Impact of Sentence length on the Readability of Web for Screen Reader Users

Bam Bahadur Kadayat[1] and Evelyn Eika[1]

[1] Faculty of Technology, Art and Design, Oslo Metropolitan University, Oslo, Norway
S310223@oslomet.no, evelyn.eika@oslomet.no

Abstract. Readability of text is generally believed to be connected to sentence length. Most studies on readability are based on visual reading. Less is known about text readability for users relying on screen readers, such as users who are blind. This study therefore set out to investigate the effect of sentence length on the readability of web texts accessed using screen readers. A controlled within-subjects experiment was performed with twenty-one participants. Participants used a screen reader to read five texts with different sentence lengths. The participants' comprehension and perceived workload were measured. The findings reveal that there is a significant effect of sentence length and most participants exhibit the highest comprehension and lowest workload with sentences comprising 16-20 words. Implications of these results are that web content providers should strive for sentence length of 16-20 words to maximize readability.

Keywords: Readability, workload, sentence length, screen reader, blind, accessibility, universal design.

1 Introduction

Readability is the measure of ease or difficulty with which the text can be read and understood by an intended reader who is reading for a specific purpose [1]. Readability is affected by several factors such as content, structure, readers' knowledge, vocabulary, layout, and design [2]. It can be challenging to read web content using screen reader software. Screen reader users also have difficulties re-tracking the reading content as software does not read it back. Users may not recall what they read, which leads to comprehension difficulties. This study investigates web readability for screen reader users, in particular, the factor of sentence length: its impact and its appropriate length.

2 Background

Most studies indicated that sentence length affects the readability where long sentences are harder to read than shorter sentences [3, 4, 5]. Shorter sentences, however, do not necessarily improve readability because of other factors such as vocabulary and coherence. Word difficulty is another factor that is commonly mentioned. A sentence with difficult words is harder to read compared to one without. The sentence and the word

length are the two attributes that are used in readability measures [3] such as the Flesch-Kincaid reading ease index. This popular readability measure is designed to quantify how difficult it is to comprehend a reading passage. The score ranges between 0 and 100, where a high score indicates easy to read and low score hard to read [6]. However, such readability measures tend to be over simplistic and are mostly used for printed text [3]. Also, text on web is read differently compared to printed text [7].

2.1 Screen reader users and the web

About 45 million blind people worldwide access websites using screen reader technology [8]. Text is a significant part of the web and reading through web applications is an especially challenging task for blind users. An accessible and readable web content thus allows blind users to access and understand its information. Also, web developers and designers were often unaware about the impact of non-visual web content for blind users hindering accessibility of Websites [8]. WCAG 2.0 offers a broad set of recommendations for making the web content accessible and readable [9].

2.2 Readability on the web

Gottron and Martin [10] employed content extraction algorithms to determine the readability of web documents. They analyzed 1114 documents from five websites and compared quantitative readability measures along with their adjusted content filters (i.e., the adapted content code blurring and document slope curves algorithms). They showed that embedding adjusted content extractions for SMOG and Flesch Reading Ease indexes yielded more accurate readability estimates. The results support a solution where corpus statistics is employed on the web to achieve language-independent measures of readability.

de Heus and Hiemstra [11] used the Automated Readability Index (ARI) to determine the mean grade level needed to understand a website. They used MapReduce for real-time calculation of the readability of more than a billion webpages. The datasets called *Common Crawl* included 61 million domain-names, 92 million PDF documents, and seven million Word documents. About 60 % of the information originated from commercial, organization, and network websites. The cumulative results showed that 12-year-olds, 23-year-olds, and 18-year-olds can comfortably comprehend 25 %, 75 %, and 50 % of the content on web, respectively.

Chung, Min, Kim, and Park [12] investigated the readability of text-based web documents for deaf people. They proposed a newscasting display technique which converted difficult sentences into easy sentences and indicated the relationship with the help of visual illustration. They developed a system consisting of a graphical representation module and a structural simplification module to visualize the relationships between simple fragmented sentences. However, the system was found not easy to use for low literacy deaf people.

