Guideline-Based Evaluation of Web Readability

Aliaksei Miniukovich
University of Trento
Trento, Italy
miniukovich@disi.unitn.it

Michele Scaltritti
University of Trento
Trento, Italy
michele.scaltritti@unitn.it

Simone Sulpizio
Vita-Salute San Raffaele University,
Milan, Italy
sulpizio.simone@hsr.it

Antonella De Angeli
Free University of Bolzano
Bolzano, Italy
antonella.deangeli@unibz.it

ABSTRACT

Effortless reading remains an issue for many Web users, despite a large number of readability guidelines available to designers. This paper presents a study of manual and automatic use of 39 readability guidelines in webpage evaluation. The study collected the ground-truth readability for a set of 50 webpages using eye-tracking with average and dyslexic readers (n = 79). It then matched the ground truth against human-based (n = 35) and automatic evaluations. The results validated 22 guidelines as being connected to readability. The comparison between human-based and automatic results also revealed a complex framework: algorithms were better or as good as human experts at evaluating webpages on specific guidelines – particularly those about low-level features of webpage legibility and text formatting. However, multiple guidelines still required a human judgment related to understanding and interpreting webpage content. These results contribute a guideline categorization laying the ground for future design evaluation methods.

CSS Keywords

• **Human-centered computing → HCI design and evaluation methods;** *Heuristic evaluations; Accessibility design and evaluation methods.*

Keywords: Accessibility; Design Guidelines; WCAG 2.1; User Experience; Web Design

ACM Reference Format:

Aliaksei Miniukovich, Michele Scaltritti, Simone Sulpizio, and An- tonella De Angeli. 2019. Guideline-Based Evaluation of Web Read- ability. In CHI Conference on Human Factors in Computing Sys- tems Proceedings (CHI 2019), May 4-9, 2019, Glasgow, Scotland UK. ACM, New York, NY, USA, Paper 508, 12 pages. https://doi.org/10.1145/3290605. 3300738

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300738

1 INTRODUCTION

Web use largely relies on reading. Improving the inherent webpage readability – readability that does not require external tools and is in-built in a webpage design – would improve both the efficiency of content communication and quality of web browsing experience. Such improvement would particularly benefit the user groups that struggle with reading, such as dyslexics and children. Dyslexia – a learning disability limiting the ability to read – affects about 7% of users [22] and makes low readability an accessibility issue, as it prevents such users from normally using the Web. Designers would not want to completely ignore such a large user group and should design for high readability.

However, the existing guidance for designers on Web readability requires further clarification and validation, because the guidelines often contradict each other or lack empirical support [30]. The use of readability guidelines in practice also needs further research, since the guidelines are too many and may be burdensome for designers to learn and meticulously apply to multiple webpages (cf., [23]). Guideline automation may alleviate these issues and aid designers, but needs more research on what guidelines can be automated and how it combines with human evaluation.

This paper contributes a validation of 39 readability guidelines, an exploration of guideline automatic evaluation, and categorization of guidelines based on their propensity for successful automation, which would inform the architecture of future semi-automatic methods of Web readability evaluation. Such methods will likely automate a part of the evaluation process to lessen the load on experts, but will still involve experts and designers to interpret and complement automatic evaluation.

2 BACKGROUND

Improving Web readability begins from understanding the determinants of readability and translating them in the intermediate-level concepts [15], such as design guidelines, which then can be used in Web design and evaluation. Substantial theoretical knowledge has been accumulated on dyslexia and reading difficulties, and guideline application.

2.1 Readability

Web readability could be positioned under the umbrella of both Web accessibility and Web usability. While the two concepts substantially differ – e.g., the usage problems reported by blind users differ from the problems reported by average users [34] – improvements in readability likely improve both accessibility and usability, as it simultaneously makes webpages more accessible to the users with reading difficulties [10,30] and reduces the user effort of regular readers [53,39]. Some research did highlight several text features to either exclusively affect dyslexics or to affect dyslexics more than average readers. These features include visual clutter and crowding [5], presence of noise and distractors [46], between-line and between-character spacing [56], and multiple other features. However, no recommended webpage adjustment for dyslexics decreased readability for the average readers, and vice versa [30].

The accumulated theoretical research on dyslexia and its causes have only rarely been translated in advice for designers, and further, in design improvements, despite graphical user interfaces (GUIs) being well suited to alleviate the issues of dyslexics due to their easy visual customizability [14]. Past attempts of alleviate such issues with specialized tools included automatically spacing out the sentences of paragraphs [53], showing simpler synonyms for complex words [37], and providing tools to change text formatting on demand [43]. Converting text to speech has also been widely marketed as a solution to the reading difficulties.

Improving Web readability requires first defining it as a concept, e.g., as the amount of mental effort needed to read and comprehend a piece of text. Literature further suggests that readability includes three different components: text legibility, text formatting, and text complexity. The first two components focus on the visual aspects of text. Legibility describes the effort to distinguish individual characters from the background and each other, and includes such visual aspects as text-background luminance contrast, letter spacing and letter case [1,21]. Text formatting applies to paragraphs rather than individual letters or words, and includes such visual aspects as between-line spacing, text justification or text column width [14,39]. Finally, text complexity focuses not on the visual, but semantic and structural aspects of text, which includes the length and structure of sentences, use of simpler synonyms for infrequent, archaic or lengthy words, or density of pronouns in a text [11,9].

2.2 Webpage Evaluation

While specialized software tools could aid dyslexics, Web designers should not rely on the hope the user has access to the right tools, as some evidence has suggested the average user with reading difficulties likely does not use them [14]. A better approach to improving Web readability would be redesigning webpages, so they are inherently more readable (cf., [10]).

Multiple design evaluation methods [40,8] could generate re-design recommendations. Empirical, user testing-based methods may suit usability re-design better than inspection-based methods, because observing the user struggle with tasks reveals critical issues [17] and such observations are relatively straightforward to interpret in re-design recommendations. However, inspection-based methods – which include experts reviewing webpages and summarizing recommendations in a report – suit readability re-design better than empirical methods because experts can rely on their theoretical knowledge of what affects readability and generate recommendations directly, whereas empirical observations of the user struggling to read register the fact that readability needs improvement, but do not tell what exactly needs to change and still require translation in recommendations by an expert.

For aiding experts in design inspections, past research particularly advocated for relying on readability guidelines [10]. Not only do the guidelines direct experts to the relevant aspects of Web designs, and thus, act as intermediate-level knowledge [15] more helpful to designers than purely theoretical knowledge, but they also summarize important aspects of readability in good design practices to be taught in class and in books. The guidelines can also serve as the basis for more advanced design inspection aids, such as design claims [47] and patterns [55,24]. Guidelines are succinct and can be applied – and thus tried out and validated – quickly. After the validation, they can be augmented with design principle explanations and application contexts to further help design evaluators, e.g., as design patterns [55].

