readability; comprehension; font size; line spacing; online reading; text presentation; eye-tracking; Wikipedia.

ACM Classification Keywords
H.5.0 Information Interfaces and Presentation: General; H.5.2 Information Interfaces and Presentation: User Interfaces—*Screen design, style guides.*

INTRODUCTION
While it may seem a little old-fashioned, reading is still one of the primary ways to interact with computing devices. And as more and more content and services move online, reading increasingly takes places on screens and in web browsers. Improving *readability* of text in the Web is one of the most simple and effective ways to improve usability and ease access to information – also for people with special needs, such as elderly people [14], or people with print disabilities [26], such as people with low vision [18] or dyslexia [23, 29].

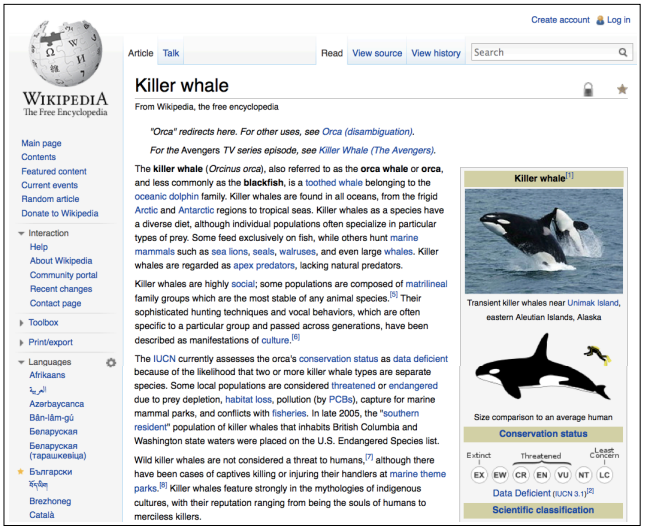

Figure 1. Example of a Wikipedia article used in the experiment.

In this work, *readability* refers to the ease with which a reader can read and understand a written text. It only refers to the properties of the presentation of the text, not the content of the text itself. One of the crucial factors for readability in this context is font size [21, 24] together with line spacing [25].

Designers who try to find the optimal font size and line spacing for their web pages face a myriad of conflicting recommendations. Jakob Nielsen, one of the experts in web usability, stresses to allow users to control the size

CHI'16, May 07-12, 2016, San Jose, CA, USA.
 ISBN/978-1-4503-3362-7/16/05...$15.00.
http://dx.doi.org/10.1145/2858036.2858204

of the font [21].[1] However, as many users will not adjust the text display settings, he suggests to use 10 points as minimum font size. According to a survey on typographic design patterns and best practices [20], in 2009 the most used font size was 13 points and the most used line spacing was 148% (line spacing of 1.23) of the font size. In a replication of the study in 2013, the font sizes 14 and 16 points were most popular [13]. Yet, these numbers represent best-practices, which are not necessarily backed up by scientific studies.

Research on font size has a long history in the HCI Community. Previous research [1, 2, 3, 4, 5, 8, 14] has extensively explored font sizes of 10, 12, and 14 points, and typically found that the biggest tested font size leads to the best results. However, this leaves a number of open questions. First, it has largely ignored line spacing as parameter [9]. Second, the studies mostly focus on readability, while comprehension, *i.e.*, whether the reader actually understood the content of the text, is mostly not measured. Third, we did not find any general-population study exploring font sizes above 16 points, even though previous findings indicate bigger fonts lead to better results.

In this paper, we report from a user study that addresses these open questions. With 104 participants, we conducted an experiment to determine the effect of font size and line spacing on readability and comprehension of texts on websites. Using a 17-inch monitor with integrated eye-tracker, the participants had to read web text in a *Firefox* browser, displayed in its default sans-serif font: *Arial*. We compared 6 font sizes (10, 12, 14, 18, 22, and 26 points) and 4 line spacings (0.8, 1.0, 1.4, and 1.8). The participants had to read articles from a popular text-heavy website: *Wikipedia* (see Figure 1).

Our main findings are:

- Font size had significant and large effects on readability and comprehension. Both aspects improved significantly with increasing font sizes until 18/22 points – far beyond the typical recommendations of 10/12/14 points.
- Line spacing had a small but significant effect on comprehension, suggesting that too small or too large spacings may impair comprehension.
- On the basis of our findings, we recommend to use 18 points font size and default line spacing if the goal is to optimize readability and comprehension of web text content.

[1] http://www.nngroup.com/articles/let-users-control-font-size/

RELATED WORK

Font Size

In his article on the *Top 10 Mistakes in Web Design*,[2] Jakob Nielsen argues that using too small font sizes is one of the most frequent mistakes made in today's website design. Preferably, users should be allowed to adjust the font size to their individual needs. Yet, Nielsen also points out that users are typically too lazy to change fonts when viewing websites. Consequently, to ensure good readability, it is essential for websites to provide appropriate defaults. Nielsen recommends to use font sizes of at least 10 points or 12 points for elderly readers. However, previous research come to different conclusions about the ideal font size:

Bernard *et al.* [3] performed a study with 60 participants measuring reading time, preference, and errors while reading the text out loud using eight different font types and 10, 12 and 14 points. Fonts of 10 points were read significantly more slowly than fonts of 12 points. In a subsequent experiment, Bernard *et al.* [4] compared the readability of two fonts –*Arial* and *Times*– and two font sizes –10 and 12 points– in an experiment with 35 participants. The experiment used the same dependent measures. 10-point *Arial* typeface again was read slower than the other conditions and the 12-point *Arial* typeface was preferred to the other typefaces.

In order to understand the impact of age on reading, Bernard *et al.* [2] studied the effects of font type and size on the legibility and reading time of online text by older adults. They compared legibility, reading time, and the participants' preferences of texts displayed with sans-serif and serif fonts, and font sizes 12 and 14 points. The 12-point serif fonts were read out loud significantly slower than 14 serif and sans-serif fonts, and participants preferred larger font sizes.

In the context of handheld computers, Darroch *et al.* [14] investigated the effect of font sizes ranging from two to 16 points, measuring the reading speed (silent reading), reading accuracy, and subjective views among two groups, 12 old and 12 young readers. They did not find any significant differences neither between the age groups, nor for the font sizes ranging from 6 to 16 points, which may be due to the rather small sample size.