2.3 Readability for blind users

Guerreiro and Goncalves [13] investigated whether increasing speech rates affected content scanning with concurrent speech. They recruited 30 visually impaired participants and focused on relevance scanning from two-hundred Portuguese news with three main topics (sports; politics and economy; and television, celebrities, and arts). The results showed that concurrent speech (two and three-voices) of a speech rate slightly higher than the default rate greatly increased scanning speed for relevant information. Their findings suggest that two-voices with a rate 1.75 times the default-rate (ca. 278 words per minute) enables the appropriate scanning without loss in performance.

2.4 Optimal sentence length

Mikk [14] examined young adults' cognitive load involving sentence length. A total of 37 students (17-18 years old) participated in their study. A total of 30 texts were taken from scientific books. Cloze tests were carried out where the participants needed to fill in the blanks with deleted words. The results showed that 50-130 characters were appropriate for these students. The findings also demonstrated that the too short and too long sentences were not suitable for participants' memory workload.

Cutts [15] did not recommend an upper limit sentence length, even though sentence lengths exceeding 40 words discouraged readers. A better goal for an average sentence length is said to be 15-20 words. Cutts argued that the word length is an average and it is not necessary for all the sentences to be in this range. Other ranges are possible.

2.5 Impact of sentence length on readability

Oelke, Spretke, Stoffel, and Keim [16] presented a tool named VisRa to assist authors to make their writing easier to read. This tool indicated complex paragraphs and sentences which were harder to comprehend. VisRa provided feedback for correcting a text. The feedback showed not only issues on sentences, but also it explained why sentences were hard to read. VisRa gives the following features: word length (the mean number of characters per word), vocabulary difficulties (the percentage of terms not listed in the common list), nominal forms (the noun ratio), and sentence length (the number of words in a sentence).

Sherman (as cited in DuBay [17]) compared older writers with modern writers and observed a trend where sentences have become shorter over time by statistically analyzing sentence lengths. His analyses showed that on average 50 words were used per sentence before the Elizabethan era, and it reduced to an average of 29 words per sentence during Victorian times and to 23 words during 1893 [17]. Currently, the average sentence length is 20 words per sentence.

In this study two questions were asked: Does sentence length affect the readability of web in terms of workload for screen reader users? What is the appropriate sentence length that makes web content readable and understandable for screen reader users? To answer the questions, two predictions were formed. First, there is a significant impact of sentence length on web readability for screen reader users. Second, a minimum of

sentence length is most suitable for screen reader users to read and comprehend web content in terms of subjective workload. We therefore formulated the following two null-hypotheses:

1. H_0: There is no significant impact of sentence length on the readability of the web for screen reader users.
2. H_0: Minimum sentence length on the web will not be appropriate for screen reader users in terms of subjective workload.

3 Method

A controlled within-subject [22, 23] experiment was conducted to collect quantitative data. The data included (a) workload perceived by participants while reading each prototype and (b) comprehension test after reading each prototype. The word length was the independent variable with five levels: 10-15 words, 16-20 words, 21-25 words, 26-30 words, and 30 or more words. The two dependent variables included the comprehension score and NASA-TLX scores. A one-way repeated measures ANOVA was employed to verify whether sentence length has impact on the readability of the web for screen reader users. It was assessed based on the workload that participants experienced while reading the content of five prototypes through a screen reader technology.

3.1 Participants

Thirteen males and eight females participated in the study (N = 21) with a mean age of 28 years (13 from 26-30 years; 5 from 31-35 years; and 1 from 20-25, 36-40, and 41 and above, respectively). The participants were recruited from Oslo Metropolitan University. All participants were non-native speakers but read English fluently (1 at intermediate level and 20 at an advanced level). Most were from the Master program of Universal Design of Information and Communication Technology, and few were from other educational background. Twenty participants were master students, and one was a bachelor student. None was recognized as a blind participant.

