How bold can we be? The impact of adjusting font grade on readability in light and dark polarities

Hilary Palmén
aitch@google.com
Google
USA

Michael Gilbert
mdgilbert@google.com
Google
USA

David Crossland
dcrossland@google.com
Google
USA

ABSTRACT

Variable font file technology enables adjusting fonts on scaled axes that can include weight, and grade. While making text bold increases the character width, grade achieves boldness without increasing character width or causing text reflow. Through two studies with a total of 459 participants, we examined the effect of varying grade levels on both glancing and paragraph reading tasks in light and dark modes. We show that dark text on a light background (Light Mode) is read reliably faster than its polar opposite (Dark Mode). We found an effect of mode for both glance and paragraph reading and an effect of grade for LM with heavier, increased grade levels. Paragraph readers are not choosing, or preferring, LM over DM despite fluency benefits and reported visual clarity. Software designers can vary grade across the tested font formats to influence design aesthetics and user preferences without worrying about reducing reading fluency.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **User studies**; *HCI design and evaluation methods.*

KEYWORDS

Reading, Readability, Typography, Typeface, Type style, Font, Variable font, Polarity, Text, Weight, Grade, Mode, Polarity. Reading fluency

ACM Reference Format:

Hilary Palmén, Michael Gilbert, and David Crossland. 2023. How bold can we be? The impact of adjusting font grade on readability in light and dark polarities. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3544548.3581552

1 INTRODUCTION

Almost every website, app, and device we use provides text to help us understand and effectively engage with the value it provides. Texts are critical information sources. Adults are frequently reading texts, whether reading signs to choose where to go next, glancing at a clock to check the time, scanning a menu to find the most appropriate choice, or immersing ourselves in a novel for pleasure. Reading is a highly diverse activity, and we read for different reasons in different ways to achieve different goals and experiences. People employ different strategies and tactics to read in different ways, for different purposes, frequently reading on phones and computer screens.

Typography is the art and process of arranging text to make written language legible, readable and generate appropriate emotional engagement with the content. Typography guides and informs readers, optimizing readability and accessibility, to ensure an excellent user experience. Typography is about how typefaces are used. A typeface is a family of related fonts. Roboto is a typeface that contains 12 font files, including fonts called "thin 100", "regular 400" where the number is a specification of the 'weight' of the font in the file. Roboto Flex is a single Variable font that can produce Roboto Thin 100 and Roboto regular 400, and all the weights in between on a weight axis. A variable font is a new font file format that provides access to multiple style variations of a typeface within a single font file. In September 2016, variable fonts were added to the OpenType 1.8 specification file format. The Roboto Flex variable font contains 12 variable axes. One file with all the necessary typeface styles, and more, is significantly smaller in size than classic families with multiple files. This shortens the font loading time in web environments, providing faster service with more style diversity. This provides a powerful design tool for detailed manipulation of the font along each of its variable axes, and by combining their effects. A designer could use a variable font to vary the weight, width, style, and optical size, each based on different contextual rules, for one piece of text. For example, a document could be programmed to increase weight when the cursor is nearby, making the word being pointed-at temporarily more bold. Variable fonts can be dynamic.

This study investigates the impact that the grade variable font characteristic has on readability, considering both light mode and dark mode. Figure 1 shows how an increase in the grade axis from 0-100 can produce an effect similar to the increase in the weight axis from 400-500. The most noticeable difference for a designer using text is that the adjustment of grade does not introduce a text reflow when compared to the adjustment of weight. As a designer increases the weight axis, either the texts gets longer and reflows over lines or the letter-spacing is reduced.

Variable font technology not only provides new design possibilities, it also provides a powerful research tool for discovering the impact of nuanced structural differences in font designs. Typically, the fonts used in research to understand the structural factors that promote reading fluency are not systematically varied on those structural factors. Researchers make manipulations using a single font as a representative of a common typographical structural quality such as serif versus san-serif. The fonts are not systematically

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581552

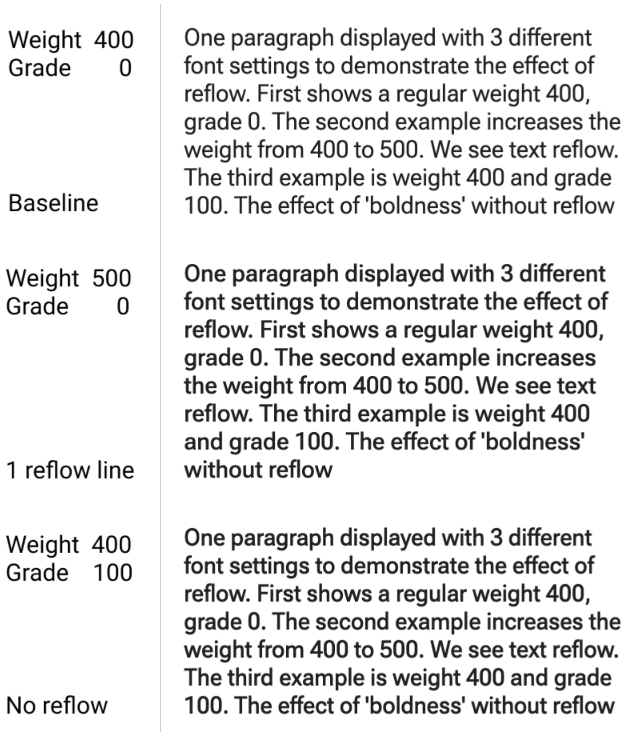

Figure 1: Demonstration of grade increasing grade without causing text to reflow across lines that require more screen real estate

varied, in ways that enable experimental manipulations of independent variables. More recently, researchers [28] have manipulated existing font structures as a way to normalize the characteristics, to enable a true comparison of the variables they are considering. This manipulation enables controlled variables but means software creators cannot confidently choose to use fonts cited in the research based on the outcomes of these studies. The control of experimental variables that enables direct generalisability to fonts that can be used in the development of products has been an ongoing challenge. Variable fonts provide a way to simultaneously bridge this gap by both facilitating experimental research manipulation of a font that is available to product designers in the forms used during the research.

Typography of print media in Europe originated as dark inks on lighter paper surfaces, a light mode without backlighting. The transition to digital media introduced the possibility of a Dark Mode (DM): Light text on a dark background. The polarities of LM and DM provide users with a personal choice that can make their reading more comfortable with preferred overall luminance levels, increasing their comfort and reading fluency. There is a difference between the experience of modes based on the reader's characteristics (age, photo-sensitivity, dyslexic thinking), general screen illumination, and typography through the possibility of halation [25]: the seeping of light across the edges of the font. Increasing the weight of a font is a potential solution that is often recommended within the design community. The pattern of interaction of mode (luminance), halation, and weight on reading fluency is not known. This knowledge can be used to improve the readability of digital products and services.