2.3 Design Guidelines

HCI research has produced numerous design guideline sets. Some of them targeted various components of GUI quality, e.g., usability [18], accessibility [26], and utility [20]; some others a specific technological platform, e.g., mobile devices [52] and interactive TV applications [6]; and yet others a specific user demographic, e.g., older users [55] and children [25].

Readability has received less attention from the relevant HCI research, as most of it focused on accessibility in general or considered readability from the perspective of blind users [49,34]. The web content accessibility guidelines (WCAG[1]) – which the industry uses as the gold-standard accessibility guidelines – largely address making webpages compatible with the text-to-speech software, but do not focus on redesigning webpages to make them inherently more readable, without specialized software. The latest version, WCAG 2.1, was published after this research was carried out and does add several new guidelines for dyslexics and readability (e.g., success criteria 1.4.11 and 1.4.12), but still misses many other relevant guidelines.

Guidelines specifically for readability and dyslexics have also been published by academics and design practitioners [2,38,42]. However, recent research has reviewed such guidelines and highlighted several major issues with them, including guidelines from different sets contradicting each other or published empirical findings, guidelines with unclear wording, and outdated guidelines [30]. In addition, only a small subset of the reviewed guidelines was validated empirically [30], which would prevent a large-scale adaptation of such guidelines in design.

2.4 Automatic Evaluation

Despite their advantages, different sets of design guidelines have only had a limited proliferation in design practice. Even when clearly worded and well articulated, guidelines can be difficult to adhere to if they become too many [23] – a common phenomenon in guideline research (cf., [4]). Past research devised systems and tools for collecting, categorizing, and hierarchically displaying relevant guidelines [13], which mitigated, but did not solve the problem of too many guidelines.

Automated inspection for guideline compliance could help designers with guideline overload and other issues, e.g., substantial training needed to master the guideline use and subjectivity of individual expert evaluations [7]. Usability research has developed several tools for automated inspection. However, most of them had limitations. For example, all but six of 157 guidelines of WebTango [16] seemed unrelated to usability [31]; and W3Touch [33] only addressed usability issues with zooming and broken navigational elements. Recent research described more advanced usability-inspection systems [12,7] that could detect a wide range of issues with webpages, but crucially, further research is needed on combining automatic evaluation with human evaluation, as automatic systems are unlikely to fully replace human evaluation, since humans and algorithms detect different issues (cf., [30]) and humans may still need to check or interpret the output of algorithms.

Automated inspection of readability has progressed less than the inspection of usability or accessibility. Multiple freely-available accessibility validators[2] check for webpage compliance with WCAG, but their ability to detect readability issues was questioned [30]. Several tools could help users, including dyslexics, read by automatically transforming webpages [37,43,19,53], but they did not focus on evaluation or rely on readability guidelines. A recent study did describe automatic readability evaluation based on guidelines [30], but it did not scrutinize combining human evaluation with automatic evaluation, which is the focus of the present study.

3 STUDY

We explored the use of readability guidelines in manual and automatic webpage evaluation, to derive the insights for future readability and accessibility evaluation protocols and systems. Such exploration required multiple guideline be involved. The chosen guidelines were all previously described in the HCI literature, which let us justifiably presume they all impacted readability. This presumption and needing to involve many guidelines – which a controlled experiment could not practically do – led us to design the exploration as a correlational study.

The study collected and compared three pieces of data for a set of webpages: the ground truth readability scores, manual evaluation with readability guidelines, and automatic evaluation on guideline-related metrics.

3.1 Stimuli

We sampled 117 webpages in Italian from the websites of news, non-profit, and governmental organizations, taking only one webpage per website to maximize the diversity in webpage appearance. The sampled webpages featured an article about health, research, new technology, or education – these topics were deemed to be sufficiently engaging for both children and adults to stay focused throughout an experimental session and read through texts without skipping. One of the authors queried a search engine (with topic-related keywords, e.g, "scientific discoveries" or "digital technology") and sampled webpages uniformly from the top-200 search results. We avoided widely-known websites to minimize familiarity effects and overly long articles (>5K characters), as long texts would unnecessarily

[1] https://www.w3.org/TR/WCAG21/

[2] E.g., AChecker, http://achecker.ca/checker/index.php

burden participants with dyslexia, particularly children. After filtering out several webpages that contained potentially inappropriate content for children (e.g., a picture of human skeleton in a health article), we sub-sampled 50 webpages, (Figure 1) while maximize the variance in visual and textual features in the final sample (the features were pre-computed for all 117 webpages, and 50 of them were selected while maximizing feature-score variance and keeping feature distributions closer to normal, across all features). All webpages were saved as screenshots (PNG 24-bit per pixel, 1600 pixel wide, full page length).

Figure 1. Examples of webpages used in the study.

3.2 Readability Guidelines

We used readability guidelines from a recent study that collected, disambiguated, categorized, and reviewed a set of guidelines with design and dyslexia experts [30]. Out of 47 guidelines listed as applicable to individual webpages, we omitted ten guidelines that appeared likely to be violated only by very few webpages (e.g., few, if any, webpages would not comply with "*Avoid using more than one whitespace after period*" and none of our sampled webpages had lists to comply with "*Use the lists of dos and don'ts, which are more useful than continuous text to highlight aspects of good practice*") or required specific tasks to be meaningful (e.g., search-related tasks for "*Always put the search box in a clear obvious position, usually the top of the page*" or navigation tasks for "*Use a breadcrumb trail (e.g., 'Home page > section 1 > sub-section 1.1') to let the user understand their location on a website*"). We also split two guidelines in four because they consisted of two parts, which resulted in the total of 39 guidelines, Table 1

3.3 Ground Truth

We measured the ground-truth readability of the 50 webpages using eye-tracking. Past research [32,45,35] explicitly linked eye-tracking features – such as, average duration of eye fixations or number of fixations – to reading performance. Eye-fixation durations depend on multiple perceptual, linguistic and typographic features of text [32], and directly correspond to reading speed: when durations are summed up and normalized by the amount of text, they describe reading time per word. Eye-fixation counts describe not only the amount of text, but also the amount of regressive saccades and re-reading, which are markers of lower readability.

Past work [30] similar to this study collected subjective readability ratings, which was less effortful than eye-tracking for researchers, but potentially introduced extra error variance in the readability ground truth, e.g., due to the interest in an article or webpage aesthetics affecting participants' judgment of article readability.

3.3.1 Eye-Tracking. Eye movements were recorded using the EYELINK 1000 Plus system (SR Research, Mississauga, Canada) with a 1000 Hz sampling rate. Data were recorded from the right eye of the participants, except for 3 cases in which the left eye granted a better tracking. Participants were seated at 75 cm from the monitor and eye-tracking camera, with their head resting on a chin-rest.

We focused on three eye-tracking metrics relevant to webpage readability: number of eye fixations on text, number of fixations on non-text, and mean fixation duration. More fixations on text could indicate re-reading or being lost within text; more fixations on non-text could indicate distraction from the main article; and longer fixations could indicate a struggle to read and decode individual words. Fixations on the main article – each webpage featured an article – were counted as text fixations; fixations on the rest of webpage were counted as non-text fixations. The mean fixation durations for text and non-text correlated strongly and we did not differentiate between the two, combining them in a single variable.

