Towards Readability Individuation: The Right Changes to Text Format make Large Impacts on Reading Speed

Shaun Wallace[1], **Rick Treitman**[2], **Nirmal Kumawat**[2], **Kathleen Arpin**[3], **Jeff Huang**[1], **Ben Sawyer**[4], **Zoya Bylinskii**[2]

[1]Brown University
{shaun_wallace, jeff_huang}@brown.edu

[2]Adobe Inc.
{treitman, kumawat, bylinski}@adobe.com

[3]Riverdale Country School
kmarpin@gmail.com

[4]University of Central Florida
sawyer@ucf.edu

ABSTRACT

In our age of ubiquitous digital displays, adults often read in short, opportunistic interludes. We consider, for the first time, whether reading outcomes in this unique Interlude Reading can be improved by tailoring typeface to the individual. Hundreds of participants provide a foundation for understanding which fonts people prefer and which make them more effective readers. Results reveal that while 77% believed their preferred font would be fastest to read in; this was only valid for 20%. Differences between best and worst font average 75 words per minute (WPM), with no significant changes in comprehension. High WPM variability for every font suggests that one font does not fit all. We here provide recommendations for favorable fonts related to higher reading speed without sacrificing comprehension and suggest that our methodological approach can be used to model for individuation, allowing digital devices to match their users' needs in-the-moment.

Author Keywords

reading, typography, text

INTRODUCTION

Effective readers imagine they are alike, but each reader struggles in their own way, and each might be provided tools to read better. Reading is a large and growing portion of adult work and entertainment, although often not in traditional forms. Many adult readers struggle to keep up; several prior studies have shown a widespread occurrence of adult readers across the United States struggling with reading speed and comprehension [2, 26, 28, 29, 43]. A study by the International Adult Literacy Survey Institute found that around 23% of adults in the United States read prose at the lowest level scored, indicating significant reading fluency problems [43].

Initial work with children has shown manipulating the font can increase accurate reading speed 25%-50% [42] These changes have also been shown to increase an adult's accurate reading speed by 20% or more [16]. This paper and past work show that, with the right reading tools, even good readers can be augmented, with increases in reading speed but no reductions in comprehension.

We are currently in the age of easy access to digital tools, big data, and as a byproduct, many opportunities for the digital personalization of content served to the individual. We posit that the reading experience can also be individualized toward significant real-world improvement via digital tools and applications (e.g., e-readers and reading applications on phones). Research by a non-profit organization, Readability Matters [42, 15], has shown that tuning the font family, character spacing, and line spacing of text can greatly improve the reading performance of school children. Notably, current settings in e-readers provide options that would have been considered miraculous by the traditional standards of printing, allowing the user to adjust the font family, size, and color of the text. However, many conventions remain from the lengthy era in which text was less flexible, and both blind spots and opportunities for exploration and enhancement result. In this era of increasingly unconventional tools, we see advantages in challenging conventional wisdom.

Our motivation comes from the idea that if the right reading tools are made readily available to all, the cumulative impacts can be significant and widespread—from an improvement to the learning outcomes of struggling readers, to the more effective ingestion of reading material by college students, and quick information intake in high-paced business settings. Speed-reading tools aside, we wonder what gains in reading performance can be obtained with less invasive changes to the reading experience. In this paper, we consider how changing the font family and size of a piece of text can change reading outcomes, as measured by reading speed and comprehension. If changing the text format can improve reading outcomes, then it is essential to know whether the user's preferences can set the format or whether it needs to be automatically inferred and suggested. This knowledge can guide the design of future reading tools, which will either hand reading controls over to the user or set them automatically based on standard guidelines.

Towards these goals, we lay a foundation for understanding systematically what people like in a text format and what makes them more effective readers. We set up large-scale reading

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

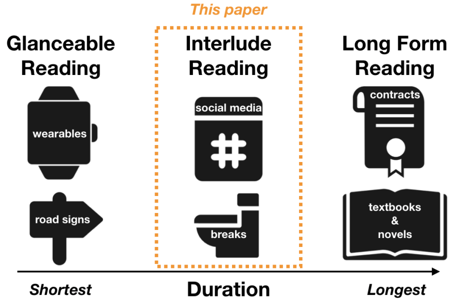


Figure 1. Reading modes as a duration continuum, in which the already well-studied domains of reading at a glance and reading at length bookend reading in short opportunistic interludes. We define Interlude Reading as the kind of reading that happens in a single brief sitting (i.e., a few paragraphs worth).

studies, covering hundreds of participants across diverse user populations, from online crowdworkers to college students and professionals. We focus our attention on paragraph reading using digital tools. Through our studies, we seek to answer the following two questions:

- **Universality VS Personalization:** Does one size fit all? Are some fonts generally preferred and effective across a population, or are there large individual differences?
- **Preference VS Effectiveness:** Do people know what's good for them? Are people most effective at reading in the fonts that they prefer most?

We build on the study designs and findings of related reading and typography work in order to present the first large-scale reading experiments to study text format preferences and effectiveness systematically. Based on the results of our studies, we make some recommendations about reading formats at the end of this paper. These recommendations can guide future reading applications and more generally, any other text-heavy digital resource.

RELATED WORK

Our work focuses on general reading by adults on digital devices. It falls between glanceable reading and long form reading. We define "Interlude Reading" (Fig. 1) to capture opportunistic reading (during short breaks), quick information gathering (research in business settings), and ingestion of content through social media platforms. This form of reading aligns with reading materials that are easily digestible (i.e., a few paragraphs worth), and fits Carver's recommended range of 138–600 words per minute for reading with comprehension [12].