The research design will initially identify if there is a significant difference in the readability of text varying polarity (LM vs DM) or context (glanceable vs longer form reading). Where a difference is found, the results can provide guidance where the variable font axis could facilitate equivalent reading experiences across polarities, within contexts. This provides data that can act as a step towards guidelines to support designers crafting reading experiences that are sensitive to the nature of the content being consumed and the context in which that consumption occurs.

This research makes multiple primary contributions, firstly introducing an approach that can be used in future readability research to consistently, systematically, and at large scale, determine the impact that modifications to variable font axes may have on readability across platforms. An approach that is sensitive to the form of content and the context in which it's presented. Secondly, focusing on a single variable font axis, grade, informing how typographical adjustments can create usable and equivalent experiences across contexts. Given the complexity possible from combinations of multiple variable font axes, this research provides one way of investigating the readability of new variable font axes to ensure that novelty does not sacrifice readability. Finally, this contribution includes a description of how mode impacts readability in these contexts.

2 RELATED RESEARCH

We will outline how reading as an activity is defined and studied to introduce our methodology. We will review literature on experiences of font weight, mode polarity and draw-out references to halation, which does not appear to have been directly studied.

2.1 Reading

2.1.1 Ways of reading. Reading is a highly diverse activity. We read for different reasons, in different ways, to achieve different goals. People employ different strategies and tactics to read in different ways, for different purposes.

"Glance" reading is a term used by researchers to describe characteristics of the task they study [10]where the reader glances to a known location to extract a meaning quickly with minimal cognitive effort. Checking the time on a watch or reading a speedometer while driving are examples of this form of reading in our everyday experience. Mandera (2020 used an internet vocabulary test with people whose first language is English to crowdsource the reading time of 62,000 English words. They established an average time to recognize a single or double syllable word with glanceable reading as 1.30 seconds (standard deviation, 0.35). In lexical decision tasks, deciding whether a string of letters are a word or not, response times are typically under 1 second.

The idea of evaluating reading speed based on a task that involves reading and recognising words when you are expecting to see a word is a persuasively realistic task to assess readability. We plan to use actual words to assess glanceable reading.

"Longform" reading where the reader is engaged with a linear storyline includes things like reading a novel, or an online news

article. A wide range of reading activities sit between glanceable and longform reading described as *Interlude* reading [28]. We will consider both glanceable reading and a subset of interlude reading that we are calling *"Paragraph"* reading. A paragraph is defined here as a text of more than one sentence set over several lines, that can be read on a mobile phone without the need for scrolling. The interactions associated with scrolling play a significant role in the reading process [14] that warrant their own classification as a category of Interlude reading, larger than a paragraph. English language long-form reading fluency is typically in the range of175–300 words per minute [6], nearly three times faster than fluency with 1 and 2 syllable words [20]. These different timings are indicative of different ways of engaging with texts of different lengths. Making differential typographical decisions to support different ways of reading is fundamental to the typographical intent. The research provides an indication of whether typographical guidance should vary based on these different types of reading.

2.1.2 Measuring readability of a surface. Readability is a property of a text, the typography, and the topic for the reader. The methods used to measure readability typically involve the presentation of a text, at minimum one word, and some form of confirmation that the text has been understood. Understanding is normally through reproduction (write, read aloud) or accurate comprehension of multiple-choice questions coupled with speed to achieve that accurate comprehension (fluency). Tasks that map to real world uses to assess reading include proof-reading accuracy [22] in identifying the number of to-be-corrected items.

Deeper understanding of differences in performance are being achieved through eye-tracking during the reading stage [24]. The surfaces used to assess reading are varied (Print, LED, Phone screens, computer screens, Head mounted displays).

We will measure readability of a surface by using readers' fluency in seconds for word recognition (glancing) and comprehension (paragraphs) while also checking on reported interest and familiarity with the text content.

2.1.3 Text contrast polarity effects on reading. Overall display Luminance impacts reading fluency [7], with higher luminance levels in LM leading to faster reading fluency when other variables are stable. Recently, Li [19] used a threshold legibility task coupled with a subjective rating task to determine that luminance of the text is the critical factor influencing legibility, where light text on a dark background delivers legibility enhancements. The type being legible is a necessary prerequisite for reading.

There is also evidence of DM having a negative effect on legibility via reading visual acuity charts (landolt 'c') that is less intense for older adult readers [23]. DM has been attributed as a positive effect for reading on a desktop screen when using an unfamiliar keyboard for the production and correction of errors in a streamlined way [22] and also on reading visual acuity with head mounted displays [11].

Dark mode has been promoted because of the ways it is understood by readers as being more comfortable for the photosensitive, and being less harsh in dark environments. Design professionals are attesting that choosing to read in DM is a matter of personal taste [4] that is more pronounced in readers with individual differences such as dyslexia, aging [5] and people with cataracts. The perspective that it saves battery life because of lower luminance levels is less of an issue with the OLED displays used in higher-end smartphonesa motivator for using DM. Recently, Dash and Hu tried to address the question "How much battery does dark mode save?" Their results show an OLED display draws between 44% and 73% of a phone's total power (averaged across the apps and devices) while in light mode at maximum brightness. For people that have their brightness settings at 50% switching only saved 8.5% battery, and at 30% brightness that saving was less than 5% [9]

Collectively these studies point in the direction of light mode facilitating reading fluency because increased luminance improves reading fluency. This effect would not be expected on devices where the operating system controls for luminance across modes. The manipulation of mode on a system without a luminance-equivalence management system should lead to improved readability in light mode, partly because of this difference in luminance.

2.1.4 Adding weight to reduce blurring. People believe that they read more effectively with bolder text and that thicker letters are more visible [17]. In vision science, studies of letter recognition have established that lower spatial frequencies, thinner text, make letters harder to identify [18]. This low spatial frequency is experienced as 'blurriness' while the two eyes are struggling to align on a small target. Thinner letter strokes and extreme boldness decreased letter recognition across font sizes with a positive effect of boldness at small visual angles that did not occur at large visual angles [3]. We intend to use the knowledge that there is an effect of weight on critical letter legibility to understand how systematic manipulation of that effect impacts relative readability for words and paragraphs.