3.3.2 Participants. We recruited 79 native Italian speakers (41 female) in four groups formed by two factors: age (adults, M = 23.08 years, SD = 3.72; and children, M = 11.51 years, SD = .91) and dyslexia (dyslexics and average readers). Children were enrolled in junior-high schools. All dyslexics had a certificate from local health services. If the certificate did not have scores for one of the reading tests used to monitor reading abilities, we administered such a test in the lab. The tests were also administered for average readers. All groups had 20 participants, except the dyslexic children group, which had one fewer participant. The purpose of the four groups was to analyze if different variables determined readability for dyslexics differently from average readers and if dyslexic adults learned to use mitigation strategies to cope with low readability. Such

analyses are not the subject of this paper and will be reported elsewhere. *3.3.3 Procedure.* After eye-tracker calibration, each participant saw five randomly selected webpages to read through. Many webpages did not fit on the screen and we sliced their screenshots in screen-sized pieces, which participants could switch among by pressing the 'up' and 'down' keys. After reading through a webpage, participants answered two questions about the webpage article and rated on a 1-7 Likert scale webpage reading difficulty, and their interest in the article topic and familiarity with the displayed website, as interest and familiarity may have influenced their reading patterns [44].

Table 1. Readability guidelines tested in the study.

ID	Guideline Text
G1	*Use left-justified text with the right edge being ragged, non-justified.*
G2	*Use an off-white color for your background, like light gray or tan; use dark gray for text instead of pure black.*
G3	*Use a plain, evenly spaced sans serif font such as Arial and Comic Sans.*
G4	*Avoid using italics in the main body of the text.*
G5	*Use bolding to highlight in order to emphasize keywords and concepts.*
G6	*Avoid underlining large blocks of text as it makes reading harder.*
G7	*Use font sizes larger than 12pt.*
G8	*Avoid capital letters, apart from the beginning of sentences, abbreviations, and where it is grammatically correct.*
G9	*If appropriate, use bullets or numbers rather than continuous prose.*
G10	*Use short, simple sentences in a direct style.*
G11	*Use active rather than passive voice.*
G12	*Avoid complex language and jargon.*
G13	*Consider using short paragraphs.*
G14	*Embed in Webpage texts the hyperlinks to the pages with the text-related concepts.*
G15	*Avoid images that are 'busy', cluttered, and include too much extra detail.*
G16	*Avoid placing images above text or text around images.*
G17	*Place the main point at the very top of page.*
G18	*Place important content in a single main column and avoid two-dimensional layouts.*
G19	*Ensure navigation menus use a text size that allows for comfortable reading.*
G20	*Avoid starting a new sentence at the end of a line.*
G21	*Keep the between-line spacing of 1.5 point.*
G22	*Use text and symbolism for navigational elements that are truly representative or a well-known concept e.g. a house for home.*
G23	*Provide clear intuitive labels for groups of links or menu sections.*
G24	*Put the main point of sentence or paragraph into the beginning of the sentence or paragraph.*
G25	*Avoid the fonts in which letters like b-d or p-q are perfectly mirrored letters.*
G26	*Ensure navigation menus group information by function.*
G27	*Ensure navigation menus differ visually from the main body of webpage.*
G28	*Limit the amount of content on a page to avoid scrolling.*
G29	*Use enough white space between webpage elements.*
G30	*Ensure high luminance contrast between text and background, with the luminance of one 7 times the luminance of the other. The rule doesn't apply to low-relevance, decorative visual elements.*
G31	*Ensure webpage elements (buttons, links, icons, etc.) that have the same function also have the same look.*
G32	*Keep the white space between paragraphs of at least 1.5 times the space between text lines.*
G33	*Avoid formatting texts in large-width columns, especially Asian logogram texts.*
G34	*Ensure Web pages have titles that describe their topic or purpose.*
G35	*Ensure headings and labels concisely describe the topic or purpose of page sections and elements.*
G36	*Use section headings to organize the content.*
G37	*User graphics that are relevant to the material and do not distract from the content.*
G38	*Use graphics, images, and pictures to break up large blocks of text.*
G39	*Place important information above the fold, so it is visible without scrolling a page down.*

Unfortunately, the scores of perceived reading difficulty were not recorded due to a programming error. The procedure was devised and controlled using E-Prime 2 (Psychology Software Tools, Pittsburgh, PA).

3.4 Manual Evaluation

Manual evaluation emulated a realistic guideline use in webpage evaluation by people. We call participants of this sub-study as experts – to differentiate them from the eye-tracking sub-study participants – even though their expertise levels varied.

3.4.1 Participants. We recruited 35 experts (27 male, m_{age} = 32.46 years, sd = 9.96, from 19 to 61), 13 of which were practitioners, eleven academics and eleven students. The sample included both beginners and true experts in Web design (m = 3.37, sd = 1.0, on a 1-5 scale), and dyslexia and reading (m = 3.37, sd = .94 on a 1-5 scale). The expertise scores were self-reported, but we did require experts describe their expertise or provide a link to their design portfolio to ground their expertise self-evaluation. Both design and dyslexia expertise scores were normally distributed.

3.4.2 Procedure. The evaluation was administered as an online study. After agreeing to the terms of their data use, experts rated five randomly chosen webpages on their compliance with each of 39 guidelines, using the 1-7 Likert-type items with anchors *not at all* and *completely*. If experts struggled to understand or apply a guideline – which happens in guideline-based evaluation [13,48] – they checked the '*I don't understand the guideline*' or '*The guideline doesn't apply*' checkbox. A full-length webpage screenshot and all guidelines were shown simultaneously on the screen, and time to rate was not limited (med = 1.64 min). As the experts familiarized themselves with the guidelines, time to rate decreased (1st webpage med = 15.35 min; 5th webpage med = 6.28 min). An average session lasted about an hour and experts noted the evaluation process as effortful. Participation was rewarded with six 25-Euro Amazon gift cards randomly allocated among the experts after the study.

3.5 Automatic Evaluation

We relied on the described measures of webpage readability features [30] and text complexity [11] to match some of 39 guidelines to automatic metrics (Table 2). To collect the input for the metrics, we developed an add-on for the Mozilla Firefox browser, which gave us access to the webpages exactly as they would be visible to the user. To estimate text-complexity metrics, we used the SUBTLEX-IT database (http://crr.ugent.be/subtlex-it/) containing Italian word frequencies and part-of-speech labels. Some metrics – although seemingly trivial to a human, such as a paragraph length – were impossible to estimate sufficiently well, in part due to a low adherence to good HTML coding practices in the sampled webpages.

Table 2. Automatic metrics for corresponding readability guidelines; Metrics were not developed for the guidelines G8, G11, G16-17, G19-20, G22-27, G31, G35, G37, and G39.