Banerjee *et al.* [1] performed a study with 40 participants, who had to read texts aloud using the font sizes 10, 12, and 14 points. Sitting at a distance of 60 to 70cm from the monitor, the 14-point fonts lead to a significant faster reading and was preferred over smaller font sizes.

Bhatia *et al.* [8] studied the effect of, amongst other factors, font size on readability. A group of 180 undergraduate students had to take part in a text-reading experiment and indicate their preferences in a survey. The font sizes that Bhatia *et al.* tested were 10, 12, and

[2] http://www.nngroup.com/articles/top-10-mistakes-web-design/, 2011, last visited Sep 07, 2013.

14 points. Unfortunately, the survey responses did not reveal any significant effects.

The related work described until now approximated text readability via preferences, reading time, and errors made while reading the text aloud. However, these measures have drawbacks. Subjective readability may not match objective readability. Fast reading times do not necessarily indicate good readability. For example, participants may simply skim texts which are difficult to read. Reading aloud may introduce unsystematic variance through the extra reading activity, and is not ecologically valid for web reading.

As a remedy, Beymer and Russel [6] explored the use of objective measures of readability: they developed *WebGazeAnalyzer*, a system to monitor reading performance with an eye tracker. This system allows, amongst other things, to record the eyes' fixations durations, which are an objective indicator of text readability [19, 27, 31]. The longer the eyes fixate text parts at a time, the higher the likelihood that the reader is encountering difficulties in reading the text. Using this system in a between-subjects design with 82 participants, Beymer *et al.* [5] studied the effect of the font sizes 10, 12, and 14 points on readability and comprehension scores. When using 10 points font size, fixation durations resulted significantly longer as compared to 14 points. They also found significant differences taking into account the mother language of the participants, non-native English subjects had significantly longer fixations.

In a similar setup, Rello *et al.* [29] studied the effect of font size on the ability of people who were diagnosed with dyslexia to read texts. They found that larger font sizes than the usual 10, 12, 14 points led to shorter fixation durations.

The findings from previous work unanimously indicate that the font sizes of 10 to 12 points, as suggested by Nielsen and other sources, might be suboptimal, and that font sizes that are larger than the traditional 10, 12, 14 points led to significant improvements in readability and comprehension.

Dependencies of Font Size

Previous work indicates that font size is interdependent with font type [3, 5]. Most of the previous work applies to the two most common fonts used on screen and printed texts, *Arial* and *Times*, respectively [12]. One of the reasons is that font size can result into different letter sizes for different font types,[3] so parts of the observed effects might be due to the actual size of the letters. In consequence, research on the effect of font size needs to consider the font type, *e.g.*, by at least making clear for which font type the findings are valid, or consider letters of the same real size event if different in point size.

[3]For the interested reader, see Figure 1 in Boyarski *et al.* [10], who compared *Times* with *Georgia* and *Verdana*.

In addition, notice that if the column width is fixed, the number of character per column depends on the font size.

Line Spacing

Line spacing refers to the vertical distance between the baselines of two text lines. The concept is also know as *leading* from the days of hand-typesetting and *line-height* in CSS. The bigger the line spacing, the further two sentences are apart vertically.

We found no specific guidelines for line spacing of web texts. By default, browsers compute the line spacing relative to the font size. A spacing of 1.0 equates to 120% of the font size.[4] In best-practice recommendations, this spacing of 1.0 is often named as "generally the most readable and doesn't require that you do anything special".[5] However, no studies are cited.

According to a review by Bix [9], the vast amount of literature indicates that the optimal amount of spacing highly depends on other factors. Except for the general recommendation to avoid too little and too much spacing, no rules are given.

Paterson and Tinker [25] studied the effect of line spacing in printed text when performing a reading test (Chapman-Cook Speed) with 400 college students. They found that bigger line spacings (1.2 and 1.4 compared to 1.1) lead to faster readings. However the authors point out that such results may depend on other factors such as the font type and the column width, similarly to font size, as already mentioned.

Rello and Marcos [28] studied the effect of line spacing (0.8, 1.0, 1.4, and 1.8) for reading raw text on the screen using eye-tracking with 92 participants. They did not find a significant effect of line spacing on fixation duration.

What is Missing

In all presented previous studies, bigger fonts led to better results, either in terms of readability [5] or in terms of preference [2]. Thus, previous work indicates that bigger font sizes will result into more readable websites. However, in the context of desktop computers, the biggest font size studied was 14 points. Thus, previous work does not answer whether this trend continues or flattens out with increasing font sizes. About line spacing, no conclusive evidence has been reported.

METHODOLOGY

To study the effect of font size and line spacing on readability and comprehension of websites, we conducted an eye-tracking experiment comparing font sizes from 10 to 26 points and line spacings from 0.8 to 1.8 times the standard spacing. We had 104 participants that read six Wikipedia entries related to animals with different font

[4]http://stackoverflow.com/questions/2262543/css-line-height-guide

[5]http://webdesign.about.com/od/styleproperties/qt/css_line_spacing.htm

sizes and line spacings. We chose Wikipedia, since it is the most-frequently visited text-heavy website.[6] Other more frequently visited websites, such as Google, Facebook, or Yahoo, contain almost no text or are multimedia focused, hence not useful for this study. Readability and comprehension were analyzed via eye-tracking, comprehension tests, and subjective perceptions of the participants.

Design

In our experimental design, *font size* and *line spacing* served as independent variables with 6 and 4 levels, respectively:

– For *font size*, we used the levels 10, 12, 14, 18, 22, and 26 points. We chose 10, 12, and 14 points to compare the results with previous studies [1, 2, 3, 7, 8]. The larger font sizes were chosen to cover a wide range of sizes, as previous work had indicated that larger font sizes improves readability, without having shown the limits of this improvement.

– For *line spacing* we tested 0.8, 1.0, 1.4, and 1.8, where 1.0 represents the browser's default line spacing – Firefox, in this case– which equals 120% points of the font size.[7] We chose 1.4 since many style guidelines suggest to use slightly increased line spacings and 1.0 because is the default in word processors.

We used a hybrid-measures design. Each participant read six texts with the same *line spacing* but six different *font sizes*. Hence, for *font size*, we collect repeated measures, while for *line spacing*, we obtain between-group data. The order in which the font sizes were presented was counter-balanced to cancel out sequence effects.