2.1.5 Adding weight to reduce halation. In popular product design media [25] the halation effect is described as negatively impacting the visual experience of DM relative to LM. Halation is light spreading beyond its boundaries and is mainly studied for controlling effects with film for lighting movies. Figure 2 shows an example of a recommended adjustment to counteract the effect of halation in DM, reducing the font weight. In this example a 64px Yanone Kaffeesatz font is displayed with, and without, an adjustment of 50 compared to the LM rendering. Research on readability and fonts does refer to halation effects. These references are predominantly used as a concept for discussing potential explanations of research results rather than as an intentionally evaluated manipulation, for example, when reading road signage [13]. The variable grade axis provides an alternative to the variable weight axis for managing halation effects. We want to know whether this visual effect impacts readability of body text sizes, making the use of an adjustment necessary to make DM equivalent to LM.

2.1.6 Reader preferences and familiarity effect. Preferences are frequently included in research looking at typographical properties and frequently excluded from experimental research with the recent notable exception where a systematic variation of 20 fonts for reading texts [28] also queried preference and explored the factors influencing preference between fonts. They established that readers' preference for a font did not indicate that they would read most fluently in that font. Software product development decisions should be informed by knowledge of both reading performance

Figure 2: The font called Yanone Kaffeesatz in dark mode 350 and 400 weight either side of light mode 400 weight to facilitate visual comparison of halation effect if viewed on an illuminated screen

and reading preferences, because both influence how users engage with the reading material.

There is an interplay between a readers' familiarity with the concepts raised in texts, how interesting those concepts are, the readers' attention and reading speed. Readers intentionally adjust their speed if they are interested in, or familiar, with the content [27].

Font familiarity, experienced as recent exposure to a typeface, improves reading fluency, while unfamiliar structural design characteristics of the typeface make it less liked [2]. We expect people to read faster in fonts that are familiar. We do not know if people are able to articulate an awareness of that familiarity in a way that aligns with their performance.

2.2 Variable Font files

The research literature on variable fonts is still in its exploratory phases with articles emerging with exciting examples of potential uses [1, 16] and investigations into how type designers approach the challenge of creating a variable font [30]. We have not yet found research that systematically looks at the impact of multiple variable font axes, and realistic interactions between those axes on reading fluency for a font.

2.2.1 Grade axis to control ink seepage. Grade is a variable font axis that originates from type setting to compensate for ink seeping at different rates on different paper types or with different presses. It served to make the typeface look visually identical across different production routes. Identical visual experiences help the designer prescribe text in a format that has a predictable reading experience when delivered in different ways at the time of reading. Grade as an axis is mainly used to visual eliminate differences caused by technology differences, such as in newspaper typography to compensate for differences in paper qualities and printing methods.

The use of grade in a digital environment is comparable when used to eliminate differences that readers might perceive by virtue of differences in luminance. Grade also used to compensate for differences between operating system rendering systems, such as the macOS renderer making the same font outline appear very slightly heavier than the Windows renderer.

Another design challenge raised by using digital media is to manage light spreading, halation, across digitally rendered boundaries without introducing the blurring associated with thinner spatial frequencies. Theoretically a designer can ensure that the rendering will always present the text with equivalent readability by adjusting grade. Reducing grade could counter the light seepage from text in DM potentially increasing visual clarity in DM without differentially impacting reflow of text layout between DM and LM. Figure 1 demonstrates the effect of text reflow when you either increase the weight axis or increase the grade axis.

3 METHOD

We conducted two studies to explore the effects of grade on readability. Initially we focused on a glanceable reading task consisting of word pairs displayed on a desktop computer screen. The second study was driven by an interest in whether there is an effect and if it is more or less detectable in the very different case of paragraph reading on a mobile device screen. The different nature of the reading activity primarily leads to differences in the study designs. In both studies, grade was systematically varied between multiple levels and both light and dark modes were tested to explore the impact o n reading speed. We expected an effect of both mode and grade on reading fluency, and readers preferences for mode. The grade axis scale varies from -200 thru to 150 in Roboto Flex. The first study evaluated the readability of short form text, varying grade at 5 equidistant points along its axis. The second study repeated this equidistant spacing on the scale, then added a point at 100 to get a 'closer' look at the variation where there may be an effect. We hypothesized that grade in DM would need to be lower than grade in LM to be equally readable by countering both halation and screen luminance effects. We hypothesized that increasing the variable font grade axis at a static character size for DM will increase readability, partly because of a reduction in the halation effects, and these effects would be apparent for both glanceable and paragraph reading. We looked at both modes (light, dark) and multiple grade levels as independent variables.

Our analysis phase consisted of ANOVAs for the impact of mode and grade on performance, one study for glance reading and the second for paragraph reading. Readability data for paragraph reading also tracked factors that are known to influence reading fluency of content familiarity, interest in the content and age. For paragraph reading on phones mode settings and preferred mode for the presented paragraphs are compared. Figure 3 shows the manipulations as text using one of the possible word pairs to facilitate visual comparisons. The letter spacing values are adopted from Material Design type system recommendations, and might not be representative of letter spacing in other systems. If the letter spacing is too small, increasing the grade could negatively impact readability.

Text type:	Glanceable	Paragraph
Weight:	500	400
Letter spacing:	0.15	0.25
Size in px:	16	14
Grade manipulation:	Repeated Mesaures	Between Groups
Scale points	equidistant grade axis points	equidistant and 100
-200.0	Jolly buffalo	Jolly buffalo
-112.5	Jolly buffalo	Jolly buffalo
-25.0	Jolly buffalo	Jolly buffalo
62.5	Jolly buffalo	Jolly buffalo
100		Jolly buffalo
150	Jolly buffalo	Jolly buffalo

Figure 3: Demonstration of relative grade variations with letter spacing and weights that are used in these studies

3.1 Glanceable reading on desktop screens

Our first study explored glanceable reading of two-word pairs on a desktop computer. Participants were recruited via Amazon's Mechanical Turk. A power analysis was performed before recruitment to inform the target participant count. Ultimately, 184 participants were recruited for this study to aim for sufficient power in analysis, and to enable removal of responses where participants may have completed the study to earn incentives without actually reading the words presented or where participants have used tools to complete the study. Given the difficulty in estimating sample sizes for a within subject design like this, we used a power analysis calculator to estimate the within-between subjects effect. Our approach is aligned with expectations for psychology research illustrated in [8], establishing a standard for beta at 80% power. We also made sure to exceed the sample size that would be needed had we run several t-test with Bonferroni corrections. The power analysis indicated that a minimum of 126 participants was required to realize statistical power of .85, with α = .05 and effect size of 0.21 for this within-subjects design. In our power analysis, we estimated effect size based partly on our previous readability findings of a small or small-to-medium effect. Although our target sample size was 126, the number of participants collected was higher, at 184, given the anticipated need for data cleaning.