GuidID	Met-ricID	Metric Description
G1	A1	Ratio of left-aligned text to all text
G2	A2a	Euclidean distance in color between text and background, weighted by each text length, normalized, centered and squared
	A2b	Contour energy (the sharpness of a contour relative to the background, cf., [54,28,28])
G3	A3	Ratio of text in sans-serif fonts to all text
G4	A4	Ratio of italic text to all text
G5	A5	Ratio of bold text to all text
G6	A6	Ratio of underlined text to all text
G7	A7	Average font size of non-header webpage texts
G9	A9	Ratio of text in bullet-point lists to all text
G10	A10a	Average number of words per sentence
	A10b	Ratio of content words (nouns, adjectives, verbs, adverbs) to all words
	A10c	Ratio of conjunctions to all words
G12	A12a	Average word length in characters
	A12b	Average word logarithmic frequency
	A12c	Average content word logarithmic frequency
G13	A13a	Ratio of white space around text to text area
	A13b	Ratio of text area heights to page length
G14	A14	Ratio of text in hyperlinks to all text
G15	A15a	Sum of image file sizes in JPG, cf., [29,50]
	A15b	Average of image file sizes in JPG
G18	A18	Number of vertical alignment points for webpage content [29]
G21	A21	Averaged ratio of text line height to font size
G28	A28a	Page length
	A28b	Count of contour pixels, cf., [41]
	A28c	Amount of page text
G29	A29a	Number of large blocks (64- and 128-pixel sized squares) after a quadtree decomposition on webpage screenshot, cf., [27,36]
	A29b	Ratio of white space around webpage elements to element areas
	A29c	Metric of visual congestion: ratio of too-close-to-each-other contours to all contours [29]
G30	A30	Average text-background luminance contrast, as in the WCAG2.1 success criterion 1.4.6
G32	A32	Ratio of white space around texts to text area sizes
G33	A33	Average width of text columns, measured in the number of characters
G34	A34	Ratio of header text to all text
G36	A36	Ratio of header text to regular text (not header or control elements)
G38	A38	Amount of text per picture

4 RESULTS

We reviewed the three pieces of data separately and then compared them against each other.

4.1 Ground Truth

Table 3 shows mean fixation durations and fixation counts across the four participant groups, suggesting that dyslexics and children used more and longer fixations than average readers and adults. After filtering out several outlier data points and log-normalizing the three eye-tracking variables – they were strongly positively skewed – we tested the significance of effect of two factors (dyslexics vs average readers, and adults vs children) in three ANOVAs, Table 4. As expected, having dyslexia significantly increased fixation duration and fixation counts. Being a child increased fixation durations and number of fixations on text, but not on non-text. Having dyslexia and being a child further increased fixation duration. The main effect of dyslexia on all three variables supports the use of these variables as indicators of readability.

Familiarity scores were strongly positively skewed with few webpages receiving a 2+ score (on a 7-point scale; m_{adult} = 1.51, SD = 1.05; m_{child} = 1.26, SD = 1.02), as we sought to minimize potential familiarity effects. Interest scores had a close-to-normal distribution, with their mean being close to the scale center for both adults (m = 4.38, SD = 1.39) and children (m = 4.05, SD = 1.85). We omit interest and familiarity from further analyses as neither of them influenced any of the eye-tracking indices: they failed to improve the fit of baseline mixed models (participantID and webpageID as random effects; eye-tracking indices as outputs; participantGroup as fixed effects). The absence of interest/familiarity effect on reading might have stemmed from the two article-related post-reading questions that ensured an article was read and not skimmed through even if interest was low.

4.2 Expert Evaluation

Few experts used the "*I don't understand the guideline*" option, and only the G33 guideline appeared problematic, with seven experts highlighting it, which still was not problematic enough to exclude it from further analyses. Experts did use the "*The guideline doesn't apply to this webpage*" option, with G9 and G38 not applying in ~22% of cases, G22 and G37 in ~15% of cases, and G15 in 11% of cases. This still left enough data to test these guidelines and we kept them for the analyses. A review of experts' scores revealed three guidelines – G1, G4, G6 – that most webpages fully complied with (mean > 6 on the 1-7 scale), which implied that our dataset did not contain sufficient variance to test these guidelines, and we excluded them from further analyses.

Table 3. Means (SDs) of eye-tracking variables for average-reader adults (AA), average-reader children (AC), dyslexic adults (DA), and dyslexic children (DC)

	Fixation duration	Text fixation counts	Non-text fixation counts
AA	248.33 (29.05)	389.47 (203.58)	74.54 (39.15)
DA	282.01 (43.57)	514.44 (340.22)	111.17 (88.76)
AC	300.65 (54.86)	444.73 (276.31)	96.44 (77.59)
DC	424.31 (110.02)	721.65 (545.37)	125.15 (163.86)

Table 4. Three ANOVAs showing the effect of dyslexia and age on eye-tracking variables, and corresponding partial eta-squared, all df = (1,396).

Factors / Ind. vars	Dyslexia		Age		Dyslexia*Age	
	F-value	eta^2	F-value	eta^2	F-value	eta^2
Mean fixation duration	143.02***	.28	252.96***	.40	28.58***	.07
# of fixation on text	25.51***	.06	7.91**	.02	2.04	.01
# of fixation on non-text	8.24**	.02	.40	.00	1.38	.00

*** p < .001; ** p < .01

We presumed the mean of several expert scores per webpage per guideline to be closer to the true score for that webpage and guideline than an individual expert score, and estimated expert's ability to adhere to the guidelines as the average of difference between mean and expert's scores. The ability to adhere negatively correlated with the self-reported design experience ($r_s(33) = -.40$, $p < .05$), but not with dyslexia experience ($r_s(33) = -.28$, $p = .10$). However, dyslexia experts tended to use more extreme scores, with the mean expert-score distance from the scale center correlating with self-reported dyslexia experience ($r_s(33) = .39$, $p < .05$), but not design experience ($r_s(33) = .23$, $p = .18$). Being more experienced did not lead to being more critical of webpages: mean per-expert scores did not correlate with design ($r_s(33) = -.19$, $p = .27$) or dyslexia experience ($r_s(33) = -.16$, $p = .37$). Being more experienced also did not lead to dedicating less time to evaluation: experts' time-per-webpage did not correlate with their design ($r_s(33) = .08$, $p = .63$) and dyslexia experience ($r_s(33) = .02$, $p = .90$).

4.3 Automatic Evaluation

After scaling and filtering out several extreme outlier data points from the computed data, we reviewed the histograms of computed variables, which showed several metrics to have only limited variance. For example, the webpages in our sample used little of underlined text and a lot of sans-serif fonts for all of their texts (Figure 2). However, we retained all metrics for further analyses. A review of strong cross-correlations ($r > .70$, $p < .001$) suggested that several metrics could have been redundant. For example, average word length (A12a) inversely

correlated with average word frequency (A12b, r(115) = -.78) and average content-word frequency (A12c, r(115) = -.80), which could be expected since longer words tend to be less frequent. Also expectedly, the amount of contours (A28c) correlated with webpage length (A28a, r(115) = .76) and amount of text (A28c, r(115) = .82), and correlated inversely with the amount of white space (A29a, r(115) = -.80). The only seemingly spurious correlation was between the amount of text per graphic (A38) and amount of text in hyperlinks (A14), r(115) = -.9. The absence of other strong spurious correlations suggested that the automatic metrics described different, unrelated aspects of webpages, which further suggested the usefulness of metrics in webpage description.