Long Form Reading

The lion's share of reading research addresses long form reading. Here, evidence for the impact of individual typeface upon reading efficacy or experience is less available. Indeed, many studies linking typography to any form of performance rely upon the presentation of short passages of text, something unhelpful in the context of our present dichotomy of durations for reading. These efforts do, potentially, shed light on considerations for Interlude Reading. For example, Rudnicky and others showed that letter size and case were influential factors in reading performance [39], a finding reinforced by Bernard and Mills [4, 13].

Glanceable Reading

In the domain of glanceable reading, performance differences between typefaces are well documented [18, 19, 37]. The concept of legibility in at-a-glance reading revolves around the ability to collect understanding under time pressure. Toward this goal, others have in both Western and symbolic Eastern languages demonstrated that psychophysical methods could be used to differentiate the utility of individual fonts [18]. Indeed, Sawyer and others have specifically called out the propensity of designers to focus on aesthetic concerns over performance concerns, especially in contexts where safety is paramount and the cost of failure high [40].

Interlude Reading, the reading mode we introduce and focus on in this paper, can not be categorized as either long form or glanceable reading. As such, present research addressing reading is insufficient, and the opportunity has arisen to explore, describe, and leverage the flexibility of digital mediums to better engineer reading experiences in this new mode. These days, Interlude Reading occupies a central role in textual consumption, as the information age increasingly drives individuals to consume information in short, opportunistic interludes.

Typography & Reading in the Digital Age

Research exploring typography as a tool to enhance reader efficacy or experience has a rich history. For example, in pursuit of better reader experiences, O'Donovan et al. developed a system to explore font selection using crowdsourced attributes [33]. However, this work does not provide a definitive ranking among all font choices. Typographical manipulations have also been used in pursuit of performance and enhanced reading acuity, or for individuals with visual impairments [1, 44]. Indeed, a large portion of this literature focuses on fonts that help or impair reading for individuals with low vision [27].

More recently, the question of which font to use in online settings has spawned several informative research efforts [3, 4, 6, 9]. These efforts focus on a small number of fonts (2–4) and point sizes ranging from 10–14 points, using different physical screen sizes and resolutions than those presently common.

Building off this work, Bernard et al. ran a study with 60 participants to measure reading time, preference, and errors reading aloud using eight different fonts with sizes 10-point, 12-point, and 14-point [5]. While they found preferences changed at different sizes, our work aims to control for perceptions in font size across a broader population with a larger more modern set of fonts. In a follow-up work, Bernard studied reading speed among a group of 35 participants, where they preferred the bigger font size (12-point over 10-point) across the two fonts (Arial and Times) tested [7]. Both of these seminal works do not capture how preference can affect reading speed and comprehension, but they point to an ongoing question of how differences in actual and percieved size can affect user's font preference.

Our work asks similar questions of effectiveness but at a much larger scale. We study a total of 16 fonts, many of which did not exist in the past. Burmistrov et al. show that light and ultra-light fonts also induce higher cognitive load [10]. Our work studies fonts with thinner stroke widths such as Montserrat and Avant-Garde that are recommended by readability experts for reading body text [15]. More recent efforts have focused on studying larger font sizes and line spacing [38]. They recommend default larger font sizes that we explore in this work at scale. They do not study how font preference and effectiveness align. Bhatia et al. studied the effect of various font and text attributes, such as size and satisfaction level with the text, on readability within a group of 180 undergraduate students [8]. Our work studies font size across a more diverse user population that is twice the size. In addition, we control for passage comprehension and font size.

STUDY OVERVIEW

Our motivation is to determine whether changing the text font on a digital device can improve the reading outcomes of adult readers. Provided it can, we need to know whether the font (i) should be universal or personalized, and (ii) whether participants' preferences can guide the font choice.

To measure what makes fonts preferable and effective for reading, this paper culminates in a large-scale reading study, run on hundreds of crowdsourced participants with 16 different fonts (Fig. 2). To answer these questions systematically, we designed a number of preliminary studies along the way, individually developing and validating the components of the final study.

In the first study, we designed an experiment to measure font preference and effectiveness. To measure preference, we developed a font toggle test that determined a participant's favorite font through a double-elimination tournament. Participants then read passages in different fonts, while their reading speeds and answers to comprehension questions were recorded. This study, run on three different user populations, provided some initial findings on the properties that drive font effectiveness.

Finding that the perceived font size (independent of the font size in pixels) partially influenced some of the results from the first study, next, we designed a perceptual task to derive a crowd-driven size normalization for each of our 16 fonts. Finally, for our large-scale reading study, we reused the components from the first two studies to create a comprehensive test of font preference and effectiveness.

Fonts: Every day, people are exposed to various fonts while reading for work or leisure across mediums ranging from printed news to websites, and magazines to books. To cover these variations in our studies, we selected four fonts each from four different sources (Fig. 2): (i) four of the most common fonts used for (digital) documents[1], (ii) four of the most popular fonts for print media [17], (iii) four fonts recommended by readability experts [41, 15], and (iv) four of the most common fonts used on websites[2]. The default font size we chose was 16px [4, 14, 30]. Modern browsers, including Firefox and Chrome, ship with a default font size of 16px. The interfaces we used to conduct our studies did not allow participants to alter the font size and were constrained to a fixed size.