For quantifying readability and comprehension, we used the following dependent measures:

Fixation Duration: We used mean fixation duration as objective approximation of readability. When reading a text, the eye does not move contiguously over the text, but alternates saccades and visual fixations, that is, jumps in short steps and rests on parts of the text. *Fixation duration* denotes how long the eye rests still on a single place of the text.

According to previous work [19, 27, 31], fixation duration is a valid and objective proxy for readability. The rationale put forward by Just and Carpenter [19] is that "*readers make longer pauses at points where processing loads are greater.*" Rayner and Duffy [27] write that "*There is now a fair amount of evidence to indicate that some of the variability [of fixation duration] is due to systematic differences in the ease of processing the words in the text.*" Hence, if the mean duration of fixations increases, the reader has encountered more difficulties, which means that the text is more difficult to read.

Comprehension Score: To measure text comprehension, we used literal and inferential questions. That is, we cover both types of text comprehension. Inferential items are questions that require a deep understanding of the text content, because the question cannot be answered straight from the text. Literal questions, in contrast, can be answered directly from the text. We used multiple-choice questions with four possible choices, one correct choice, two wrong choices, and "I don't know". To compute the text comprehension score, the correct choice counted 100% and the rest 0%.

For each text, we created two questions: one literal and one inferential. Questions were asked right after each text had been read to avoid order and memory effects, e.g. the last text scoring a higher comprehension score.

Literal questions could be directly answered from the text such as "the giant turtle lives? (a) in the Seychelles, (b) Caribbean, etc." Inferential questions required a deeper comprehension of the text (see Figure 3 in the paper). Using literal and inferential questions resemble questions about the *main idea* and *main facts* used by Dyson and Hazelgrove [16].

Subjective Perception Rating: In addition, we asked the participants to provide their subjective perceptions. For each of the six texts, the participants rated on two five-point Likert scales, how easy it was for them to read and to understand the text. This defines the *Subjective Readability Rating* and *Subjective Comprehension Rating*, respectively. Sauro and Dumas [30] showed that, given a sufficiently large sample, single-item scales are a very easy-to-use, yet effective measurement tool. Hence, we used two single-item scales, of which on is shown in Figure 2.

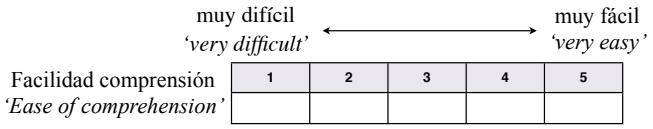

Figure 2. Comprehension perception scale rating.

Participants

Following our IRB requirements, we make a public announce and sent it to the schools and universities of a city district. Potential participants contacted us and came to our lab at the Universitat Pompeu Fabra (UPF). Minors were accompanied by a parent. Most of the participants were students. Participation was voluntary.

104 volunteers (61 female, 43 male) took part in the study. Their ages ranged from 14 to 54 ($\bar{x} = 30.24$, $s = 9.13$) and they all had normal or corrected-to-normal vision. Most of the participants had higher education. Except from 5 participants, 7 participants were attending school or high school and 92 participants were studying

[6] Wikipedia is the seventh most popular website worldwide, according to Alexa ranking: `http://www.alexa.com/topsites` (consulted Dec 18, 2015).

[7] That is, 0.8 equals to 96% , 1.0 equals to 120%, 1.4 equals to 168%, and 1.8 equals to 216% of the font size.

or had already finished university degrees. All participants were frequent readers: per day, 47.83% read less than four hours per day, 39,13% read between four and eight hours per day, and 13.04% participants read more than eight hours.

Materials

To isolate the effects of the text presentation, the texts themselves need to be comparable in complexity. In this section, we describe how we designed the texts that were used as study material.

Wikipedia Entries

Since Wikipedia entries are heterogeneous, it is challenging to find many similar entries. We decided against modifying articles to increase ecological validity. Thus, we went through the articles of Wikipedia and chose 24 articles which share the following comparable characteristics:

(a) All texts used in the experiment cover the same *genre* and the same *topic*, namely animals. We chose animals because they are a topic of general interest, not technical or academic.

(b) They all have a similar number of words in the first and the second paragraphs, ranging from 40 to 60 words for each of the paragraphs.

(c) They have a similar discourse structure: title, the first paragraph presents the animal and the place where it lives, and the second and third paragraphs provide more details.

(d) The layout was always the same: the paragraphs were located in roughly the same position of the screen. Each article contained one image on the top-right of the content pane (see Figure 1).

(e) All texts had low frequencies (ranging from two to five) of numerical expressions, acronyms, and foreign words, because these type of words are processed differently than regular words [15, 31].

(f) All the entries used the *sans-serif* font *Arial*, which Wikipedia uses as default on Firefox and other browsers on MS Windows.

For each of the selected Wikipedia articles, we obtained the HTML source code. To alter the presentation, we used a browser plug-in (StyleBot) to modify the style sheet (CSS) to change font size and line spacing.

Comprehension Questionnaires

The comprehension questions were administered in form of questionnaires. There was one questionnaire for each of the Wikipedia articles containing six multiple-choice questions. An example of each type of item is given in Figure 3.

Equipment

The eye-tracker that we used was the Tobii 1750 [33], which has a 17-inch TFT monitor with a resolution of 1024x768 pixels. The time measurements of the eye-tracker have a precision of 0.02 seconds. The eye-tracker was calibrated for each participant and the light focus was always in the same position. The distance between the participant and the eye-tracker was constant (approximately 60 cm. or 24 in.) and controlled by using a fixed chair. Figures 4 and 5 show the setup as used during the study.

Segun lo que acabas de leer en la Wikipedia *'According to what you just read on Wikipedia:'*

- El gorila tiene un ADN muy similar al de los humanos. *'The gorilla's DNA is similar to the humans' one.*
- El gorila vive en los bosques del sur de África. *'The gorilla lives in the forests of southern Africa'.*
- El gorila es un primate carnívoro. *'The gorilla is a carnivorous primate'.*
- No lo sé, creo que lo no ponía o al menos yo no lo recuerdo. *'I do not know, I think it was not in the text, or at least I do not remember it'.*

Figure 3. Comprehension item.