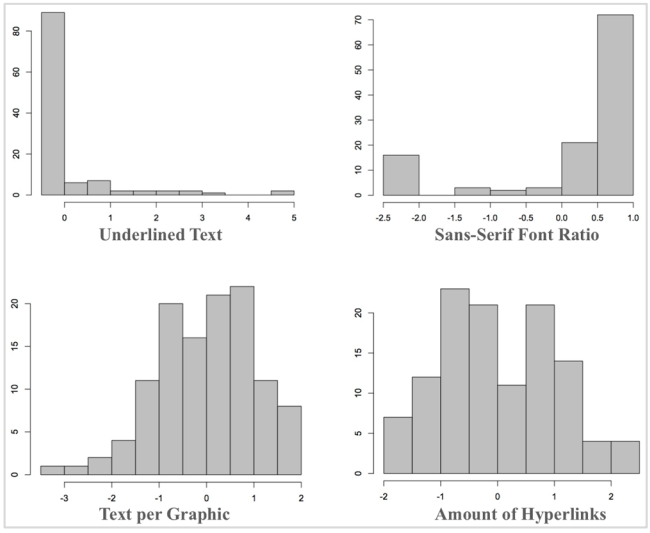

Figure 2. Histograms of metrics A6 (Underlined text), A3 (Sans-serif fonts), A14 (Hyperlinks), and A38 (Text per graphic). Metric scores (horizontal axes) were centered and scaled.

4.4 Expert Evaluation and Ground Truth

We relied on the linear mixed models as the main tool to explore the relationships between guidelines and readability variables [3]. To test each guideline, we put its scores aggregated across experts in a model and compared model fit relative to a baseline model without the guideline scores. A significant improvement in model performance – measured as a decrease in the Akaike Information Criterion (AIC), which penalizes having more predictors while accounting for model fit – indicated that a guideline was indeed related to a readability aspect. The three eye-tracking variables were independent variables, and participant ID and webpage ID were random-effect variables. The baseline models contained both factors (age and dyslexia) and trial index (the order of webpage evaluation); an actual model also contained the scores of one guideline and its interaction product with the two factors. Table 5 summarizes significant relationships between guidelines and eye-tracking variables. The significance was measured with an F-test comparing a model with a guideline against a baseline model.

Table 5. Series of mixed models (one model per row) show guidelines linked to readability. The link was determined as a decrease in AIC relative to the baseline models without a corresponding guideline. An F-test generated significance levels, comparing each model against its baseline model. Betas show the magnitude of each guideline effect on reading for average-reader adults (AA), average-reader children (AC), dyslexic adults (DA), and dyslexic children (DC).

Independent variable	Guid. ID	Betas AA	AC	DA	DC	Change in AIC
Mean fixation duration	G7	-.11	-.16	-.12	-.24	-15.68***
	G28	.06	.12	.11	.22	-7.43**
	G22	-.07	-.12	-.05	-.14	-5.74**
	G35	-.00	.03	.05	.08	-3.53*
	G30	-.04	-.00	-.04	-.04	-3.47*
Number of non-text fixations[3]	G29	-.21	-.40	-.25	-.62	-17.32***
	G24	.08	.07	.15	.47	-3.07*
	G11	.14	-.14	.06	-.29	-3.61*
	G16	-.11	-.14	-.07	.11	-4.37*
	G28	-.03	-.35	-.01	-.41	-4.54*
	G15	-.04	.02	.09	.59	-6.29**
	G21	-.11	-.33	-.18	-.62	-6.61**
Number of text fixations	G28	-.34	-.34	-.39	-.34	-8.26**
	G38	-.31	-.28	-.32	-.41	-5.90**
	G12	-.36	-.31	-.34	-.19	-5.44**
	G20	-.22	-.32	-.26	-.41	-3.47*
	G33	-.22	-.22	-.28	-.41	-2.63*
	G13	-.28	-.29	-.29	-.26	-1.91*
	G24	-.10	-.11	-.12	-.01	-1.54*

*** p < .001; ** p < .01; * p < .05

4.5 Automatic Evaluation and Ground Truth

The automatic variables were entered in the series of linear mixed models the same way as experts' scores above, except for the models of number of text fixations, which contained the amount of webpage text (A28c) already in the baseline model – we found that the number of text fixations strongly correlated with the amount of text, which would be expected, and decided to exclude its influence from metric evaluation by including it in the baseline. Table 6 lists successful metrics that produced scores linked to an improvement in one of the eye-tracking variables.

4.6 Expert and Automatic Evaluation and Ground Truth

For several guidelines, both expert and automatic evaluation revealed a link between the guidelines and readability variables.

[3] Webpage length was included as an extra fixed effect already in the baseline model.

Table 6. Series of mixed models (one model per row) show automatic metrics related to readability. The link was determined as a drop in AIC relative to the baseline models without a metric. An F-test generated significance levels, comparing each model against its baseline model. Betas show the magnitude of each metric relationship to reading for average-reader adults (AA), average-reader children (AC), dyslexic adults (DA), and dyslexic children (DC).

Independent variable	Metric ID	Betas				Change in AIC
		AA	AC	DA	DC	
Mean fixation duration	A7	-.10*	-.18	-.10	-.22	-22.95***
	A2b	.10	.15	.09	.13	-16.68***
	A28b	-.09*	-.12	-.10	-.18	-9.78**
	A10c	.03	.05	-.02	-.10	-9.41**
	A28a	-.07	-.09	-.10	-.13	-8.23**
	A29a	.08	.13	.10	.19	-5.43**
	A13b	-.06	-.05	-.06	-.10	-4.77*
	A13a	-.06	-.10	-.08	-.13	-1.95*
Number of fixations on non-text[4]	A28c	.25	.45	.29	.94	-24.55***
	A29c	.10	.39	.15	.81	-11.11**
	A15a	.12	-.13	.06	-.66	-9.09**
	A13b	.19	.27	.15	.45	-7.93**
	A28a	.28	.42	.19	.51	-7.23**
	A18	.05	-.23	-.03	-.74	-6.37**
	A34	-.05	-.25	-.14	-.72	-5.23*
	A21	-.00	-.23	-.06	-.49	-3.20*
	A33	-.11	-.16	-.12	-.43	-1.93*
Number of fixations on texts[5]	A38	.22	.27	.28	.56	-2.98***
	A14	-.20	-.24	-.27	-.52	-15.78***
	A18	-.18	-.12	-.26	-.28	-7.95**
	A33	-.067	-.106	-.14	-.38	-6.05**
	A34	-.08	-.12	-.15	-.37	-4.33*
	A29c	.15	.11	.20	.26	-2.58*
	A1	-.20	-.14	-.22	-.12	-2.37*

*** p < .001; ** p < .01; * p < .05

To test if expert evaluation could be substituted with automatic evaluation or if expert evaluation captured some readability variance that automatic evaluation could not, we explored such guidelines in a series of mixed linear models (Table 7).