3.1.1 Word-pair presentation. Participants were timed reading word-pairs presented in Roboto Flex [12] lowercase 16px, 500 weight 0.15px letter spacing. Participants were shown a single word pair and told to press the space bar after they had read it. This stopped the timer, and they were then taken to a new screen and asked to type the word-pair to demonstrate that they had read it. They were required to type at least 40% of the string correctly to continue. This presentation and test was repeated 56 times for each participant as we randomly and systematically varied grade between 5 equidistant points on the full scale (-200, -112.5, -25, 62.5, 150.0) and mode (dark, light). Each of the 126 participants read all 56 word experiencing:

- 3 Training questions to become familiar with the process
- 5 Questions per grade
- 5 Grades per mode

Word pairs were constructed from the Dale-Chall word list [26] which contains 3,000 words that should be familiar to at least 80% of children at the US fourth grade level. Word pairs were always a single syllable word paired with a two-syllable word. The words and word-order were randomized.

Estimated time to complete this task was based on feedback from colleagues as approximately 6-8 minutes; participants were paid $1.50 to participate, aiming for an average rate of pay of $12-15/hour.

3.1.2 Data cleaning. This design allowed us to identify and filter-out responses from participants that were not attempting to read the word-pairs by requiring a set number of *"Read"* responses before the task was complete. *Read* responses were identified by calculating the Levenshtein distance [21] between the source phrase and the text each participant entered. Given the Levenshtein distance, we then calculated a ratio to determine the percentage of the string that was correctly typed. Any string that was more than 40% aligned with the source was determined to be *Read*. This level of flexibility was introduced to allow for misspelled words where we could be confident that the respondent grasped the gist of the source phrase. After data cleaning, 126 participants were included in the final analysis. Table 1 shows the age distribution.

Table 1: Age distribution of 126 people reading word-pairs

Age ranges	Respondent percentage
18-24	7.1%
25-34	42.7%
35-44	36.5%
45-54	7.9%
55-64	4.8%

3.2 Paragraph reading on phones

Having completed the glanceable reading study we wanted to determine whether the same effects would hold on an entirely different reading task and device. If findings were consistent across diverse reading environments, we could feel more confident in making general recommendations to enhance readability. For this study, we switched to smaller text (16px to 14px) on a smaller surface (phone) with a different reading activity (word pairs to paragraphs).

3.2.1 Paragraph presentation. For this study each participant read two short texts (103 words vs 108 words), a subset of the open sourced texts used by [28] and provided by the Virtual Readability Lab [1], in a random order in differing font grades and modes. A Qualtrics survey was used, specifying the .CSS properties to typeset Roboto Flex at 14px, 400 weight, 0.25px letter spacing, with reverse foreground and background color polarities of #121212 and #FFFFFF at varied grade levels (-200, -112.5, -25, 62.5, 150.0). These settings align with the Material Design type system[2]. We had the opportunity to add a survey that covered another comparison point between DM and LM, and decided to add a grade level of 100 to add granularity at this point in the axis. Qualtrics measured the presentation time of each question. The survey respondents were asked to read quickly, to understand, and knew they were being timed and tested. Respondents answered two questions about the paragraph text content to demonstrate the understanding to convert reading from speed into fluency. Everyone read in LM and DM in randomized presentation order across respondents.

3.2.2 Survey structure. Before timing the readers we asked them if they normally read with LM, DM, or switched between modes. We also asked them to estimate their reading speed compared to other people. We are not hypothesizing a relationship between self-assessed reading speed and grade. This question is across different reading studies to help build an insight of the relationship between self-awareness, performance and individual differences. After reading each text, respondents were asked to answer 2 multiple choice questions, each with 4 choice options about the text, then rate the paragraph content on unidimensional Likert scales for familiarity and interestingness.

After reading both texts, respondents were asked *"Which paragraph was visually clearer?"*. This question is intended to reveal whether they notice the halation effect. Respondents were asked about their preferred mode for reading these specific paragraphs, in case their current phone settings do not map to their preferences. Finally we asked respondents to rate their familiarity with the font that displayed the paragraphs. This will enable us to investigate whether a feeling of familiarity could be an indicator of familiarity related performance effects in future research.

Before distributing the survey we evaluated it with remote unmoderated usability testing. Four usertesting.com panel members completed the survey with a talk out loud protocol. This confirmed that people were able to understand the questions, and complete the survey in between 4 and 6 minutes with the added load of reading out loud.

[1]https://readabilitylab.xyz
[2]https://material.io/design/typography/the-type-system.html

Table 2: Age distribution of 333 people reading paragraph texts

Age ranges	Respondent percentage
18-19	3.0%
20-29	10.8%
30-39	19.5%
40-49	18.3%
50-59	24.3%
60-69	19.5%
70+	4.5%

3.2.3 Data cleaning. Respondents were recruited from Cint's US panel[3] to be gender balanced and representative of the US adult age-range and screened for completion on an Android phone. The panel screened-in 410 participants. Responses were progressively removed from the set where their responses did not provide evidence that they had read the survey. We started by removing the 48 responses that completed the survey in less than 100 seconds. This fast time was inconsistent with Qualtrics predictions and our own pilot testing. These fast completions are probably indicative of an automated survey completion, not a person engaging with the texts. Next we excluded the 10 responses where one of the 2 texts was read at more than 699 words per minute, more than 2x the normal adult speed [6]. Finally, we looked at how each person answers questions for each paragraph to ensure that we had repeated measures data and enabled comparative comments. If a respondent answered both questions for one of the paragraphs inaccurately, we removed all their data for both paragraphs. This removed 19 responders. After data cleaning, 333 responses remained for use in the analysis. The age distribution of the respondents is shown in Table 2.

4 RESULTS

We will begin by presenting results from the desktop study of word-pairs and then move on to explore the paragraph reading study on phones.