3.2 Materials

Five webpages were chosen for each test prototype. Five comprehension tests were constructed for each reading task. The NVDA (non-visual desktop access) screen reader was used as assistive tool.

Fig. 1 shows the five prototypes (webpages) of different sentence length that were created for the experiment (Prototype A, 10-15 words; Prototype B, 16-20 words; Prototype C, 21-25 words; Prototype D, 26-30 words; and Prototype E, 30 words or more). The contents of all the prototypes were taken from online news portals including BBC, Yahoo, Norway today, New York Times, and The Local. All the pages addressed different topics (e.g., technology, education, and entertainment) but they had similar layout. Each prototype consisted of two same-length sentences.

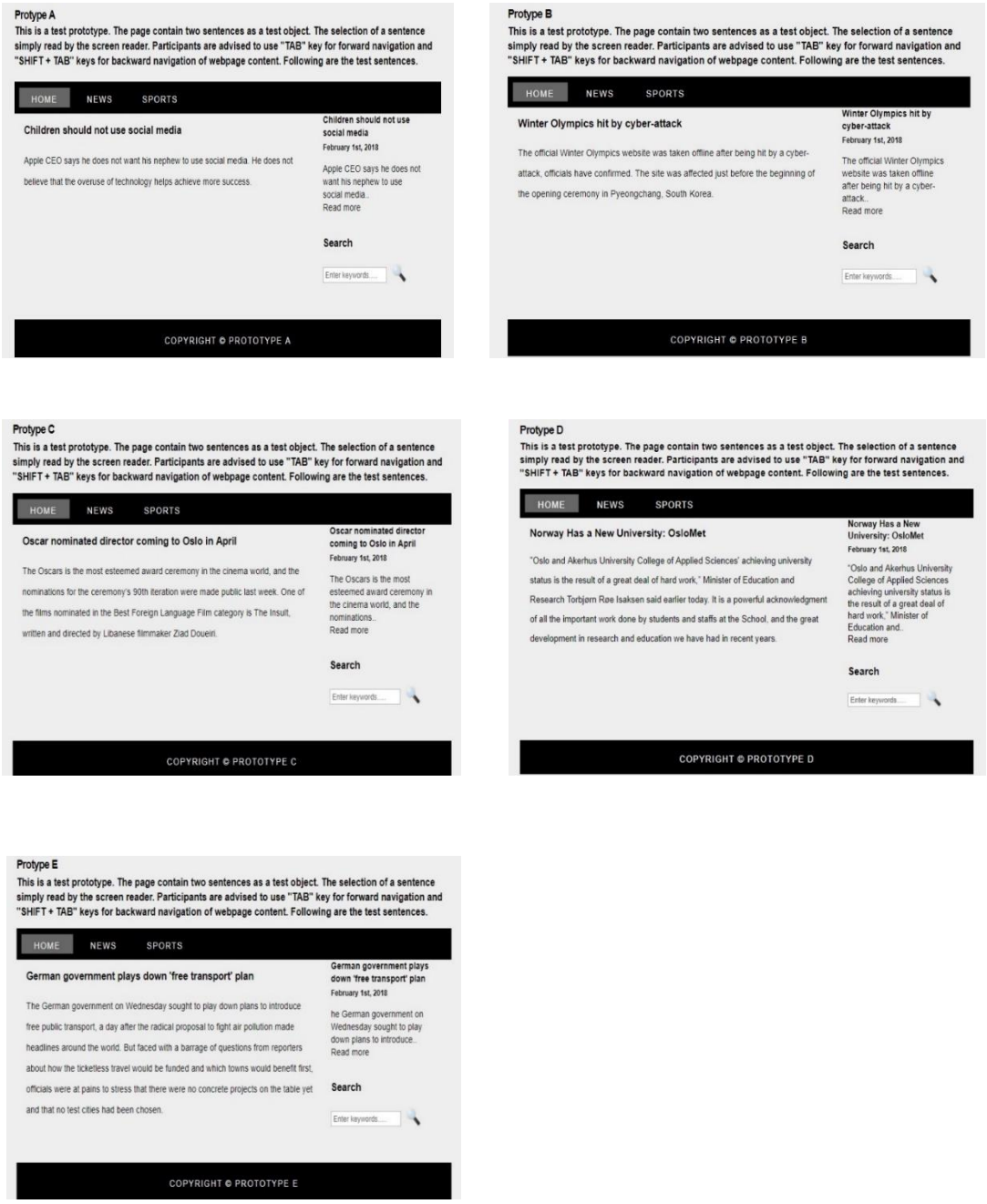

Fig. 1. The five prototype webpages with different sentence lengths: (a) 10-15 words, (b) 16-20 words, (c) 21-25 words, (d) 26-30 words, and (e) 30 words or more.

3.3 Observations

A Comprehension test was used to assess the participants' ability to read and understand the content of the prototypes. Each test consisted of two multiple-choice questions.

NASA-TLX was used to measure the perceived workload of the participants in the experiment process. It is a standardized tool used by many studies in Human Factors and Ergonomics [18]. It is shown to be highly reliable and valid [19]. Moreover, it is used for subjective multidimensional workload assessment along the dimensions of Mental Demands (Md), Physical Demands (PD), Temporal Demands (TD), Own Performance (OP), Effort (EF), and Frustration (FR) [20]. It helps determine the perceived workload of a participant while performing a task since mental workload varies among individuals [23].