We entered expert scores and algorithm scores in models – individually and both together, which gave us three models per guideline – and checked if model performance improved. If adding algorithm scores as a predictor in addition to expert scores did improve the performance, we considered expert evaluation to be replaceable with automatic evaluation. If both algorithm and human scores contributed significantly to full model performance, we considered automatic and manual evaluation to be complementary. Table 8 summarizes the results of such expert-algorithm comparison and the other guidelines, for which a comparison was not possible or needed.

Table 7. Comparison of expert scores VS algorithm scores: if adding algorithm scores on top of expert scores improved model fit (significant values in column AIC Alg.) while adding expert scores on top of algorithm scores did not improve model fit (non-significant values in column AIC Expert), expert evaluation could be replaced with algorithms ("Repl" and "ShouldRepl" Recommendation). Otherwise, algorithms complement ("Compl") or cannot replace experts ("NonRepl").

GuidID /AlgID	AIC Alg	AIC Expert	Indep-t variable	Recommen-dation
G7/A7	-15.51***	-6.31**	Mean fixation duration	Compl
G28/A28a	-3.50*	1.181		ShouldRepl
G28/A28b	-4.74**	1.39		ShouldRepl
A67	-7.58	-5.71**	# of non-text fixations	Compl
G28/A28c	-24.14***	-7.70**		Compl
G15/A15a	-2.80*	-0.53		ShouldRepl
A29/A29c	-4.71*	-12.81**		Compl
A21/A21	1.582	-3.24*		NonRepl
G28/A28c	-35.75***	1.71	# of on-texts fixations	ShouldRepl
A38/G38	-4.84*	-2.42*		Compl
A33/G33	0.82	2.65		Repl
G13/A13b	-4.41*	-7.79**		Compl

*** p < .001; ** p < .01; * p < .05

Finally, we compared experts' ability to consistently apply the guidelines that algorithms performed better on or as good as evaluators (Table 8, guidelines marked as 'Alg' and 'Both') against the guidelines that algorithms could not use as well as human evaluators (Table 8, guidelines marked as 'Exp'). A paired t-test showed that the difference between average and individual expert scores was higher for the algorithm-better guidelines (m = 1.05) than for the expert-better guidelines (m = .96; t(34) = 2.51, p < .05).

5 DISCUSSION

The traditional guideline-based webpage evaluation has several issues, e.g., past research asserted that experts disliked and resisted using long checklists [8]. Our evaluators also regularly mentioned in the free-form optional feedback after the evaluation that it was tiresome and repetitive, even though we asked them to evaluate only five webpages each. Our analysis suggests that computation and algorithms could indeed mitigate such issues. However, completely replacing humans with algorithms does not appear beneficial or easily achievable, as Table 8 and related work [51] show, and we believe the future webpage evaluation will combine both human-based and automatic methods.

[4] Webpage length was included as an extra fixed effect already in the baseline model.
[5] Amount of text was included as an extra fixed effect already in the baseline model.

Table 8. Comparison of performances of expert-based and automatic evaluation; 'x' means a guideline was supported with a method; '-' means a guideline was not tested with a method. The guidelines in gray were not supported as relevant for Web readability.

Guid ID	Guideline Label	Ex-perts	Algo-rithms	Win ner
G11	Active voice writing	x	-	Exp
G12	Complex words	x		Exp
G16	No text around images	x	-	Exp
G20	New sentence at new line	x	-	Exp
G22	Symbolic navigation items	x	-	Exp
G24	Start sentences from the main point	x	-	Exp
G30	Luminance contrast	x		Exp
G35	Meaningful titles	x	-	Exp
G7	Font size	x	x	Both
G13	Short paragraphs	x	x	Both
G21	Between-line space	x	x	Both
G29	White space	x	x	Both
G1	Left-aligned text	-	x	Alg
G2	Moderate text-background contrast		x	Alg
G10	Simple sentences		x	Alg
G14	Hyperlinks		x	Alg
G15	Simpler images	x	x	Alg
G18	Single column		x	Alg
G28	Limit scrolling	x	x	Alg
G33	Narrow text columns	x	x	Alg
G34	Have titles		x	Alg
G38	Not only text	x	x	Alg
G3	Sans serif font			
G4	Italic text	-		
G5	Bold text			
G6	Underlined text	-		
G8	Capital letters		-	
G9	Bullet lists			
G17	Main point at the top		-	
G19	Text in menus		-	
G23	Labels for groups		-	
G25	Symmetrical fonts		-	
G26	Menu item grouping		-	
G27	Menus differ from the webpage body		-	
G31	Same elements, same look		-	
G32	Between-paragraph spacing			
G36	Titles and other text			
G37	Relevant graphics		-	
G39	Important info before scrolling		-	

5.1 Guideline Categorization

Overall, a review of Table 8 suggests that the guidelines could be categorized in three groups. First, some guidelines should be automated and humans may re-check automatic evaluation if an algorithm flags a guideline as violated for a webpage. This group includes guidelines about the low-level text-legibility or text-formatting aspects of readability, which human evaluators were not only worse at utilizing than algorithms, but also struggled to apply these guidelines consistently. Such aspects include, e.g., the amount of white space (G29), left aligned text (G1), text-background contrast (G2), sentence (G10) and paragraph length (G13), complexity of images (G15), width of text column (G33), font size (G7) and other similar aspects.

Second, some guidelines could be automated or partially automated, but humans need to re-check the results of automatic evaluation. This group contains some of the text-complexity guidelines that algorithms could measure, but only indirectly, or text-formatting guidelines with a clause defining their scope, which may be problematic for algorithms to handle. Such guidelines include the luminance contrast guideline (G30), which has a clause defining its scope to non-decorative elements only, the guideline about hyperlinks (G14), which specifies that hyperlinked pages should be related to the main content, the single-column guideline (G18), which specifies that only important, main content should be in the single column, and the have-headings guideline (G34), which the algorithm evaluates indirectly – by measuring the amount of header text and presuming that such texts is relevant in the context of a webpage.

Finally, some guidelines are difficult to automate, and human evaluators would need to use them without the help of algorithms. This final group contains the guidelines that require understanding and interpreting the topic of webpage, which algorithms would struggle to accomplish. Such guidelines include the use of jargon (G14), symbolical icons for navigational elements (G22), main point of sentences being put upfront (G24), and labels and menu items being concise (G35). These guidelines would require interpreting what qualifies as jargon for a particular audience, if an icon meaningfully describes an item, what the main point of sentences and paragraphs is, and if menu items are actually linked to the webpage topic.