Figure 4. Setup: the eye-tracker used during the experiment showing a Wikipedia article used.

Procedure

The study took place at the end of 2012. The sessions were conducted at the Universitat Pompeu Fabra, and lasted around 20 minutes for each participant. Each session took part in a quiet room, where only the interviewer (first author) was present, which ensured that the participants could concentrate. Each participant performed the following four steps.

First, we handed out the questionnaire that was designed to collect demographic information: age, gender, native languages, education, and reading hours per day. Second, we asked the participants to read six Wikipedia articles in silence. For each article, the participants read freely covering at least the first 3 paragraphs. During this time, the reading was recorded by the eye-tracker. After finishing each article, the participants completed the corresponding comprehension questionnaire. Finally,

Figure 5. Heat map of the eye fixations.

we presented all the articles again to the participants, and asked them to provide their ratings regarding the texts readability and comprehension.

RESULTS

In this section, we present the analysis of the data from the eye-tracker (fixations), the comprehension tests, and the subjective perception ratings.

Mean Fixation Duration

Font Size: In some cases, the eye-tracking did not function correctly throughout the whole study. Hence, we only analyze the 93 participants that successfully contributed to all six font-size conditions. Figure 6 and Table 1 show the mean durations of eye fixations for each of the *font size* conditions.

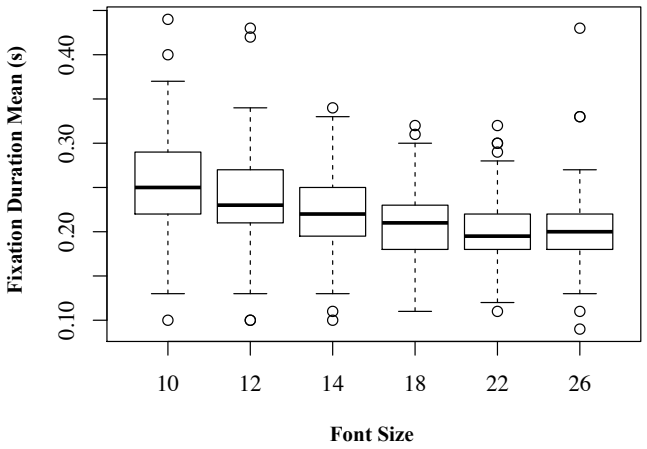

Figure 6. Mean fixation duration by font size (lower fixation durations indicate better readability).

A Two-Way ANOVA revealed a significant main effect of *font size* on *mean fixation duration* ($F(5, 445) = 66.825, p < .001$). The Eta-Square effect size value ($\eta^2 = 0.159$) suggests a high practical significance. Holm-corrected pairwise, repeated-measures t-tests revealed the following significant differences:

Font size	Mean	SD
10	.255	.059
12	.241	.054
14	.224	.048
18	.208	.040
22	.199	.037
26	.204	.045

Table 1. Mean and standard deviation of the average fixation durations per font size. Smaller mean values indicate better readability.

- For 10 points font size, participants had significantly longer fixation durations than for all larger font sizes ($p < .01$ for 12 pts, $p < .001$ all other comparisons).
- For 12 points font size, participants had significantly longer fixation durations that for all larger font sizes ($p < .001$, each).
- For 14 points font size, participants had significantly longer fixation durations than for all larger font sizes ($p < .001$, each).
- For 18 points font size, participants had significantly longer fixation durations than for 22 points ($p = .003$).
- Otherwise, there were no significant differences between the larger font sizes.

The data indicates an absolutely continuous decrease of the mean fixation duration until font size 18. Mean fixation duration was lowest for 22 points.

Spacing: Figure 7 and Table 2 show the mean fixation duration for each of the *line spacing* conditions. The number of participants per condition were 24, 29, 26, and 25 for the conditions 0.8, 1.0, 1.4, and 1.8 respectively. We did not find a significant effect of *line spacing* on *fixation duration* ($F(3, 89) = 0.064, p = .978$).

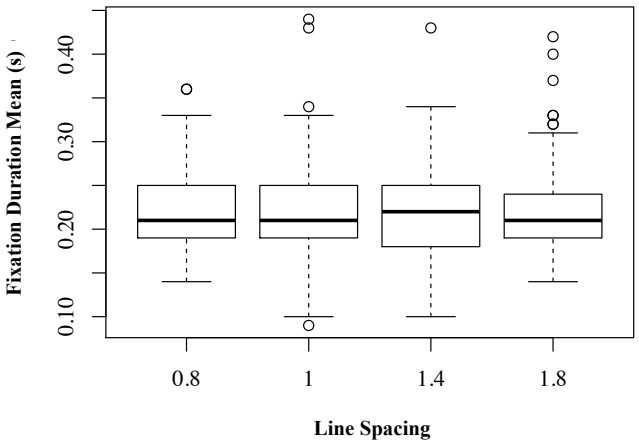

Figure 7. Mean fixation duration by line spacing (lower fixation durations indicate better readability).

Spacing	Mean	SD
0.8	.224	.046
1.0	.220	.057
1.4	.220	.054
1.8	.223	.050

Table 2. Mean and standard deviation of the average fixation durations per line spacing. Smaller mean values indicate better readability.

Interaction Fontsize x Spacing: There was a significant interaction between *font size* and *spacing* ($F(15, 445) = 4.098, p < .001$). The Eta-Square effect size value ($\eta^2 = .034$) suggests low practical significance. In the interaction plot (Fig. 8) with Spacing as group, we see all 4 lines dropping in parallel for increasing font size. Only for 26 points, the lines visibly diverge. We ran tests on subsets of the data. The interaction effect disappears when only keeping font sizes 10-18. A marginal effect remains when keeping font sizes 10-22. This indicates that the two largest tested fonts (22 and 26 points) were affected by line spacing.

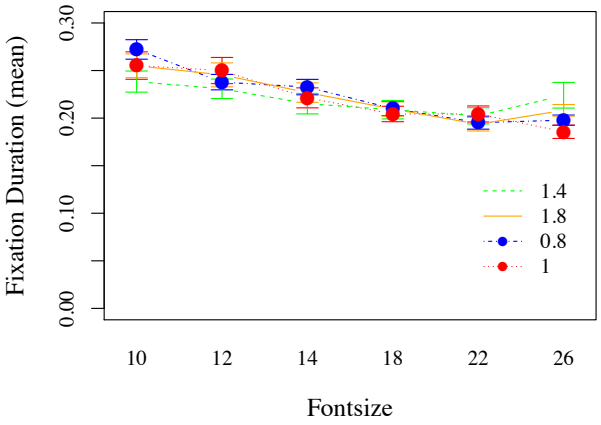

Figure 8. Mean fixation duration interaction plot.