Twenty step bipolar scales (semantic differentials) were applied to get ratings for the dimensions [24]. Bipolar is a specific type of rating scale characterized by a range between two opposite endpoints. A score ranged from 0 to 100 (allocated to the closest point 5) was taken on each scale (Ibid.). After the participants' ratings, 15 possible pairwise comparisons were conducted in terms of six scales [18].

3.4 Procedure

The participants were first given an information sheet and then familiarized with the NVDA screen reader tool. They were asked to use an eye-mask during the experiment to cover their eyes (blindfolded) during the reading task. The prototypes were started for the participants to read their contents through the screen reader. After reading each prototype, the participants were to remove their eye-mask and take the comprehension test consisting of two multiple-choice questions. Immediately they were to rate their perceived workload using NASA-TLX on paper. They were then to take a short rest (1-2 minutes) before reading the next prototype. The whole experiment took about 40 minutes for each participant. Five participants at a time participated in the experiment. The prototype presentation order was randomized to minimize bias [24].

3.5 Analysis

One-way repeated measures ANOVA were employed to verify sentence length impact on web readability using SPSS version 24.0 for Windows [29].

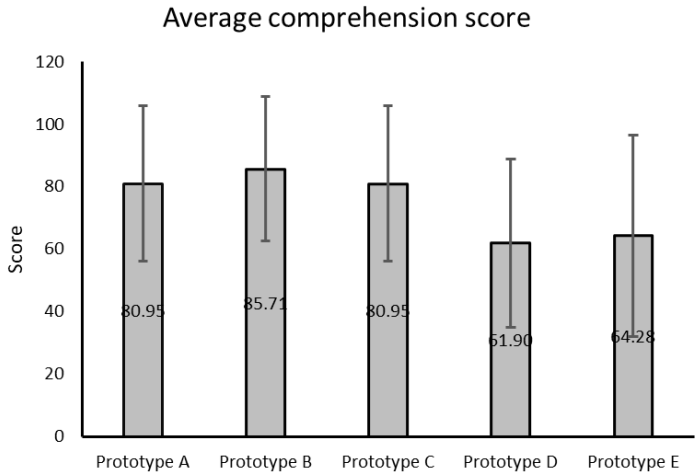

Fig. 2. Mean comprehension test score for the five prototypes. Error bars show standard deviation.