4.1 Glanceable reading on desktop screen results

A repeated measures two-way ANOVA showed that there are significant differences in fluency between LM and DM (p=0.004), details in Table 3. As shown in Figure 4, at grade level 62.5 we observed a mean reading time for dark mode of 2.06 seconds and a mean for

[3]https://www.cint.com/

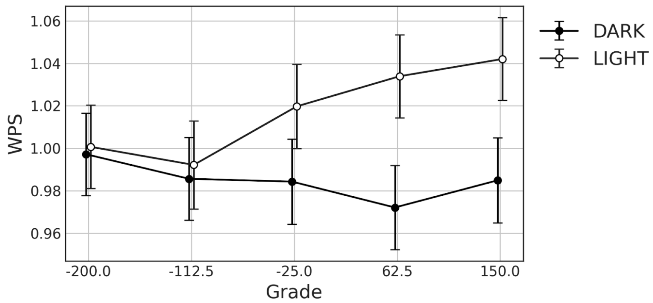

Figure 4: Mean scores in words per second with standard error for Roboto Flex 16px reading speed in light mode and dark mode at different grade levels

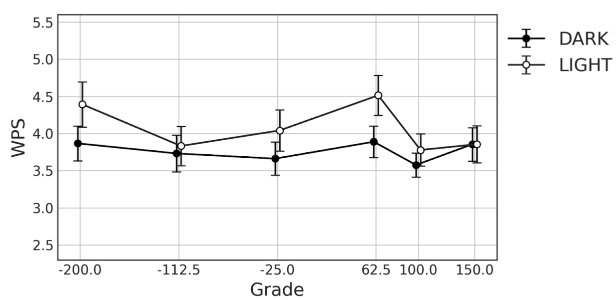

Figure 5: Mean scores in words per second with standard error for Roboto Flex 14px reading speed in light and dark mode at different grade levels

light mode of 1.93 seconds (p<0.01, d=0.16). At grade level 150.0 we observed a mean for dark mode of 2.03 seconds and for light mode of 1.92 seconds. Increasing grade in DM does not bring word-pair reading advantages.

Grade significantly impacted reading time for LM. The significant levels for grade -112.5 (average reading time 2.02 sec) and grade 150.0 (average 1.92 sec). Readability was lower in DM compared to LM and it was significantly lower for the heavier grade levels of 62.5 (Mean for DM=2.06 and LM=1.93) and 150.0 (Mean for DM=2.03 and LM=1.92). Figure 5 demonstrates this effect, plotting words per second against grade.

4.2 Paragraph reading on Android phone results

In the paragraph reading study, we observed a significant main effect of mode with a repeated measures 2-way ANOVA (F=12.8, p<0.01, ^2_p=0.038). LM is more readable t(332)=3.5, p<0.01, with a 95% confidence interval for mean difference and Holm's correction to the post-hoc tests for self estimated fluency. There are no significant interaction effects and in contrast to the word-pair study we observed no significant effect of grade changes. Figure 5 demonstrates this effect, plotting words per second against grade.

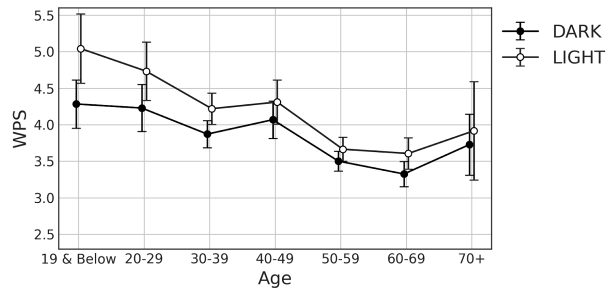

Figure 6: Mean words per second with standard error for dark mode and light mode at different age groups

4.3 Mode settings, clarity and preferences

Respondents' current phone mode settings are reported as balanced between LM

(34%) and DM (36%), with some reporting that they switch modes (25%). Table 4 shows answer distributions to 3 questions we asked of respondents, about light and dark mode, what their current phone typically displays, which of the 2 paragraphs they read was visually clearer and which mode they preferred to read.

When tasked to give a judgment on relative visual clarity of the 2 paragraphs, participants had equally mixed opinions (LM=41%, DM=42%). The assertion of visual clarity does not follow through to their preference for a paragraph mode.

Readers did not disproportionately have more LM settings on their phone, believe reading in LM was more visually clear, or express a preference for reading in LM. Most strikingly, people are not stating that they prefer to read in the font that they perceive to be most visually clear.

We looked at the number of people that read faster in the mode that mapped to their current settings as a way to explore the impact of in-context familiarity. We tested and accepted the null hypothesis that 'mode consistency with settings to paragraphs does not enable respondents to read the paragraphs faster'. Using a chi square this produced a Pearson's probability of 0.31 for DM and 0.18 for LM. Mode consistency does not have a positive effect, and mode changes do not have a detrimental effect on paragraph fluency.

4.4 Age, interestingness and familiarity effects

We tracked the impact of age, content familiarity and interest in the content, as a way to ensure that our data is representative of established variations. We saw the expected trends for age and interestingness, but not familiarity. These respondents are more likely to read fluently in LM than DM (F=10.06, p=0.002, η_p^2=0.030) and there is a between group effect of age (F=3.0, p=0.007, η_p^2=0.053) between people in their 20s and people in their 60s. Figure 6 shows these effects.

We are seeing the known effect of people reading slower in a topic that they find more interesting (LM: F=2.5, P=0.04; DM: F=4.7 p<0.001). We are not seeing an effect of reported content familiarity; whether people are familiar with the topic does not indicate their fluency for that text.

Table 3: Within Subjects Effects

Cases	Sum of Squares	df1	df2	Mean Square	F	p	p-corr	η_p^2	eps
MODE	$1.392e+06$	1	106	$1.392e+06$	8.578	0.004	0.004	0.075	1
GRADE	626181	4	424	156545	1.855	0.117	0.017	0.017	0.955
MODE*GRADE	778811	4	424	194703	2.406	0.048	0.022	0.022	0.952

Table 4: Respondents report current phone mode settings, paragraph mode's visual clarity and preference for reading using that mode

Question answered by participants	Dark background with light text	Light background with dark text	Other options	Total
`Does your phone screen normally show`	36%	34%	25% (switches) 7% (dont know)	100%
`Which paragraph was visually clearer?`	42%	41%	17% (neither)	100%
`How did you prefer to read the paragraphs?`	33%	33%	34% (either)	100%

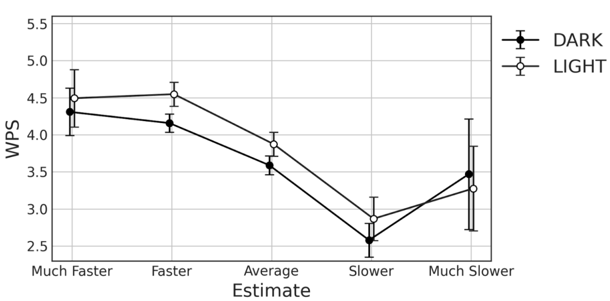

Figure 7: Mean words per second with standard error with self-assessed reading fluency for dark mode and light mode

4.5 Self-assessed fluency

Adult readers are self-aware of their relative reading speed. When asked to estimate their reading speed for an unfamiliar English text, the results map to actual performance for reading fluency. This between-groups effect is significant (F=7.4, p<0.001, η_p^2=0.083) with post-hoc t-tests detailed in Table 5 uncovering significant differences (p<0.01) between *"Much faster"* with *"Slower"*, between *"Faster"* with *"Average"* and with *"Slower"*, and p<0.05 between *"Average"* with *"Slower"*. These results are illustrated in Figure 7.