Number of fixations: While mean fixation duration is an established proxy for readability, it might be assumed that shorter fixations led to more fixations, which intuitively would appear as counter-evidence to any readability improvement. Hence, we analyzed the number of fixations. A Two-Way ANOVA revealed the existence of a significant main effect of *font size* on the *number of fixations* ($F(5, 445) = 5.249, p < .001$). However, the Eta-Square effect size value ($\eta^2 = .025$) suggests low practical significance. The post-hoc comparisons, using holm-corrected, pairwise t-tests, revealed only a single significant difference: the number of fixations were significantly lower for 10 points ($M = 115.7, SD = 61.4$) than for 12 points ($M = 149.4, SD = 87.3$) ($p < .001$). Other comparisons are not significant. Neither plots nor post-hoc tests do not reveal a clear trend. Therefore, the reduced mean fixation duration observed for increasing font size cannot be explained by a an increase in the number of fixations.

Comprehension Score

Figure 9 shows the comprehension score distribution for each of the *font size* conditions. A Levene Test showed that the variances in the comprehension scores were not sufficiently equal to use parametric statistics ($F(5, 532) = 5.696, p < .001$). A Friedmann Test revealed a significant effect of *font size* on the comprehension score ($\chi^2(5) = 27.29, p < .001$). Holm-corrected pairwise Wilcoxon Signed-Rank tests revealed the following differences:

- For 10 and 12 points, participants had significantly lower comprehension scores than for 18 points ($p < .01$, both).
- For 12 points, participants had significantly lower comprehension scores than for 26 points ($p < .05$).

Hence, our data indicates that, **comprehension was negatively affected by small font sizes.**

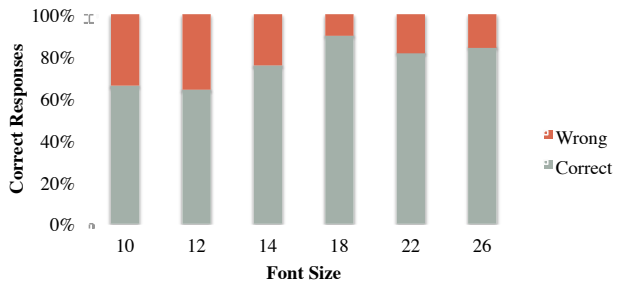

Figure 9. Mean comprehension scores by font size.

Figure 10 shows the comprehension score distribution for each of the *line spacing* conditions. A Levene Test showed that the variances in the comprehension scores were not sufficiently equal to use parametric statistics ($F(3, 534) = 6.729, p < .001$). A Kruskal-Wallis Test revealed a significant effect of *line spacing* on the comprehension score ($\chi^2(3) = 19.56, p < .001$). Holm-corrected pairwise Mann-Whitney U tests revealed the following effects:

- For 0.8 line spacing, participants had significantly lower comprehension scores than for 1.0, 1.4. and 1.8 ($p < .05$, $p < .05$, $p < .001$, respectively).

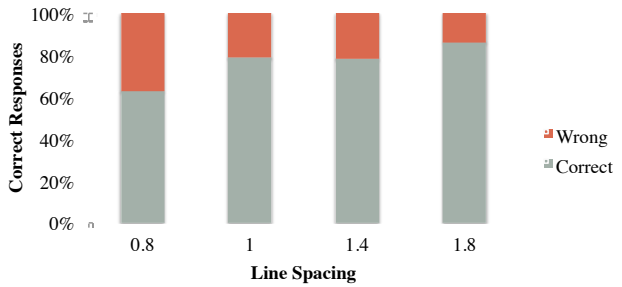

Figure 10. Mean comprehension scores by line spacing.

Hence, in our data set, **the smallest line spacing led to lower scores in the text comprehension tests.**

Subjective Readability Ratings

Figure 11 shows the distribution of the subjective readability ratings by *font size*. A Levene Test showed that the variances in the perception ratings were not sufficiently equal to use parametric statistics ($F(5, 363) = 6.705, p < .001$). A Friedman Test revealed a significant effect of *font size* on subjective readability ($\chi^2(5) = 135.85, p < .001$). Holm-corrected pairwise Wilcoxon Signed-Rank tests revealed the following differences:

- For 10 points, readability ratings were significantly lower than for all other sizes ($p < .001$, each).
- For 12 points, readability ratings were significantly lower than for all larger sizes ($p < .001$, each).
- For 14 points, readability ratings were significantly lower than for 18 points ($p < .01$).
- For 26 points, readability ratings were significantly lower than for 18 points ($p < .05$).

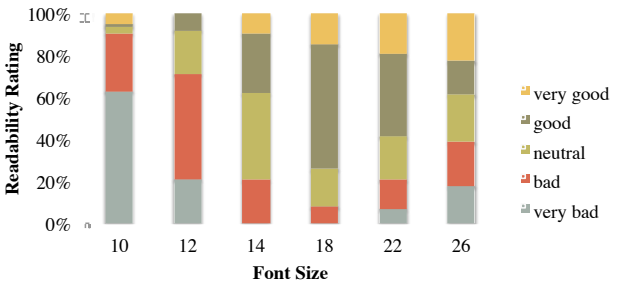

Figure 11. Subjective readability ratings increased with increasing font sizes until 18 points.

Thus, for the participants of this study, **the subjective perception of readability continuously improved and was highest for 18 points font size**.

Figure 12 shows the distribution of the readability ratings by *line spacing*. A Levene Test showed that the variances in the perception ratings were sufficiently equal to use parametric statistics ($F(3, 365) = 1.568, p = .197$). A independent-measures ANOVA did not reveal any significant effect on the subjective readability ($F(3, 365) = 2.074, p = .103$).