4 Results

4.1 Comprehension scores

Fig. 2 showed the average comprehension test score of each prototype. The total score was 100, and all participants scored above 50% in each test. There was a significant difference in comprehension score among prototypes B, D, and E but prototypes A and C had the same mean score. The results showed that prototype B had the highest mean score (M = 85.7, 16-20 words) whereas the prototype D had the lowest mean score (M = 61.9, 26-30 words). Based on comprehension scores, the results suggested that web contents with shorter sentences tended to be easier to comprehend than those with longer sentences.

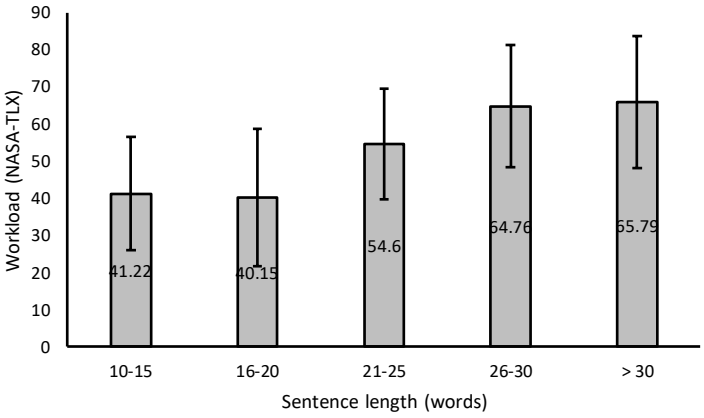

Fig. 3. NASA-TLX workload scores for the five prototypes. Error bars show standard deviation.

4.2 NASA-TLX workload scores

The NASA-TLX scores are shown in Fig. 3. Mauchly's test of sphericity revealed that the assumption of sphericity was violated ($\chi^2(9) = 28.17$, $p < .001$) for the TLX observations. A Greenhouse-Geisser correction was therefore applied [25, 26] since the epsilon was less than 0.75. The results showed that there was a significant effect of sentence length on average workload ($F(2.26, 45.27) = 19.77$, $p < .001$). Descriptive statistics for the five-level independent variables (prototypes A to E) showed that the participants used less workload on prototype B ($M = 40.15$, $SD = 18.45$) compared to other prototypes. Post-hoc tests revealed that prototype B was significantly different compared to the other prototypes ($p < .001$) apart from prototype A ($p = .67$). Prototype A ($M = 41.22$, $SD = 15.30$) had just a slightly higher workload than prototype B.

The highest workload was observed for prototype E ($M = 65.79$, $SD = 17.77$) which was significantly different to the other prototypes ($p < .001$) apart from prototype D ($p = .76$). Similarly, prototype C ($M = 54.60$, $SD = 14.79$) and prototype D ($M = 64.76$, $SD = 16.51$) exhibited a significantly mean difference of 10.14 ($p < .01$). Prototype B had the lowest workload mean among all the prototypes.

The results indicated statistically significant differences of mean workload across the five prototypes, except in between prototypes A and B (towards the shortest length), and between D and E (towards the longest length). The overall findings illustrated that the participants experienced significantly less workload with prototype B (second shortest) while reading sentences on web compared to prototypes A, C, D, and E. This evidence supports the hypothesis that sentence length significantly impacts readability of the web for screen reader users.

5 Discussion

This study investigated how sentence length impacts the readability of web for screen reader users. Comprehension tests were conducted before measuring the workload of the participants. The comprehension tests helped verify whether a participant could read and comprehend the prototype contents through multiple-choice questions. The comprehension tests showed that most participants understood the prototype contents as their comprehension scores were above 60 %. Most of the participants were students in the Master of Universal design of ICT study program and they therefore had knowledge about screen readers. Their education level might have affected the comprehension results positively as they were more experience with comprehending complex texts compared to the general population. However, they had difficulties recalling all the words of the sentences as a screen reader read the content only once. They experienced even more challenges for the longer sentences. Most participants understood the prototype B content (16-20 words), related to "The official Winter Olympics website." This could be attributed to greater attention of the participants towards the cyber-attack of Olympics website because young people may be highly attracted to games and sports.