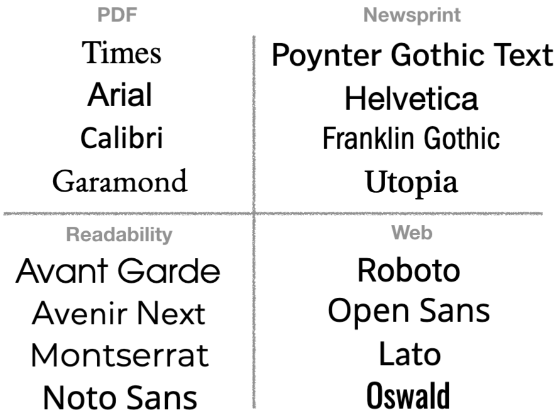


Figure 2. The 16 fonts used throughout our experiments, chosen because they are popular fonts that span different use categories.

PRELIMINARY STUDY

Starting with the hypothesis that people's font preferences can point to more effective fonts, we first designed and validated a method to determine a participant's preferred font using pairwise comparisons. The full study design alternated between preference elicitation and reading evaluation.

Guiding Questions:

- Which are the highly rated fonts?
- What factors influence font preference?
- Do people have similar preferences?

Study design: Pairwise comparisons are a standard method across different fields to derive personal preference [21, 36, 46]. However, determining a participant's preferred font among 16 fonts can be a time-consuming task if they make every possible pairwise comparison. Methods exist to synthetically fill-in a pairwise matrix to speed-up this process [24, 36]. However, this does not eliminate ties between fonts due to the transitive property. It would be ideal if the most preferred font also had the highest number of pairwise comparisons per participant to add further weight to their preference. A double-elimination style tournament, where a font is eliminated after a participant picks against it twice, is a method to decrease the number of total match-ups, remove the possibility of ties, and to arrive at a definitive winning font per participant.

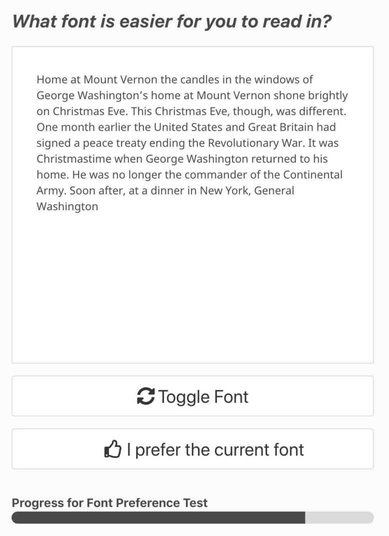
The total number of pairwise comparisons in a double elimination tournament can be computed as $(N-1)\times 2+1$, where N is the total number of fonts in this study. The pairing of fonts is randomized before each round of pairwise comparisons. In this work, participants also make additional random comparisons to validate previous results.

We designed an interface to toggle between two fonts in order to choose the preferred one, using the prompt: "What font is easier for you to read in?" (Fig. 3). Toggling between pairs of options at a time provides a simple and efficient method for assessment, motivated by other pairwise comparison tasks in the wild, such as eye exams and hearing aide adjustments [32].

The full study alternated between (i) preference tests, where participants would perform the toggle test to compare a set

[1] Obtained from an Adobe corpus of 2302 PDF documents.

[2] https://fonts.google.com/analytics



What font is easier for you to read in?

Home at Mount Vernon the candles in the windows of George Washington's home at Mount Vernon shone brightly on Christmas Eve. This Christmas Eve, though, was different. One month earlier the United States and Great Britain had signed a peace treaty ending the Revolutionary War. It was Christmastime when George Washington returned to his home. He was no longer the commander of the Continental Army. Soon after, at a dinner in New York, General Washington

Figure 3. Font preference toggle test: A participant toggles between pairs of fonts to decide which one is easier to read in. The interface is always a fixed width of 420px regardless of the device. This toggle test is done repeatedly within a double-elimination tournament over pairs of fonts to determine a participant's preferred font. A participant toggles the font family used to display the sample text, then they stop on the font of the pair they prefer, and click to indicate their preference.

of four fonts and (ii) effectiveness tests, where participants would read two sets of passages and answer two sets of comprehension questions. Half of the time, participants read passages in their preferred font, and the other half the time in another randomly-assigned font. Participants' answers to the comprehension questions, and reading speed on all the reading portions, were recorded as measures of effectiveness. Study phases (i) and (ii) alternated multiple times until all 16 fonts were compared. Participants also completed a practice phase and filled out pre- and post-surveys. The full study description and survey results can be found in the supplemental material.

Participants: We recruited 63 participants: 12 participants via university mailing lists, 15 participants from the UserTesting.com platform, and 36 crowdworkers from Amazon's Mechanical Turk. Participants across all groups ranged in age from 18 to 55 years (average = 31). Overall, 51% of participants identified as female. It took participants 40 minutes on average to complete the study. Participants were compensated in accordance with the pricing guidelines of each platform.