5 DISCUSSION

We've looked at how the variable font grade axis can be used to relieve an impact of halation on reading fluency in DM without causing text reflow. We've looked at the impact of grade for different reading contexts, glanceable and paragraph reading. Grade changes in DM had no detected effect on reading fluency (glance or paragraph), while glance reading in LM with grade levels over 60 improved fluency. We found the expected effects of people reading more fluently in LM than DM [7, 23] and reading interesting [27] paragraphs less fluently as they moderate their speed and slow down to engage with the content.

The exciting thing about these results is the message that we can give to designers when making typographical decisions with Roboto Flex 14px as body text. Designers can vary the grade to be as positive, or negative, as suits their typographical goals, without concern about a negative impact on paragraph reading fluency, as detectable by this research approach. Grade is not typically used as a form of emphasis in the way that bold or italics can be, yet grade could be applied in this way to provide a subtle emphasis. Designers can adjust grade for aesthetics, user comfort, user preference, or to create emphasis for other typographical goals. Examples of typographical uses are many, grade could be used to convey an auditory rhythm in a poem by increasing and decreasing with the beat of the words. Grade could be used to emphasize the location of the cursor on the screen without causing reflow. While this study did not detect an effect, it is possible that there is an effect, and this effect is small relative to the other factors that influence readability, including age, mode, and fluency self-assessments.

For glancing word-pairs in 16px, the story is the same, except that we'd recommend using LM grade over 60 to increase the readability in LM, and this would enable the best possible fluency. Grade as an axis is mainly used to eliminate differences caused by technology, such as halation, and we've not found evidence of performance or perceptual impact of the technological changes we manipulated within this experimental set-up, by presentation mode and reading context (glanceable, paragraph). Grade can be used for reasons outside of enabling consistency across different technologies, and the case we focused on during this research was body-text reflow.

Designers making typographical decisions can be confident that for 14px, 400 weight, 0.25px letter spacing then 68% (one standard deviation) of their adult readership will be reading between 3.0 - 4.5 words per second. This is a powerful message of typographical flexibility, produced by font flexibility, for engaging with digitally delivered texts on users' own devices. The impact of grade will be

Table 5: Post Hoc Comparisons - ESTIMATE

		Mean Difference	SE	t	p_{holm}
Much Faster	Faster	2.910	19.623	0.148	0.882
	Average	40.156	18.920	2.122	0.207
	Slower	100.598	25.716	3.912	0.001
	Much Slower	61.661	31.896	1.933	0.216
Faster	Average	37.247	11.516	3.234	0.011
	Slower	97.689	20.879	4.679	$< .001$
	Much Slower	58.751	28.143	2.088	0.207
Average	Slower	60.442	20.220	2.989	0.021
	Much Slower	21.505	27.657	0.778	0.875
Slower	Much Slower	−38.938	32.684	−1.191	0.703

affected by the letter spacing, and smaller letter spacing than those we used could have negative effects on legibility. Identifying how grade works with smaller letter spacing is a possible area for future research.

Designers can use grade to address reflow of body text, with the Material Design letter spacing tested, without being concerned about a negative effect on content readability. Designers can select the grade levels that work most effectively for their layouts and preferred visual impacts. This will enable designers to create a shift between LM and DM that maintains consistent experiential qualities across modes without introducing reflow for DM [24] on users' current devices and operating systems. We have not unraveled the extent to which the impact of LM on increased fluency was attributable to increased luminance.

These 2 studies are an early step towards crafting a framework for describing a font's 'reading surface' across all its variable axes. Does grade act as an effective form of emphasis without reflow at different optical sizes? Does a different font at weight 400, size 14px have design characteristics that mean the experience of grade is effectively the same? For all variable font axes, how can they be described in ways that enable designers to use them to achieve the most effective readability, expressiveness and comfort for their readers? How should we be using the variable qualities in the design system in ways that support more effective and satisfying engagement with the user interface?

5.1 Readability as fluency

Our adjustment of ways of reading was external to the reader, the difference between glance and paragraph reading. We saw the reading fluency effect of glanceable reading (16px on computer) being slower than paragraph reading (14px on phone) within ranges consistent with existing research [6, 20]. We saw known effects where the reader's age and values, interest in the topic, are influencing their reading fluency. We also saw that the respondents were not relying solely on their impressions of visual clarity to identify their preference for a mode, or using the fluency levels they achieved as a leading driver of their mode preference choices.

These adult readers are aware of their relative reading speed capabilities compared to other people, the individual differences are most clearly consistently described by their self-assessment of relative reading speed. This meta-level performance knowledge is valuable for exploring abilities and tactics consciously applied to ensure they read how they want to read. Increased fluency is not necessarily the most valuable approach to font selection for a reader. We know that people report the font they find easiest to read in is not actually the font that supports them reading most fluently [28], so maybe those fonts are easier to read for some other reason than fluency? Slower readers might be applying personally relevant reading strategies to create their preferred engagement with the text. Maybe some readers intentionally choose slower reading strategies as an overall approach, while everyone adjusts fluency for interesting or unfamiliar topics [27]. Diverse reading strategies of this nature could account for the range of fluencies we are seeing, but don't explain the results for font, or mode, preference. Systematically describing the ways that people intentionally approach engaging with different texts to unravel how they are intentionally managing their reading. Eye-tracking will play a key role in observationally verifying these potentially different reading behaviors, to describe how people engage with the texts, potentially defining ways of using type to encourage valuable engagement effects, aligned with encouraging fluency, but not treating it as the ultimate success metric.

5.2 No evidence of halation as visual clarity not influencing preferences

Asking our readers' about the relative visual clarity of the modes gave insight into their experience of the halation effect. The 41% who indicated that LM is clearer may be experiencing halation effects in DM coupled with the effects of increased luminance levels provided in LM. The equivalent proportion who cite DM as clearer and the 18% who found neither mode clearer are not aware of experiencing a relative effect of halation in DM. We did not uncover a performance effect or a subjective assessment that indicated any effect on fluency attributable to halation. While an effect may exist, people are not reporting an experience, and it is not evident through this performance metric which does detect the effects of mode and age.