The ANOVA results indicate a significant impact of sentence length over five prototypes. There was a considerable mean difference across five prototypes in terms of subjective workload. The results based on the subjective workload mean that the first null hypothesis (H_0) can be rejected and instead the alternative hypothesis (H_1) can be accepted. It suggests that screen reader users' comprehension can be affected by sentence length. Thus, it is advisable that web content authors use appropriate sentence length to create web content to assist screen reader users. The post-hoc tests indicate that prototypes D and E had the smallest mean difference, followed by A and B (second smallest mean difference). The increasing mean difference across prototype A through prototypes B, C, D and E signifies an increasing degree of complexity in readability of the text.

Overall prototype B (with sentence length of 16-20 words) had the desired preferences for reading by the participants as all the participants exhibited the least physical demand workload. This might be because of less body movement required of the participants. The temporal demand was also the lowest with prototype B compared to other prototypes. Also, prototype B exhibited the overall lowest workload thereby featuring the maximum readability. However, the sentence length of prototype B was not the shortest among the prototypes. Hence, this result does not support the second hypothesis that the shortest sentence length will be most appropriate for the screen reader users on the readability of the web. Nevertheless, it is evident that the length of sentences on the web for the screen reader users has impact on readability. Concerning the appropriate sentence length, it also depends on the readers' language proficiency, reading skills, and memory workload.

One limitation of this study concerns recruiting fully blind participants, which is a challenging task. It was thus decided to recruit sighted participants which were blindfolding with an eye-mask. The findings may have been different if actual blind users were recruited as blind users depend mainly on their auditory and touch senses to substitute their lack of visual hints while interacting with the environment [27] and blind

users are likely more experienced using screen readers. As also noted, some participants found it awkward to be blindfolded. Another factor that might have affected the results is the complexity of the measurement scales. As the NASA-TLX workload measurement scale relies on subjective perceptions of the participants, there might be individual differences in understanding and completing the measurement scales. As observed, some participants experienced difficulties using the NASA-TLX bipolar scales (20 steps) ratings from 0 to 100 scores. Immediately after the reading tests, the participants needed to circle the factor which affected the workload of the task. A skilled participant may find it easy to perform these tasks whereas others might find it challenging to perform the same task in the same situation. Some participants also found that the physical demand factor was not relevant because they did not perform any physical strenuous task during the experiment (except using the screen reader and manually filled in the forms).

6 Conclusion

This study examined the impact of sentence length on the readability of the web for screen reader users and explored suitable word lengths. The results indicate that there is a significant difference in the workload of the participants over five prototypes (websites) of varied sentence lengths. Regarding the appropriate length of sentences, it is not mandatory for sentence length to be as short as possible. Prototype B with sentence lengths of 16 to 20 words shows the lowest workload thereby exhibiting maximum readability. The result thus suggests that using sentences of 16-20 words may be appropriate for screen reader users when performing reading tasks on the web. Future work may address additional web content types with more varied word lengths and sentence lengths. Future studies might also address in-depth analysis of NASA-TLX.

7 References

1. Pikulski, J. J.: Readability. Retrieved January, 10, (2020).
2. Owu-Ewie, C.: Readability of comprehension passages in Junior High School (JHS) English textbooks in Ghana. Ghana Journal of Linguistics 3(2), 35-68 (2014).
3. Eika, E., Sandnes, F. E.: Assessing the Reading Level of Web Texts for WCAG2. 0 Compliance—Can It Be Done Automatically? In Advances in Design for Inclusion (pp. 361-371): Springer (2016).
4. Eika, E., Sandnes, F. E.: Authoring WCAG2. 0-compliant texts for the web through text readability visualization. In International Conference on Universal Access in Human-Computer Interaction, pp. 49-58. Springer, Cham (2016).
5. Eika, E.: Universally designed text on the web: towards readability criteria based on anti-patterns. Stud. Health Technol. Inform 229, 461-470 (2016).
6. Kincaid, J.P., Fishburne, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. Research Branch Report 8–75. Chief of Naval Technical Training: Naval Air Station Memphis (1975).
7. Gottron, T., Martin, L.: Readability and the Web. Future Internet 4(1), 238-252 (2012).