Data preprocessing: Across all studies presented in this paper, participants answered several voluntary pre-survey questions to ensure their data was not affected by any diagnosed disabilities (e.g., dyslexia), medical and neurological conditions (macular degeneration, diabetes, ADD, memory disorders, LPD, dyspraxia, etc.), and any other external factors (reading environment, caffeine, nicotine, etc.). For each participant who self-reported any of the above factors, we tested if their overall words-per-minute (WPM) or comprehension score fell outside the normal distribution of data using the IQR method. Participants could also be removed if their average dwell time per font during the preference test fell outside the normal distribution of data. This removal method covers the scenarios where participants iterated between pairwise comparisons too quickly, or always chose the second option of the pair.

To establish a range of reading speed indicative of "interlude reading," we expand on Carver's recommended range of 138–600 WPM to account for standard error, to remove any individual WPM measurements and the corresponding reading comprehension scores that are outside the range of 100–650 WPM [11, 12]. In this study, 9% of the participant data was removed based on the pre-processing methods described above.

Evaluation metrics

Win Rate: The percent of pairwise match-ups each font won during the font preference test.

Elo Rating: Since the font preference test consisted of a double-elimination tournament, participants did not make every possible pairwise comparison. To account for the strength of each font in a pairwise comparison, an Elo Rating [20] was computed per font, per participant. Elo Ratings were averaged across participants to create an average Elo Rating per font. The initial Elo Rating per font was 1500, and the system ran with a K value of 64, which is higher than usual to account for a small number of pairwise comparisons.

Why Elo Rating: Elo Ratings have been used for rating chess players [20], in educational settings [34], and to help mitigate cold-start problems in recommender systems [45]. By design, the Elo Rating system aims to minimize the differences between expected and actual outcomes of competitions, or in our case, pairwise comparisons. Elo Rating is an appealing option to determine user preference given small sample sizes [22]. Other ranking alternatives include TrueSkill [25] and Rank Centrality [31].

Disagreement: This is the standard deviation across all participants' Elo Ratings per font. The greater the number, the less consensus there was among overall preference per font.

Preference Consistency: The rate a participant's current font preference given a unique pairwise comparison matches their previous preference for the same pairwise comparison.

Results

Which are the highly rated fonts? Noto Sans (chosen by 9 participants), Montserrat (chosen by 8), and Garamond (chosen by 8) were the fonts selected most frequently as the overall winners in the preference tests (Table 1, 'Most Preferred'). The rest of the fonts were similar, in that they were mostly chosen by 1–4 participants each (with two exceptions: Franklin Gothic had 0, Roboto had 5). In other words, there are clear winners, but there is also a lot of diversity in font preferences across individuals.

Apart from the overall winners of the preference tests, we also considered the percent of pairwise match-ups each font won across all participants (Table 1, 'Win Rate'). For a fairer evaluation, 'Elo Rating' is a related metric that additionally takes into account the strength of the opponent in each pairwise match-up. These metrics were computed based on many data points because each participant compared each font multiple times against other fonts. Within these metrics, Noto Sans and

Montserrat are the top fonts, but Garamond is in the bottom 5. How to reconcile this with the previous result? Garamond led to split opinions across participants - those who liked it, liked it a lot (it was their top font); others disliked it (voting it down in pairwise match-ups). Garamond has a high inter-participant disagreement score (Table 1, 'Disagreement'). For that matter, so does Montserrat, although it is still a top font because it won more pairwise match-ups against other fonts than it lost.

Taking all these results together, Noto Sans is a clear favorite. It was the absolute favorite font of 9 participants and was in the top 5 fonts for 46 participants (almost 80% of the participant pool). With the highest win rate and average Elo Rating, it was the most consistently preferred font.

During the font preference tests, we included validation rounds to repeat some of the pairwise match-ups and check for participant response consistency (see supplemental material for the details). Preferences across unique font pairings were consistent 79% of the time. The IQR method was used to compare individual participant consistency values, and no participants were found to be inconsistent according to this metric.

What factors influence font preference? Participants remarked that they liked bigger (P6, P11, P13, P22, P26, P42, P45, P54) and bolder fonts (P6, P22, P23, P42, P43, P45), fonts that were more modern (P13), with good kerning (P13), "linear and not curvy" (P6), with good spacing between the letters and lines (P23, P43), "wide and spaced out" (P51). A few participants remarked that their preference might depend on the context of what was being read (P3, P18).

Looking at the results of the present study, Noto Sans beats Open Sans in all three preference metrics (Table 1). The main difference between the two fonts is that Noto Sans has a heavier stroke, reflecting participants' preferences. The fonts with the highest Elo Rating also tend to be larger than other fonts.

Does familiarity drive font preference? Participants completed a pre-survey where, among other things, they indicated which content they commonly read for work and leisure. In the post-study, participants rated their familiarity with each font on a 5-point Likert scale after reading a sample of text written in the font. Arial, Times, and Helvetica were rated as most familiar, on average, while Lato, Poynter Gothic Text, and Montserrat were rated least familiar. From the pre-survey responses, we can confirm that if people mostly read novels for leisure, then the most familiar rated fonts make sense (Fig. 4). However, familiarity was not predictive of font preference. Participants' familiarity with their recommended font followed a normal distribution. Pearson's Correlation between font familiarity and Elo Rating per font per participant was 0.02 ($p = 0.6$). The most preferred font, Noto Sans, was also among the least familiar fonts to participants.

Do people have similar preferences? The fact that all but one font was selected as a favorite by at least one of the participants (Table 1, 'Most Preferred') points to individual differences.