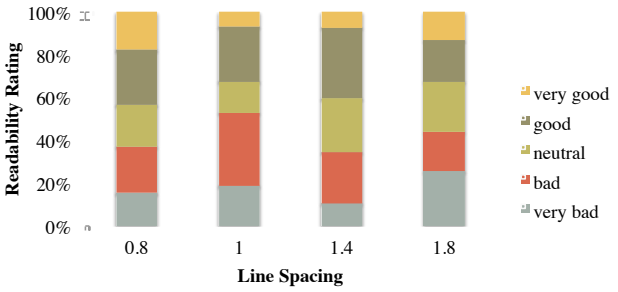

Figure 12. No significant effect of line spacing on readability was found.

Comprehension Perception Ratings

Figure 13 shows the distribution of subjective comprehension ratings by *font size*. A Levene Test showed that the variances in the perception ratings were sufficiently equal to use parametric statistics ($F(5, 363) = 2.189, p = .055$). There was a significant effect of *font size* on comprehension ratings ($F(5, 363) = 18.614, p < .001$). The eta-square effect size ($\eta^2 = .193$) indicates large practical significance. Holm-corrected pairwise, repeated-measures t-tests revealed the following effects:

- For 10 points, comprehension ratings were significantly lower than for all larger fonts ($p < .001$, each).
- For 12 points, comprehension ratings were significantly lower than for all larger fonts ($p < .001$, each), as well.
- No significant differences were found between 10 and 12 points.
- Similarly, no significant differences were found between 14, 18, 22, and 26 points.

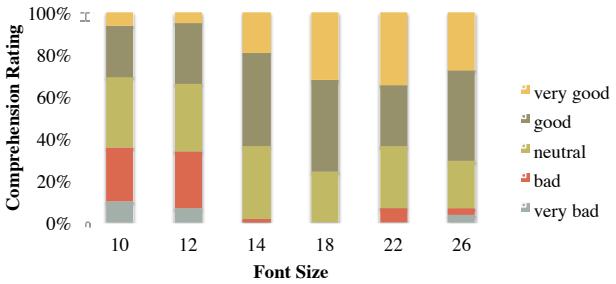

Figure 13. Subjective comprehension ratings increase with increasing font sizes.

Hence, in our study, **subjective comprehension of the texts was significantly lower for small font sizes (10 and 12 points)** compared to the larger tested font sizes.

Figure 14 shows the distribution of comprehension ratings by *line spacing*. A Levene Test showed that the variances in the perception ratings are sufficiently equal to were parametric statistics ($F(3, 365) = 2.209, p = .087$). There was a significant effect of *line spacing* on comprehension ratings ($F(3, 365) = 3.249, p = .022$). The eta-square effect size ($\eta^2 = .018$) indicates low practical significance. Holm-corrected pairwise t-tests revealed the following effects:

- For 1.0 comprehension ratings were significantly higher than for 1.8 ($p < .05$).

These results hint that our participants felt that their **comprehension was impaired by the largest line spacing**.

DISCUSSION

Font Size

Font size had significant effects on all dependent measures. The observed effects are consistent.

The average fixation durations decreased steadily until 22 points, which indicates that readability improved with increasing font size. The subjective measures confirmed

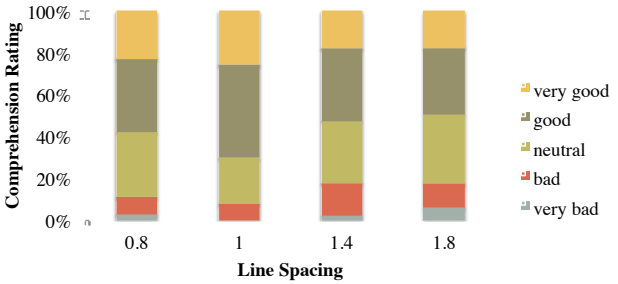

Figure 14. Subjective comprehension ratings were higher for 1.0 than 1.8 line spacing.

the objective measures: subjective readability was higher for the larger font sizes (14 to 26 points) than for the very small sizes (10 and 12 points), and it was best for 18 points. The results are inline with previous work comparing font sizes [1, 3, 4, 5, 6, 8], where the largest font size was found to be the best, either in terms of reading time or preference. However, those works on studied font sizes up to 14 points. Thus, it was not clear whether the readability of texts would improve beyond 14 points, and at what point the letters would become too large so that the effects get reversed.

Beyond readability, we also showed that comprehension was impaired by smaller font sizes. Our subjects gave more wrong answers for 10 and 12 than for 18 and 26 points. This shows that the measured, objective readability translated into measurable, objective improvement of comprehension. This is a notable insight, as 10 and 12 points happen to be font sizes which historically were very commonly used in websites [20].

Thanks to testing font sizes beyond the one that were typically studied in the past, we showed that the improvement in readability continues with increasing font size beyond 14 points. However, since no further significant improvements were observed beyond 22 points, and 18 points scores the best subjective readability. The results indicate that a local maximum might for objective readability exist between 18 to 26 points.[8] A local maximum is to be expected, as increasing font size will required to have less and less text in a single line, which leads to more frequent eye jumps, scrolling, and the loss of overview [17].

This results could also be expected for target populations with reading disorders such as people with dyslexia. In fact, we replicated this experimental setting with smaller group of 28 people who were diagnosed dyslexia and found similar effects of font size [29].

These findings advance our knowledge in two important aspects: first, beyond readability, this study is –to the best of our knowledge– the first to prove an effect of font sizes on both, objective and subjective readability and comprehension for a general target population. Second, since we tested larger font sizes as used in previous studies, we could show that the positive effect continues until 18 points, before it flattens out. This finding is in contrast with common recommendations, which suggest to use 10, 12, or 14 points.

[8]For the fixation duration the minimal value was attained for 22 points.

Line Spacing

The effects of *line spacing* were less pronounced. Our study revealed significant effects on comprehension, but not on readability.

Objective comprehension was lower for small line spacings: in the 0.8 line spacing condition, less comprehension questions were answered correctly. Further, subjective comprehension was higher for the standard spacing compared to the largest spacing (1.8). Thus, our study provides evidence that comprehension of texts may be impaired when line spacings are too small or too large.

Yet, the data did not reveal effects as pronounced as by font size. This corroborates the assumption by Bix [9] that line spacing is not a major factor on readability and that the ideal line spacing depends on other factors. Nevertheless, our work extends previous work in one important aspect: while previous work argues over readability, this work is –again to the best of our knowledge– the first to show that line spacing affects the comprehension of texts as well.