8. Babu, R., Singh, R., Ganesh, J.: Understanding blind users' Web accessibility and usability problems. AIS Transactions on Human-Computer Interaction 2(3), 73-94 (2010).
9. World Wide Web Consortium.: Web content accessibility guidelines (WCAG) 2.0 (2008).
10. Gottron, T., Martin, L.: Estimating web site readability using content extraction. Paper presented at the Proceedings of the 18th international conference on World wide web (2009).
11. de Heus, M., Hiemstra, D.: Readability of the Web: A study on 1 billion web pages. Paper presented at the DIR (2013).
12. Chung, J. W., Min, H. J., Kim, J., Park, J. C.: Enhancing readability of web documents by text augmentation for deaf people. Paper presented at the Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics profile data 24(2), 95-112 (2013).
13. Guerreiro, J., Gonçalves, D.: Faster Text-to-Speeches: Enhancing Blind People's Information Scanning with Faster Concurrent Speech. Paper presented at the Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (2015).
14. Mikk, J.: Sentence length for revealing the cognitive load reversal effect in text comprehension 34(2), 119-127 (2008).
15. Cutts, M.: Oxford guide to plain English: OUP Oxford (2013).
16. Oelke, D., Spretke, D., Stoffel, A., Keim, D. A.: Visual readability analysis: How to make your writings easier to read. IEEE Transactions on Visualization and Computer Graphics 18(5), 662-674 (2012).
17. DuBay, W. H.: The Principles of Readability. Online Submission (2004).
18. Hart, S. G.: NASA-task load index (NASA-TLX); 20 years later. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 50, No. 9, pp. 904-908). Sage CA: Los Angeles, CA: Sage Publications (2006).
19. Longo L.: On the Reliability, Validity and Sensitivity of Three Mental Workload Assessment Techniques for the Evaluation of Instructional Designs: A Case Study in a Third-level Course (2018).
20. Hart, S. G., Staveland, L. E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Advances in psychology (Vol. 52, pp. 139-183): Elsevier (1988).
21. De Alwis Edirisinghe, V.: Estimating Mental Workload of University Students using Eye Parameters. Masters thesis, NTNU (2017).
22. MacKenzie, I. S.: Within-subjects vs. Between-subjects Designs: Which to Use? Human-Computer Interaction: An Empirical Research Perspective 7, 2005 (2002).
23. Raluca, B.: Between-Subjects vs. Within-Subjects Study Design. Retrieved from https://www.nngroup.com/articles/between-within-subjects/ (2018).
24. Suresh, K.: An overview of randomization techniques: an unbiased assessment of outcome in clinical research. Journal of human reproductive sciences 4(1), 8. (2011).
25. Greenhouse, S. W., Geisser.: On methods in the analysis of profile data 24(2), 95-112 (1959).
26. Huynh, H., Feldt, L. S.: Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. Journal of educational statistics 1(1), 69-82 (1976).
27. Nielsen, J.: Cloze Test for Reading Comprehension. Retrieved from https://www.nngroup.com/articles/cloze-test-reading-comprehension/ (2011).
28. Rony, M. R.: Information Communication Technology to support and include Blind students in a school for all An Interview study of teachers and students' experiences with inclusion and ICT support to blind students (2017).

29. Green, S. B., Salkind, N. J.: Using SPSS for Windows and Macintosh, Books a la Carte: Pearson. (2016).