Considering the three user populations separately, and keeping in mind that there was a different number of participants per group, the most preferred 3 fonts across 36 MTurk participants were Noto Sans, Open Sans, and Montserrat (ordered by Elo Rating); across 15 UserTesting participants: Roboto, Noto Sans, and Open Sans; across 9 university participants: Utopia, Times, and Noto Sans. University participants spend a lot more time reading textbooks and papers, explaining the difference in results. We assume that the MTurk and UserTesting participants when taken together represent the broader population better than the university participants.

Font	Most Preferred	Win Rate	Avg Elo Rating	Disagreement
Noto Sans	9	62%	1635	88
Montserrat	8	56%	1598	124
Garamond	8	43%	1421	124
Roboto	5	56%	1583	78
Lato	4	54%	1553	73
Helvetica	4	53%	1525	103
Arial	4	55%	1567	88
Poynter Gothic	3	54%	1542	63
Times	3	47%	1471	99
Avenir Next	3	48%	1502	82
Utopia	3	51%	1505	90
Open Sans	2	57%	1598	71
Oswald	2	22%	1236	110
Calibri	1	46%	1512	80
Avant Garde	1	40%	1422	142
Franklin Gothic	0	34%	1329	78

Table 1. Results from the preliminary study. Noto Sans consistently performed highly: it was both the most preferred (including highest win rate and Elo Rating) across all 16 fonts. 'Most preferred' refers to the total number of participants (out of 60) for whom the selected font was the absolute favorite. 'Win Rate', 'Average Elo Rating', and 'Disagreement' refer to the toggle-based font preference test. A high disagreement score means participants had highly varying opinions of the font. Green (versus red) cells correspond to fonts with the best (correspondingly, worst) scores according to each metric.

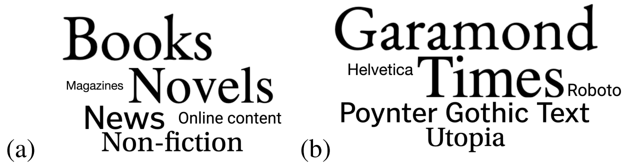


Figure 4. A proxy for the fonts people commonly encounter. (a) Content participants reported reading most for leisure. Text size approximately maps to frequency of response. (b) Fonts commonly used for the corresponding reading content. Font name is rendered in its own font type.

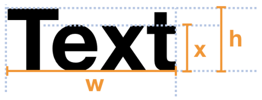
Take-aways:

- Noto Sans is a top-rated font across participants
- Effective font size influences font preference
- Font preference is not driven by familiarity
- Participants vary in their preferences for fonts

FONT SIZE NORMALIZATION STUDY

Not all fonts are created equal. In the preliminary study we found a strong effect of effective font size on people's preference. Previous work has discussed how Georgia was designed

Figure 5. Text can be normalized by adjusting its x-height (x), height (h), or width (w) to match a reference font.



to have a larger x-height compared to similar fonts to give a perceived advantage over Times [9] and how x-height can increase legibility [35]. Here we attempt to correct for the perceived size differences of fonts when they are all initialized at 16px, using Times as a reference. While prior work proposes to normalize font sizes according to a particular attribute (e.g., x-height [5, 9]), here we take a crowdsourced approach to finding the attribute, per font, that perceptually normalizes its size the best.

Guiding Question: What is the best way to normalize a font's size?

Procedure

Font normalization: Taking Times at 16 px as our reference font, we computed three new font sizes for each font in our set, corresponding to matching the reference in each of x-height, height, or width (Fig. 5). The first of these, x-height, is the height of the main body of lowercase letters (excluding ascenders and descenders). For our fonts, it was either already listed in OpenType[3], or as commonly done in typography, set to be equal to the height of a small letter 'x'. Normalizing by x-height means adjusting the font size of the target font until its x-height matches the x-height of the reference font. Normalizing by width or height corresponds to adjusting the font size of the target font until the width (or correspondingly, height) of a set of characters (in Fig. 5, the word "Text") matches the width of the same characters rendered in the reference font. Further details and all the normalized font sizes computed are provided in the supplemental material.

Study Design: Participants completed a perception study to select the best normalization method for each of our 16 study fonts and 4 tutorial fonts – used for practice sessions to familiarize participants with our study (Georgia, Verdana, Raleway, and Comic Sans). Our interface presented participants with two screens side-by-side, with the same piece of text rendered in two different fonts (Fig. 7). One of the screens always contained the reference font (Times), and the other screen contained a target font. Participants could click to toggle between four possible settings of the target font, one of which was the original font setting, and the rest corresponded to normalizing the font by x-height, height, or width. The four settings were shuffled per font, per participant. After toggling through all the settings, a participant would click to select the setting most similar in size to the reference. Participants could also swap the reference and target fonts, which facilitated quickly toggling back and forth between the fonts. There was also an option to change the underlying text passage.

Participants: We recruited 61 participants: 23 via university mailing lists, 18 professionals ranging from designers to engineers, and 20 crowdworkers from Amazon's Mechanical Turk. We did not collect demographic information for this study.