5.2 Expertise in Evaluation

The results suggested that less-experienced evaluators may benefit from relying on a guideline checklist more than expert evaluators: we observed a negative correlation between expertise and one evaluator scores being closer to the true scores. This correlation could not be explained by the less-experienced evaluators spending more time and putting more effort in evaluation than experts, since time-per-page and expertise did not correlate. It also could not be explained by our sample containing more novices than experts, and thus, novices impacting the true scores – estimated as the average – more than experts, since the distribution of expertise scores was close to symmetrical and mean of expertise was above the scale average. Finally, the correlation also could not be explained by experts being

more critical of problematic aspects of webpage, since the propensity to set more extreme ratings did not correlate with expertise. We may speculate that the less-experienced evaluators could better adhere to the guidelines than experts, and additional experience may have interfered with applying a guideline as written, without interpretation.

5.3 Readability Guidelines

Readability guidelines could instruct designers in creating better webpages, but they still require additional clarification and validation. This study explored 39 guidelines and almost all of them were clear to the evaluators. However, the study found no evidence of 17 guidelines being connected to readability (Table 8, highlighted in gray). Some of these 17 should indeed be discarded as useless, but we expect that other guideline could receive empirical support in a different context or with a different, larger sample of webpages. For example, the guidelines about navigational menus (G26 and G27) might show their influence if participants need to not only read through an article, but also browse and search for information; and the guidelines about sans-serif fonts (G3) and underlined text (G6) could be supported if the sample of webpages contained more variance in the features related to them.

Out of the 22 guidelines that were connected to readability, almost all performed as expected, with higher compliance with a guideline resulting in a decrease in reading effort (Table 5). Only few guidelines performed unexpectedly. For example, limiting webpage content (G28) increased average fixation duration, which was unexpected, but also offset by a large decrease in the number of on-text fixations, which was expected. Putting the main paragraph point upfront (G24) increased the number of non-text fixations, particularly for dyslexic children. Finally, avoiding dense images also increased the number of non-text fixations (G15) for dyslexic children, which may have stemmed from the children liking and looking at such images more since simplicity and aesthetics are connected [50,29].

Improved Web readability could help all groups of users, but vulnerable groups, such as dyslexics and children, could benefit particularly. Crucially, the improvement for such groups does not lead to reduced readability for the average readers (cf., [30]), as the direction of effect for all guideline scores was the same across groups (Table 5), with a possible exception of the use of active voice sentences (G11), which decreased the number of non-text fixations for dyslexics, but increased for the average readers.

6 FUTURE WORK

Future research will rely on the described three-group categorization of guidelines to develop systems and protocols for semi-automated webpage evaluation. Algorithms will automatically highlight the issues with the low-level easily countable aspects of webpages, which will lower the burden on evaluators and allow them to focus on the guidelines that require meaning and context understanding and interpretation. Future research will further validate such systems and protocols, and test if novice evaluators can learn good design practices from them, and whether expert evaluators resist and disagree with automatic evaluation and prefer to rely on their experience.

Future research will also address the limitations of this study, including enlarging the samples of users, involving other user groups besides dyslexics and children, testing the guidelines with functional websites instead of webpages, and using both reading and non-reading tasks.

7 CONCLUSION

This paper investigated 39 Web readability guidelines and their use in expert-based and automatic evaluation. Algorithms appeared helpful in highlighting several problematic webpage aspects, including the low-level visual aspects that human experts struggled to rate consistently. However, the experts would still need to manually apply a subset of guidelines that the algorithms could not successfully utilize, including the guidelines based on content understanding and interpretation.

ACKNOWLEDGMENTS

The research was funded by the EU Horizon 2020 programme (grant agreement No 643644 for *ACANTO: A CyberphysicAL social NeTwOrk using robot friends*) and by Cassa di Risparmio di Trento e Rovereto (*LEILA - Leggere i libri digitali appropriati*).

REFERENCES

[1] Arditi, A. and Cho, J. Letter case and text legibility in normal and low vision. *Vision research*, 47, 19 (2007), 2499-2505.

[2] Aziz, F. A. and Husni, H. Interaction Design for Dyslexic Children Reading Application: A Guideline. In *Knowledge Management International Conference (KMICe)* (2012), 682-686.

[3] Bates, D., Maechler, M., Bolker, B. & Walker, S. Fitting linear mixed-e ects models using lme4. *Journal of Statistical Software*, 67 (2014), 1–48.

[4] Brown, C. M. *Human-computer interface design guidelines*. Intellect Books, 1999.

[5] Cassim, R., Talcott, J. B., and Moores, E. Adults with dyslexia demonstrate large effects of crowding and detrimental effects of distractors in a visual tilt discrimination task. *PloS one*, 9, 9 (2014).

[6] Chorianopoulos, K. User interface design principles for interactive television applications. *Intl. Journal of Human–Computer Interaction*, 24, 6 (2008), 556-573.

[7] Dingli, A., & Cassar, S. An intelligent framework for website usability. *Advances in Human-Computer Interaction*, 5 (2014), 1-13.

[8] Dumas, J. S. and Salzman, M. C. Usability assessment methods. *Reviews of human factors and ergonomics*, 2, 1 (2006), 109-140.

[9] Flesch, R. A new readability yardstick. *Journal of applied psychology*, 32, 3, 1948.

[10] Friedman, M. G. and Bryen, D. N. Web accessibility design recommendations for people with cognitive disabilities. *Technology and Disability*, 19, 4 (2007), 205-212.

[11] Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36, 2 (2004), 193-202.

[12] Grigera, J., Garrido, A., Rivero, J. M., & Rossi, G. Automatic detection of usability smells in web applications. *International Journal of Human-Computer Studies*, 97 (2017), 129-148.

[13] Henninger, S. A methodology and tools for applying context-specific usability guidelines to interface design. *Interacting with computers*, 12, 3 (2000), 225-243.

[14] Henry, S. L. Developing text customisation functionality requirements of PDF reader and other user agents. In *International Conference on Computers for Handicapped Persons* (2012), Springer, 602-609.

[15] Höök, K. and Löwgren, J. Strong concepts: Intermediate-level knowledge in interaction design research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19, 3 (2012), 1–18.

[16] Ivory, M. Y. and Hearst, M. A. Improving web site design. *Internet Computing*, 6, 2 (2002), 56-63.

[17] Jeffries, R. and Desurvire, H. Usability testing vs. heuristic evaluation: was there a contest? *ACM SIGCHI Bulleti*, 24, 4 (1992), 39-41.

[18] Ji, Y. G., Park, J. H., Lee, C., & Yun, M. H. A usability checklist for the usability evaluation of mobile phone user interface. *International journal of human-computer interaction*, 20, 3 (2006), 207-231.

[19] Jo, J., Kim, B., & Seo, J. EyeBookmark: Assisting recovery from interruption during reading. In *the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), ACM, 2963-2966.