Limitations of the Study

One of the limitations of our study is that we only considered the first three paragraphs of Wikipedia articles. When using eye-tracking to study reading, it has been found that the initially measured fixation durations are longer, since users are still in a familiarization phase [22]. The heat map in Figure 15 shows that this effect occurred in our setup, too. However, the heat map also shows that the fixation durations normalize when reading on. Yet, since we assume that people often only read parts of web pages, we conclude that despite the short lengths of the texts, our findings have high ecological validity, that is, this familiarization also happens when people read web pages [11].

In comparison to other previous work [1, 3, 4, 5, 14], we did not measure reading time. We did so for two reasons. First, we wanted to create a natural setting, in which reading as fast as possible is neither a goal nor an indicator for readability. Second, reading fast can, in our opinion, be misleading. For example, in case of bad readability, participants might become frustrated and start skimming the text instead of reading it with full attention. Our decision is backed up by findings from Beymer *et al.* [5], who found significantly longer fixations for smaller fonts but no significant effect of font size on reading speed. Since we used comprehension tests, our evidence indicates that participants did not skim texts.

Another limitation of our study is that we used a fixed line length, as the browser window was maximized throughout the study. Previous research [16, 32] has

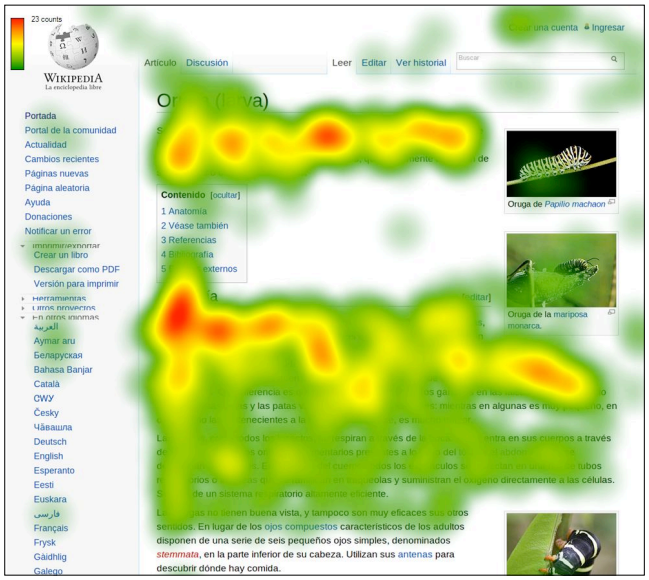

Figure 15. Heat map of a Wikipedia article used in the study (18 points and 1.0 lines for line spacing).

found that line length affects reading, and in the real world, people may shrink or enlarge the browser window freely. Yet, the typical browser will not change its window size when changing the font size. Some websites use a fixed width to display texts. Hence, our design has high ecological validity and allows applying our findings to typical reading settings.

Previous work [3, 10] has shown that readability of texts also depend on the font type. Since we only used a single font, namely *Arial*, our findings might not be generalize to other font types. However, *Arial* is one of the most widely encountered fonts in the web, as it is the default sans-serif font in most modern web browsers. More important, we believe that our work shows a clear indication that bigger sizes of similar font types lead to better reading and comprehension, encouraging designers to, regardless of the font type, think about and argue for bigger font sizes.

RECOMMENDATIONS

On the basis of our results, we recommend to use at least 18-point font size for the text body of websites. This value strikes the balance between having the best readability, comprehension, subjective perception scores, and allowing to fit as much text on the screen as possible. In our experiment, increasing font size beyond 22 points led to no significant improvements. Best subjective readability was achieved with 18 points. This forms a stark contrast to existing recommendations and guidelines, which typically suggestion font sizes from 10 to 14 points.

Regarding line spacing, our data suggests that it is best not to deviate too much from the standard line spacing (1.0). However, moderately larger line spacing, such as the widely used 1.5 spacing (e.g. in journal manuscripts), might be equally well to ensure readability and comprehension.

CONCLUSIONS

We tested the effect of font size and line spacing on objective and subjective readability and comprehension of Wikipedia articles. Up to a font size of 18 points, subjective and objective readability as well as comprehension improved continuously. Beyond 22 points, there were no further effects for the objective measures, and a decrease in subjective readability. Line spacing, in contrast, had no effect on the objective readability, but extreme spacings (0.8 and 1.8) negatively affected objective and subjective comprehension.

Our work advances previous knowledge, as it is the first work (1) to study reading with a general population in the context of the Web and (2) to show that readability improves past the typically studied font sizes (10, 12, and 14 points). It demonstrates that a simple increase in font size is a cheap and efficient way to improve access to textual information. Thus, it is a great and welcome development, that today's (2016) browsers, such as Firefox or Chrome, ship with a default font size of 16 points.

Future work needs to explore whether these findings are stable when other parameters, such as font type or column width, are altered. In particular, since more and more reading is taking place on tablets and mobile phones with much smaller screens, additional studies are required to verify our findings for those devices.

ACKNOWLEDGMENT

We thank the participants of the study.

REFERENCES

1. Jayeeta Banerjee, Deepti Majumdar, Madhu Sudan Pal, and Dhurjati Majumdar. 2011. Readability, subjective preference and mental workload studies on young indian adults for selection of optimum font type and size during onscreen reading. *Al Ameen Journal of Medical Sciences* 4 (2011), 131–143.
2. Michael Bernard, Chia Hui Liao, and Melissa Mills. 2001. The Effects of Font Type and Size on the Legibility and Reading Time of Online Text by Older Adults. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems (CHI EA '01)*. ACM, New York, NY, USA, 175–176. DOI: http://dx.doi.org/10.1145/634067.634173
3. Michael Bernard, Bonnie Lida, Shannon Riley, Telia Hackler, and Karen Janzen. 2002. A comparison of popular online fonts: Which size and type is best. *Usability News* 4, 1 (2002), 2002.
4. Michael L Bernard, Barbara S Chaparro, Melissa M Mills, and Charles G Halcomb. 2003. Comparing the effects of text size and format on the readibility of computer-displayed Times New Roman and Arial text. *International Journal of Human-Computer Studies* 59, 6 (2003), 823–835.