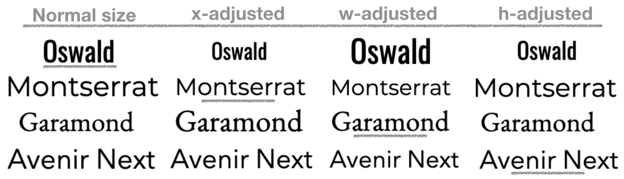
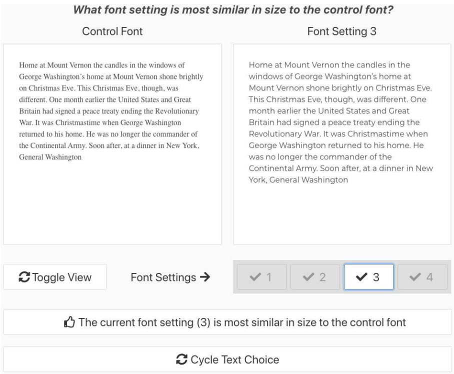


Figure 6. Different fonts, when rendered with the exact same font size, can have quite different effective sizes (first column above). Depending on which criteria is used for normalizing a font (x-height, width, or height) the adjusted effective font size can also vary drastically. The assumption that a single criterion can be used to normalize all fonts proves sub-optimal (compare fonts across any one column), although normalizing by height work best overall (last column). We take the approach of selecting the optimal normalization strategy on a font-by-font basis, as determined by a population of participants. This corresponds to using the font sizes on the diagonal in this figure (i.e., normal size for Oswald, x-adjusted size for Montserrat, etc.).



What font setting is most similar in size to the control font?

Control Font

Home at Mount Vernon the candles in the windows of George Washington's home at Mount Vernon shone brightly on Christmas Eve. This Christmas Eve, though, was different. One month earlier the United States and Great Britain had signed a peace treaty ending the Revolutionary War. It was Christmastime when George Washington returned to his home. He was no longer the commander of the Continental Army. Soon after, at a dinner in New York, General Washington

Font Setting 3

Home at Mount Vernon the candles in the windows of George Washington's home at Mount Vernon shone brightly on Christmas Eve. This Christmas Eve, though, was different. One month earlier the United States and Great Britain had signed a peace treaty ending the Revolutionary War. It was Christmastime when George Washington returned to his home. He was no longer the commander of the Continental Army. Soon after, at a dinner in New York, General Washington

The current font setting (3) is most similar in size to the control font

Figure 7. Our perception study for normalizing font size. Participants see a control (reference) font in one panel and a target font in the second panel. They can toggle between 4 different settings of the target font, which corresponds to adjusting the font size to match the reference in x-height, width, height, and no adjustment. These settings are shuffled for each participant. In this figure, Times is on the left and Montserrat is on the right.

Results

Normalization can have a big effect on font size: The base font sizes for different fonts vary significantly (Fig. 6). For instance, Garamond is a smaller font, while Montserrat is naturally quite large. Comparing blocks of text rather than individual words demonstrates how the size difference gets amplified (Fig. 7).

Re-analyzing the data in the preliminary study, we computed which font was larger per pairwise comparison according to each of: x-height, width, and height. We ran two-tailed t-tests of unequal variance. For x-height the winning font was larger 52% of the time, for width the winning font was larger 57% of the time, and for height the winning font was larger 53% of the time ($p < 0.05$ for all). While size proved significant, the relatively small difference in mean win rate indicates size was not the only factor driving font preference.

What is the best way to normalize a font's size? Previous papers that have mentioned the need to account for a font's effective size have indicated that x-height should be the criteria used [5, 9]. However, our study results go against this common wisdom and show that the preferred way to normalize a font actually

[3] https://docs.microsoft.com/en-us/typography/opentype/spec/os#sxheight

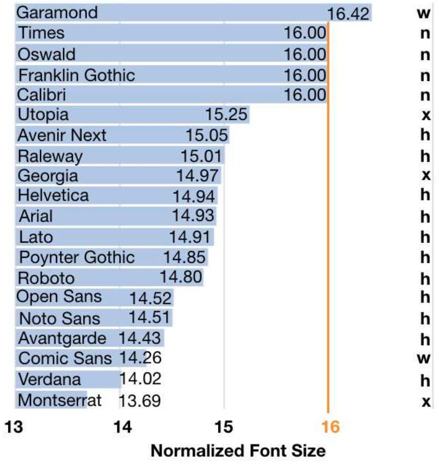


Figure 8. Crowdworkers voted on which of three normalizing quantities (x = x-height, w = width, h = height) or no adjustment (n), when applied to a font, best matched the perceived size of Times at 16px. The majority vote per font (indicated to the right of the graph) was used to determine the final size to use for each font, plotted on the bars.

depends on the font itself. Fig. 8 shows the normalization factor that a majority of participants picked. In most cases, height was the most frequently-picked criterion. However, it's not always the best, and in some cases the differences between the best normalization factor for a font and another normalization factor can be quite large (Fig. 6). Rather than choose a single normalization strategy for all the fonts, we selected the most frequently chosen normalization strategy on a font-by-font basis, as determined by the crowd. This resulted in the final normalized font sizes plotted in Fig. 8. These final sizes are still above the recommended minimum font size for reading on a digital device [30].

Take-aways: There is not a single effective way to normalize a font's size. Normalization strategies depend on the font.

FONT STUDY ACROSS A LARGE POPULATION

The motivation of the present study was to determine whether people are most effective at reading in the fonts that they prefer, after controlling for font size and reading comprehension. To answer this question, we ran experiments on hundreds of participants on Amazon's Mechanical Turk.

Guiding Questions:

- Which are the highly rated fonts (controlled for font size)?
- Is people's preferred font their most effective font?
- What gains in reading are achievable by font choice?