Current mode settings did not demonstrate readers' tendency to favor a mode, visual clarity did not favor a mode, nor did preference for the specific modes used with paragraphs in this study. These 3 reader assessments of mode are all balanced, despite the stable, subtle reading fluency performance advantage of LM. These results are consistent with choosing to read in a mode as being a matter of personal taste [4] , where personal taste of the group doesn't lean towards favoring either mode.

More people attribute visual clarity to a specific mode than cite a preference for using that mode. Our findings are consistent with legibility research [29] that indicates visual clarity (direction of a Landolt-c judgements) and preference judgements are independent. Our respondents are choosing a mode that they judge to be less clear, and we do not know why. Those reasons could be related to aesthetics, comfort, fatigue [11]beyond fluency provided by screen clarity in a relatively short reading task.

5.3 Measuring reading activities that readers value

We've focussed on fluent reading, timings are for relatively successful content reproduction (glance) or multiple choice questions about the paragraphs. These demonstrations of knowledge enabled us to know that the reader had engaged with text and had grasped the gist of the meaning it presented. To establish a systematic, reproducible, analysis we made decisions that addressed non-trivial epistemological questions, such as *"when does a mis-spelt word signify 'not-read?"* and *"when is a multiple-choice option not referred to directly in the paragraph denote not having understood the content?"*. We established a precise definition of understanding for our research texts. The creation and standardization of assessments of how people understand is a complex challenge, and using pre-standardised materials is an appealingly effective approach. Our designs tackled 'equivalence' of reading material by using counter-balanced repeated measures and previously established standard texts from the Dale-Chall word list [26], paragraphs like those we used from the Virtual Readability Lab[4]. Sophisticated approaches have been developed for standardization across languages [15], and we should pay equivalent attention to standardizing our ways of assessing readers' understanding as we do to developing standard texts, ensuring that the assessments are tied to actual ways of understanding.

The way that researchers set-up their assessment of reading-understanding sets the task demands for how people read both glance and paragraph. There are many more ways in which people read in real world environments, so the classification of glance, interlude and longform reading is probably too broad to effectively map the diversity of strategies that people employ in different contexts. Scrolling plays a significant role in the reading process [14] and this is excluded from our paragraph design. Our research classification would yet cover that realistic activity. With our paragraph reading activity, we didn't ask readers questions about 'how' they were reading, nor did we draw them to engage with texts that require scrollable reading, email threads, chat conversations, blog posts, or books. How should the typography adapt across different devices and surfaces to facilitate the interplay of reading and scrolling? We want to encourage people to read in ways that will help them to effectively engage with the digital environment. This raises the question of why people are reading, what are they trying to achieve? Setting-up research task-demands to reflect the things that people are trying to achieve by reading will be the most effective way to establish typographical effects on interactions. Using proof-reading as a task [22] has this solid ecological validity, especially if embedded in a design layout such as form completion. We should also be evolving reading fluency assessments that are embedded in common human interactions. Eye-tracking technology coupled with crafting more realistic task-demands for the reading activities, and more realistic content, will help give us a deeper understanding of how to most effectively use typographical characteristics for engaging readers in task-appropriate ways.

[4]https://readabilitylab.xyz

6 CONCLUSION

How bold can we be with grade? Performance and preferences did not reveal any significant differences for Roboto Flex at size 14px with grade applied. We can be as positive or negative as suits the typographical context, without concern about a detectable negative impact on paragraph reading fluency on a phone. This neither supports nor negates the use of grade to manage the effects of halation. This means we can manipulate the grade variable for aesthetics, other typographical effects, reader comfort or other reader-preferences. This is a powerful message of typographical flexibility produced by font flexibility for engaging with texts. This approach does detect the positive impact of LM, possibly attributable to higher luminance levels, on reading fluency.

The same story applies to 16px for glanceable reading in DM using Roboto Flex lowercase 16px, 500 weight, 0.15px letter spacing, in DM on a computer. We should be using values of grade above 60 in LM to get the best possible readability for glanceable texts on a computer.

Variable fonts, and remote unmoderated evaluations with users, has vastly opened up the possibilities for systematically describing how reading experiences are changed by subtle yet powerful typographical decisions. As a research community, we can systematically vary different aspects of typography (tracking, weight, letter-spacing, line-spacing), observing users' behaviors (words per second, eye-tracking fixation duration) and listening to how this changes readers' strategies and experience. This understanding can be structured into a framework for choosing and using type to achieve the best effects. We will evolve our research with variable fonts towards an articulation of a typographical space that enables crafting of more adaptive and readable surfaces.

ACKNOWLEDGMENTS

We would like to thank Julia Feldman for her guidance on analysis contributions, from all the anonymous readers, The Readability Consortium members, the Google Design Platform researchers, Material Design team, and the Google Fonts team. Finally, thanks to Google Fonts for funding this research.

REFERENCES

[1] Barış Atiker. 2021. Exploring the Future of Spatial Typography in Immersive Design Applications. In *Game + Design Education*, Özge Cordan, Demet Arslan Dinçay, Çağıl Yurdakul Toker, Elif Belkıs Öksüz, and Sena Semizoğlu (Eds.). Springer International Publishing, Cham, 295–310.

[2] Sofie Beier and Kevin Larson. 2013. How does typeface familiarity affect reading performance and reader preference? *Information Design Journal* 20 (10 2013), 16–31. https://doi.org/10.1075/idj.20.1.02bei

[3] Sofie Beier and Chiron A.T. Oderkerk. 2019. Smaller visual angles show greater benefit of letter boldness than larger visual angles. *Acta Psychologica* 199 (2019), 102904. https://doi.org/10.1016/j.actpsy.2019.102904

[4] Olivier Berni. 2019. *The past, present, and future of Dark Mode.* Retrieved August 26, 2022 from http://uxdesign.cc/the-past-present-and-future-of-dark-mode-9254f2956ec7

[5] BoIA. 2021. *Dark Mode Can Improve Text Readability — But Not for Everyone.* Retrieved August 26, 2022 from https://www.boia.org/blog/dark-mode-can-improve-text-readability-but-not-for-everyone

[6] Marc Brysbaert. 2019. How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language* 109 (2019), 104047. https://doi.org/10.1016/j.jml.2019.104047

[7] Axel Buchner, Susanne Mayr, and Martin Brandt. 2009. The advantage of positive text-background polarity is due to high display luminance. *Ergonomics* 52 (2009), 882–886. https://doi.org/10.1080/00140130802641635/

[8] Doğan Yücel Ceyhan Ceran Serdar, Murat Cihan and Muhittin A Serdar. 2021. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochem Med (Zagreb)* 31(1) (2021). https://doi.org/10.11613/BM.2021.010502

[9] Pranab Dash and Y. Charlie Hu. 2021. How Much Battery Does Dark Mode Save? An Accurate OLED Display Power Profiler for Modern Smartphones. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services* (Virtual Event, Wisconsin) *(MobiSys '21)*. Association for Computing Machinery, New York, NY, USA, 323–335. https://doi.org/10.1145/3458864.3467682

[10] Jonathan Dobres, Nadine Chahine, and Bryan Reimer. 2017. Effects of ambient illumination, contrast polarity, and letter size on text legibility under glance-like reading. *Applied Ergonomics* 60 (2017), 68–73. https://doi.org/10.1016/j.apergo.2016.11.001

[11] Austin Erickson, Kangsoo Kim, G. Bruder, and Greg Welch. 2020. Effects of Dark Mode Graphics on Visual Acuity and Fatigue with Virtual Reality Head-Mounted Displays. *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (2020), 434–442.