[20] Johannessen, G. H. J., & Hornbæk, K. Must evaluation methods be about usability? Devising and assessing the utility inspection method. *Behaviour & Information Technology*, 33, 2 (2014), 195-206.

[21] Kember, P. and Varley, D. The legibility and readability of a visual display unit at threshold. *Ergonomics*, 30, 6 (1987), 925-931.

[22] Kirkwood, T. B., Bond, J., May, C., McKeith, I., and Teh, M. M. *Foresight mental capital and wellbeing project*. The Government Office for Science, 2008.

[23] Kurniawan, S. and Zaphiris, P. Research-Derived Web Design Guidelines for Older People. In *7th ACM Conference on Computers and Assessibility* (2005), ACM, 129-135.

[24] Lanzilotti, R., Ardito, C., Costabile, M. F., and De Angeli, A. Do patterns help novice evaluators? A comparative study. *International journal of human-computer studies*, 69, 1 (2011), 52-69.

[25] Large, A., Beheshti, J., Nesset, V., & Bowler, L. Web portal design guidelines as identified by children through the processes of design and evaluation. *Proceedings of the American Society for information Science and Technology*, 43, 1 (2006), 1-23.

[26] Mi, N., Cavuoto, L. A., Benson, K., Smith-Jackson, T., & Nussbaum, M. A. A heuristic checklist for an accessible smartphone interface design. *Universal access in the information society*, 13, 4 (2014), 351-365.

[27] Miniukovich, A., Sulpizio, S., & De Angeli, A. Visual complexity of graphical user interfaces. In *the 2018 International Conference on Advanced Visual Interfaces* (2018), ACM, 1-9.

[28] Miniukovich, A. and De Angeli, A. Pick Me! Getting Noticed on Google Play. In *CHI'16* (2016), ACM.

[29] Miniukovich, A. and De Angeli, A. Quantification of Interface Visual Complexity. In *the 2014 International Working Conference on Advanced Visual Interfaces* (Como 2014a), ACM, 153-160.

[30] Miniukovich, A., De Angeli, A., Sulpizio, S., and Venuti, P. Design Guidelines for Web Readability. In *the 2017 Conference on Designing Interactive Systems* (2017), ACM, 285-296.

[31] Montero, F., Vanderdonckt, J., & Lozano, M. Quality models for automated evaluation of web sites usability and accessibility. In *International COST294 workshop on User Interface Quality Models* (2005).

[32] Morrison, R.E. and Inhoff, A.-W. Visual factors and eye movements in reading. *Visible Language*, 15, 2 (1981), 129-146.

[33] Nebeling, M., Speicher, M., & Norrie, M. W3touch: metrics-based web page adaptation for touch. In *the SIGCHI Conference on Human Factors in Computing Systems* (2013), ACM, 2311-2320.

[34] Petrie, H. and Kheir, O. The relationship between accessibility and usability of websites. In *SIGCHI conference on Human factors in computing systems* (2007), ACM, 397-406.

[35] Rayner, K. Eye movements and the perceptual span in beginning and skilled readers. *Journal of Experimental Child Psychology*, 41, 2, 211-236.

[36] Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., and Liu, J., Gajos, K. Z. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *CHI* (Paris 2013), ACM, 2049-2058.

[37] Rello, L. and Baeza-Yates, R. Evaluation of Dyswebxia: A reading app designed for people with dyslexia. In *the 11th Web for All Conference* (2014), ACM.

[38] Rello, L., Kanvinde, G., and Baeza-Yates, R. Layout guidelines for web text and a web service to improve accessibility for dyslexics. In *International cross-disciplinary conference on web accessibility* (2012), ACM.

[39] Rello, L., Pielot, M., and Marcos, M. C. Make it big! The effect of font size and line spacing on online readability. In *CHI'16* (2016), ACM, 3637-3648.

[40] Rosenbaum, S. The future of usability evaluation: increasing impact on value. In Law, E. et al., eds., *Maturing usability*. Springer, 2008.

[41] Rosenholtz, R., Li, Y., and Nakano, L. Measuring visual clutter. *Journal of vision*, 7, 2 (2007), 1-22.

[42] Santana, V. F., de Oliveira, R., Almeida, L. D. A., and Baranauskas, M. C. C. Web accessibility and people with dyslexia: a survey on techniques and guidelines. In *International Cross-Disciplinary Conference on Web Accessibility* (2012), ACM.

[43] Santana, V. F., Oliveira, R., Almeida, L. D. A., and Ito, M. Firefixia: An accessibility web browser customization toolbar for people with dyslexia. In *10th International Cross-Disciplinary Conference on Web Accessibility* (2013), ACM.

[44] Schiefele, U., & Krapp, A. Topic interest and free recall of expository text. *Learning and individual differences*, 8, 2 (1996), 141-160.

[45] Slattery, T. J. Eye movements: From psycholinguistics to font design. In Dyson, M. C. and Yuen, C. Y., eds., *Digital Fonts and Reading*. World Scientific, 2016.

[46] Sperling, A. J., Lu, Z. L., Manis, F. R., and Seidenberg, M. S. Motion-perception deficits and reading impairment: it's the noise, not the motion. *Psychological Science*, 17, 12 (2006), 1047-1053.

[47] Sutcliffe, A. G. and Carroll, J. M. Designing claims for reuse in interactive systems design. *International Journal of Human-Computer Studies*, 50, 3 (1999), 213-241.

[48] Tetzlaff, L., & Schwartz, D. R. The use of guidelines in interface design. In *SIGCHI Conference on Human Factors in Computing Systems* (1991), ACM, 329-333.

[49] Theofanos, M. F. and Redish, J. Guidelines for accessible and usable web sites: Observing users who work with screen readers. *Interactions*, 10, 6 (2003), 38-51.

[50] Tuch, A. N., Presslaber, E. E., Stocklin, M., Opwis, K., and Bargas-Avila, J. A. The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. *International Journal of Human-Computer Studies*, 70 (2012), 794-811.

[51] Vigo, M., Brown, J., & Conway, V. Benchmarking web accessibility evaluation tools: measuring the harm of sole reliance on automated tests. In *the 10th International Cross-Disciplinary Conference on Web Accessibility* (2013), ACM.

[52] Yáñez Gómez, R., Cascado Caballero, D., & Sevillano, J. L. Heuristic evaluation on mobile interfaces: A new checklist. *The Scientific World Journal* (2014).

[53] Yu, C. H. and Miller, R. C. Enhancing web page readability for non-native readers. In *the sIGCHI conference on human factors in computing systems* (2010), 2523-2532.

[54] Yu, H. and Winkler, S. Image complexity and spatial information. In *Quality of Multimedia Experience (QoMEX)* (2013), IEEE, 12-17.

[55] Zajicek, M. Successful and available: interface design exemplars for older users. *Interacting with computers*, 16, 3 (2994), 411-430.

[56] Zorzi, M., Barbiero, C., Facoetti, A. et al. Extra-large letter spacing improves reading in dyslexia. *Proceedings of the National Academy of Sciences*, 109, 28 (2012), 11455-11459.