5. David Beymer, Daniel Russell, and Peter Orton. 2008. An Eye Tracking Study of How Font Size and Type Influence Online Reading. In *Proceedings of the 22Nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 2 (BCS-HCI '08)*. British Computer Society, Swinton, UK, UK, 15–18. http://dl.acm.org/citation.cfm?id=1531826.1531831
6. David Beymer and Daniel M. Russell. 2005. WebGazeAnalyzer: A System for Capturing and Analyzing Web Reading Behavior Using Eye Gaze. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05)*. ACM, New York, NY, USA, 1913–1916. DOI: http://dx.doi.org/10.1145/1056808.1057055
7. David Beymer, Daniel M Russell, and Peter Z Orton. 2007. An eye tracking study of how font size, font type, and pictures influence online reading. *Proceedings INTERACT 2007* (2007), 456–460.
8. Sanjiv K Bhatia, Ashok Samal, Nithin Rajan, and Marc T Kiviniemi. 2011. Effect of font size, italics, and colour count on web usability. *International Journal of Computational Vision and Robotics* 2, 2 (2011), 156–179.
9. Laura Bix. 2002. The Elements of Text and Message Design and Their Impact on Message Legibility: A Literature Review. *Journal of Design Communication* 4 (2002).
10. Dan Boyarski, Christine Neuwirth, Jodi Forlizzi, and Susan Harkness Regli. 1998. A Study of Fonts Designed for Screen Display. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '98)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 87–94. DOI: http://dx.doi.org/10.1145/274644.274658
11. Georg Buscher, Ralf Biedert, Daniel Heinesch, and Andreas Dengel. 2010. Eye tracking analysis of preferred reading regions on the screen. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*. ACM, 3307–3312.
12. C. Chapman. 2011. The most popular fonts used by designers. http://www.webdesignerdepot.com/2011/08/the-most-popular-fonts-used-by-designers/. (August 2011). last accessed Jan 8, 2016.
13. Jan Constantin. 2013. Typographic Design Patterns And Current Practices (2013 Edition). Smashing Magazine.http://www.smashingmagazine.com/2013/05/typographic-design-patterns-practices-case-study-2013/. (May 2013). last accessed Jan 8, 2016.
14. Iain Darroch, Joy Goodman, Stephen Brewster, and Phil Gray. 2005. The Effect of Age and Font Size on Reading Text on Handheld Computers. In *Proceedings of the 2005 IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'05)*. Springer-Verlag, Berlin, Heidelberg, 253–266. DOI: http://dx.doi.org/10.1007/11555261_23
15. S. Dehaene. 1992. Varieties of numerical abilities. *Cognition* 44 (1992), 1–42.
16. Marc C. Dyson and Mark Haselgrove. 2001a. The influence of reading speed and line length on the effectiveness of reading from screen. *Int. J. Human-Computer Studies* 54 (2001), 582–612. DOI: http://dx.doi.org/doi:10.1006/ijhc.2001.0458
17. Mary C Dyson and Mark Haselgrove. 2001b. The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human-Computer Studies* 54, 4 (2001), 585–612.
18. L. Evett and D. Brown. 2005. Text formats and web design for visually impaired and dyslexic readers-Clear Text for All. *Interacting with Computers* 17 (2005), 453–472. Issue 4.
19. M.A. Just and P.A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review* 87 (1980), 329–354.
20. Michael Martin. 2009. Typographic Design Patterns and Best Practices. Smashing Magazine.http://www.smashingmagazine.com/2009/08/20/typographic-design-survey-best-practices-from-the-best-blogs. (2009). last accessed Jan 8, 2016.
21. J. Nielsen. 2012. Lower-literacy users. Jakob Nielsen's Alertbox. http://www.useit.com/alertbox/20050314.html. (March 2012). last accessed Jan 8, 2016.
22. Jakob Nielsen and Kara Pernice. 2010. *Eyetracking web usability*. New Riders Pub.
23. Beth A O'Brien, J Stephen Mansfield, and Gordon E Legge. 2005. The effect of print size on reading speed in dyslexia. *Journal of Research in Reading* 28, 3 (2005), 332–349.
24. Donald G. Paterson and Miles A. Tinker. 1929. Studies of typographical factors influencing speed of reading. II. Size of type. *Journal of Applied Psychology* 13, 2 (1929), 120.
25. Donald G Paterson and Miles A Tinker. 1932. Studies of typographical factors influencing speed of reading. VIII. Space between lines or leading. *Journal of Applied Psychology* 16, 4 (1932), 388.
26. Christopher Power, Helen Petrie, David Swallow, Emma Murphy, Bláithín Gallagher, and Carlos A Velasco. 2013. Navigating, Discovering and Exploring the Web: Strategies Used by People with Print Disabilities on Interactive Websites. In *Human-Computer Interaction–INTERACT 2013*. Springer, 667–684.

27. K. Rayner and S.A. Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition* 14, 3 (1986), 191–201.

28. Luz Rello and Mari-Carmen Marcos. 2012. An Eye Tracking Study on Text Customization for User Performance and Preference. In *Proceedings of the 2012 Eighth Latin American Web Congress (LA-WEB '12)*. IEEE Computer Society, Washington, DC, USA, 64–70. DOI: http://dx.doi.org/10.1109/LA-WEB.2012.13

29. Luz Rello, Martin Pielot, Mari-Carmen Marcos, and Roberto Carlini. 2013. Size Matters (Spacing Not): 18 Points for a Dyslexic-friendly Wikipedia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility (W4A '13)*. ACM, New York, NY, USA, Article 17, 4 pages. DOI: http://dx.doi.org/10.1145/2461121.2461125

30. Jeff Sauro and Joseph S. Dumas. 2009. Comparison of Three One-question, Post-task Usability Questionnaires. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 1599–1608. DOI: http://dx.doi.org/10.1145/1518701.1518946

31. S.C. Sereno and K. Rayner. 2003. Measuring word recognition in reading: eye movements and event-related potentials. *Trends in Cognitive Sciences* 7, 11 (2003), 489–493.

32. A. Dawn Shaikh. 2005. The Effects of Line Length on Reading Online News. *Usability News* 7, 2 (2005), 1–4.

33. Tobii Technology. 2005. Product description Tobii 50 Series. (2005).