Procedure

Study materials: A reading specialist collected a set of 15 text passages from Project Gutenberg[4], a repository of creative commons e-books. The passages were chosen to span different topics (history of science, biography, botany, etc.), with 12 non-fiction and 3 fiction passages. Passages were curated down to 300–500 words, with minor adjustments to sentence structure and vocabulary to be at approximately an 8-th grade level (Lexile range[5]: 800–1200, Flesch score[6]: $60.5 - 79.8$). The reading specialist also prepared 4-6 similar-level comprehension and inference questions per passage. As such reading material has not previously been made available in the public domain, we will be releasing it along with our paper for future reading studies. For the present study, we selected 9 non-fiction and 3 fiction passages, and further cut them down to 160–178 words. Then we split each passage approximately in half to be presented across 2 consecutive reading screens, without breaking apart sentences (69–93 words each). We selected 2 comprehension questions per passage, one corresponding to each half of the passage. In this way, participants would need to read both halves of the passage carefully enough to answer both comprehension questions correctly.

Study design: Participants could complete the study on a device of their choice: computer, tablet, or mobile screen. Participants began the study with a pre-survey asking a range of questions, including about demographics (age, education, native language), reading experience (frequency, type of content, device of choice), vision (normal/corrected), disabilities (learning or reading), state (under the influence of drugs, medications, alcohol), and environment (lighting, time of day). The full survey and aggregate results are provided in the supplemental material.

After an instructional screen, participants proceeded to the practice phase, with short versions of both the preference test and effectiveness test, to get acquainted with the study flow (Fig. 9). The first phase of the main study was a preference test, similar to the one in the preliminary study, but run as a double-elimination tournament over all 16 study fonts. The preference test was split into a competition block of 30 comparisons, followed by a validation block of 6 comparisons, which were randomly selected repeat comparisons from the competition block to measure a participant's self-consistency. Participants used the toggle interface (Fig. 3) for the pairwise comparisons.

After the preference test, participants completed 10 rounds of the effectiveness test. Each round consisted of reading a passage split across two consecutive reading screens (69-93 words per screen), followed by two multiple-choice comprehension questions, and a mini questionnaire asking participants about their reading technique, as well as familiarity and interest in the topic matter presented, using a 5-point Likert scale. Each participant read a total of two passages in each of 5 fonts. Assignment of fonts to passages was randomized per participant. The 5 fonts used were as follows: Noto Sans (best overall font from the preliminary study), Times (baseline font used in the preliminary study), the participant's preferred font (from the preference test), and two randomly-selected fonts out of the remaining 13 study fonts[7]. As a result, across the 10 rounds of reading, each font was used for two different reading passages (different topic, different content, similar length). We recorded the time spent per reading screen and the responses to the study questions.

[4] https://www.gutenberg.org

[5] https://hub.lexile.com/analyzer

[6] http://www.readabilityformulas.com/free-readability-formula-tests.php

[7] In cases where the participant's preferred font was one of Noto Sans or Times, we would sample three, instead of two, randomly-selected fonts out of the remaining 14 study fonts.

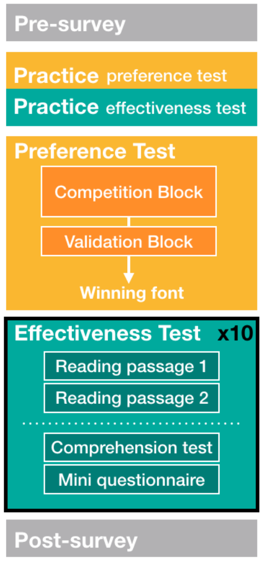


Figure 9. Apart from surveys and a practice session, this study design is split into two distinct phases. In the first phase, participants complete a preference test: (i) a double-elimination tournament with 16 fonts, leading to 30 pairwise comparisons (competition block), and (ii) 3 repeated pairwise comparisons (validation block). In the second phase, participants complete an effectiveness test: 10 rounds of (i) reading 2 passages (average 70 words each), (ii) answering 2 comprehension questions (comprehension test), and (iii) answering 3 additional questions about interest, familiarity, and reading technique (mini questionnaire).

The study ended by showing participants: (i) the font that won the most match-ups across all the preference tournaments, i.e., their most preferred font, and (ii) the font that they read the fastest in. A post-survey asked participants about their familiarity with each of the 20 study fonts (including the 4 practice fonts), their experience with the toggle interface, their reaction to their preferred font, and how effective they think their preferred font would be to read in.

Participants: We recruited 500 participants on Amazon's Mechanical Turk. Of the 386 participants (46% female) that remained after data preprocessing, ages ranged from 18 to 71 years (average = 33): 8 were younger than 20, 150 were in their 20s, 148 in their 30s, 51 in their 40s, and 29 were older than 50. Participants took on average 34 minutes to complete this study and were compensated \$5 for their time.

Data preprocessing: Participants were removed from the study if they met one of the following conditions: (i) did not submit both pre and post surveys, (ii) did not self-report being "very comfortable" reading in English, (iii) self-reported being diagnosed with any reading or learning disability, medical or neurological condition, (iv) self-reported being under the influence of any drugs, medications, or alcohol, (v) had either of their WPM or reading comprehension scores outside the normal distribution as computed using the IQR method, or (vi) had their preference consistency outside the normal distribution as computed using the IQR method. We also removed individual data points outside the range of Interlude Reading as in the preliminary study (100–650 WPM, derived from Carver [11, 12]). After this filtering (removing 22% of participants), the data of 386 participants was used for reporting the results of this study.