[12] Santiago Orozco Irene Vlachou Ilya Ruderman Yury Ostromentsky Mikhail Strukov David Berlow Santiago Orozco Irene Vlachou Ilya Ruderman Yury Ostromentsky Mikhail Strukov Font Bureau, David Berlow. 2022. *Roboto Flex specimen.* Retrieved August 26, 2022 from https://fonts.google.com/specimen/Roboto+Flex/

[13] Philip M. Garvey, Martin T. Pietrucha, and Donald Meeker. 1997. Effects of font and capitalization on legibility of guide signs. *Transportation Research Record* 1605 (1 Jan. 1997), 73–79. https://doi.org/10.3141/1605-09

[14] Sian Gooding, Yevgeni Berzak, Tony Mak, and Matt Sharifi. 2021. Predicting Text Readability from Scrolling Interactions. In *Proceedings of the 25th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Online, 380–390. https://doi.org/10.18653/v1/2021.conll-1.30

[15] GA Hahn, D. Penka, C. Gehrlich, André Messias, M. Weismann, Lea Hyvarinen, Markku Leinonen, M. Feely, Gary Rubin, C. Dauxerre, F. Vital-Durand, Sam Featherston, Klaus Dietz, and Susanne Trauzettel-Klosinski. 2006. New standardised texts for assessing reading performance in four European languages. *Br J Ophthalmol* 90 (05 2006), 480–4. https://doi.org/10.1136/bjo.2005.087379

[16] Iván Huelves and L. Pujaol Marco. 2020. Sensor Variable Font: A Model to Improve the Process of Signification of Graphic Interfaces Through Variable Fonts and Data Collected by Sensors. In *Perspectives on Design and Digital Communication.* Springer, 65–90.

[17] Eivind Arnstein Johansen Jonny Nersveen, Anne Kristin Kvitle. 2018. Legibility in Print Text for People with Impaired Vision. *Studies in Health Technology and Informatics* 52 (2018), 862–869.

[18] MiYoung Kwon and Gordon E. Legge. 2012. Spatial-frequency requirements for reading revisited. *Vision Research* 62 (2012), 139–147. https://doi.org/10.1016/j.visres.2012.03.025

[19] Yibin Li, Yanqun Huang, Xu Li, Jie Ma, Jian and; Zhang, and Jutao Li. 2022. The influence of brightness combinations and background colour on legibility and subjective preference under negative polarity. *Ergonomics* 65 (2022), 1046–1056. https://doi.org/10.1080/00140139.2021.2013546

[20] Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2020. Recognition times for 62 thousand English words: Data from the English Crowdsourcing Project. *Behavior Research Methods* 52 (2020), 741–760. https://doi.org/10.3758/s13428-019-01272-8

[21] Ethan Nam. 2019. *Understanding the Levenshtein Distance Equation for Beginners.* Retrieved Sept 10, 2022 from https://medium.com/@ethannam/understanding-the-levenshtein-distance-equation-for-beginners-c4285a5604f0

[22] Lasse Apalnes Pedersen, Svavar Skuli Einarsson, Fredrik Arne Rikheim, and Frode Eika Sandnes. 2020. User Interfaces in Dark Mode During Daytime – Improved Productivity or Just Cool-Looking?. In *Universal Access in Human-Computer Interaction. Design Approaches and Supporting Technologies: 14th International Conference, UAHCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I* (Copenhagen, Denmark). Springer-Verlag, Berlin, Heidelberg, 178–187. https://doi.org/10.1007/978-3-030-49282-3_13

[23] Cosima Piepenbrock, Susanne Mayr, Iris Mund, and Axel Buchner. 2013. Positive display polarity is advantageous for both younger and older adults. *Ergonomics* 56:7 (2013), 1116–1124.

[24] Luz Rello, Susana Bautista, Ricardo Baeza-Yates, Pablo Gervás, Raquel Hervás, and Horacio Saggion. 2013. One Half or 50%? An Eye-Tracking Study of Number Representation Readability. In *Human-Computer Interaction – INTERACT 2013*, Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 229–245.

[25] Robin Rendle. 2020. *Dark mode and variable fonts.* Retrieved August 26, 2022 from http://css-tricks.com/dark-mode-and-variable-fonts

[26] Brian Scott. 1995. *The Dale-Chall 3,000 Word List for Readability Formulas.* Retrieved August 26, 2022 from http://readabilityformulas.com/articles/dale-chall-readability-word-list.php

[27] Todd A. Shimoda. 1993. The Effects of Interesting Examples and Topic Familiarity on Text Comprehension, Attention, and Reading Speed. *The Journal of Experimental Education* 61, 2 (1993), 93–103. https://doi.org/10.1080/00220973.1993.9943854

[28] Shaun Wallace, Zoya Bylinskii, Jonathan Dobres, Bernard Kerr, Sam Berlow, Rick Treitman, Nirmal Kumawat, Kathleen Arpin, Dave B. Miller, Jeff Huang, and Ben D. Sawyer. 2022. Towards Individuated Reading Experiences: Different Fonts Increase Reading Speed for Different Individuals. *ACM Trans. Comput.-Hum. Interact.* 29, 4, Article 38 (mar 2022), 56 pages. https://doi.org/10.1145/3502222

[29] An-Hsiang Wang and Ming-Te Chen. 2000. Effects of polarity and luminance contrast on visual performance and VDT display quality. *International Journal of Industrial Ergonomics* 25, 4 (2000), 415–421. https://doi.org/10.1016/S0169-8141(99)00040-2

[30] Maíra Woloszyn and Berenice Santos Gonçalves. 2022. The Design Process of Variable Fonts: A Prospective Survey-Based Investigation with Type Designers. In *Advances in Design and Digital Communication II*, Nuno Martins and Daniel Brandão (Eds.). Springer International Publishing, Cham, 54–65.