Evaluation metrics

For analysis, we used the same metrics as in the preliminary study, along with a few additional ones we define here.

WPM: We measured reading speed in words-per-minute (WPM) computed as $(w \times 60)/s$ where w is the number of words in a passage and s is the number of seconds spent reading the passage.

Speed Rank: Per participant, we compared their WPM on the 5 study fonts they were presented. We treated this as an implicit pairwise comparison by sampling each pair of fonts out of the 5 fonts used and keeping score of which font of the pair had the highest WPM. Across all participants, this produced a win percentage of each font against every other font, which we can interpret as a speed rank for that font over all other fonts.

Comprehension score: We measured comprehension as the percent of questions answered correctly, by choosing one out of three multiple choice answers. Each participant read 2 sets of passages per font and answered 2 comprehension questions per passage. When we report comprehension score as a percentage, it is based on a total of 4 questions per font, per participant.

Results

Factors controlled for: During our data preprocessing procedure, we removed participant data with poor comprehension scores, below 0.71 according to the IQR method. Two-tailed t-tests of unequal variance with Bonferroni correction showed no significant differences in reading comprehension scores across fonts in the remaining participant data. We also found no effect of content on reading speed. We conducted a two-tailed t-test of equal variance to compare the average words per minute for fiction passages (M = 273) compared to non-fiction (M = 277) passages. There was no significant difference in reading speed between the type of passage. The data presented has effectively been controlled for reading comprehension.

We also measured and found either none or minor effects of font familiarity ($r = 0.042$), passage familiarity ($r = 0.033$), passage interest ($r = -0.053, p < 0.05$), and age ($r = -0.16, p < 0.05$) on reading speed (with Bonferroni correction). Similarly, there were no effects of font familiarity ($r = -0.004$) and passage familiarity ($r = 0.02$) on reading comprehension. There was a minor effect of passage interest ($r = 0.11, p < 0.05$) on reading comprehension.

Which are the highly rated fonts? Noto Sans and Times were each chosen by 56 (17% of all) participants as the overall winners in the preference tests (Table 2, 'Most Preferred'). Avenir Next (chosen by 41) led another group of fonts (Helvetica, Calibri, Garamond, Arial, Open Sans, and Roboto) which performed similarly well across preference metrics. The rest of the fonts won less than half of their overall matchups. Despite this, every single font was the preferred font of at least 4 participants. This points to a lot of diversity in font preferences across individuals.

Another indication of diversity in preferences are the inter-participant disagreement scores (Table 2, 'Disagreement'). For instance, while Times and Garamond were highly rated fonts overall, they led to split opinions across participants, some of which consistently voted them up (correspondingly,

Font	Most Preferred	Win Rate	Avg Elo Rating	Disagr-eement	Font Familiarity	WPM	Speed Rank	SD WPM	Compre-hension
Noto Sans	56	62%	1639	90	1.89	272	48%	108	91%
Times	56	58%	1596	115	2.50	277	50%	108	91%
Avenir Next	41	54%	1554	97	1.74	264	45%	106	93%
Helvetica	36	59%	1608	87	2.22	283	50%	102	89%
Calibri	35	55%	1573	90	2.34	276	56%	102	92%
Garamond	34	52%	1543	103	1.90	310	48%	120	91%
Arial	33	57%	1591	86	2.40	270	47%	103	93%
Open Sans	19	56%	1585	77	2.03	255	54%	91	90%
Roboto	14	53%	1556	84	1.83	268	47%	106	94%
Montserrat	13	42%	1451	90	1.77	271	57%	109	87%
Utopia	13	44%	1464	105	1.81	274	48%	116	86%
Avant Garde	11	38%	1398	104	1.83	261	29%	90	94%
Oswald	10	16%	1154	127	1.70	295	58%	99	89%
Lato	6	49%	1519	72	1.73	293	54%	99	91%
Poynter Gothic	5	44%	1473	78	1.82	265	52%	97	93%
Franklin Gothic	4	27%	1296	87	1.79	270	56%	107	89%

Table 2. Results from our large scale font study. Noto Sans consistently performed highly across preference: it was both the most preferred (including highest win rate and Elo Rating). 'Most preferred' refers to the total number of participants for whom the selected font was the absolute favorite. 'Win Rate', 'Average Elo Rating', and 'Disagreement' refer to the toggle-based font preference test. A high disagreement score means participants had highly varying opinions of the font. 'Font Familiarity' was a 5-point Likert scale question from the post-survey (5 = very familiar). 'Times Read' is a measure of how many times participants read a given font in a reading passage. This is not uniform across fonts, because font preference was used to guide which fonts participants read in. 'WPM' and 'Comprehension' refer to the reading effectiveness test.

down) in the preference tests. On the other hand, Open Sans and Arial were generally likeable, as witnessed from their low inter-participant disagreement scores.

Controlling for size generally led to smaller fonts (Times, Garamond, Helvetica, Calibri) performing consistently better than in the preliminary study. As a notable example, Times, which was in the bottom 5 fonts according to Elo Rating in the preliminary study, was in the top 5 fonts in the present study. However, Noto Sans, was relatively stable in performance across both studies. It was in the top 5 fonts for 80% of participants in the preliminary study, and 77% of participants in the present study. With the highest win rate and average Elo Rating, it was the most consistently preferred font, overall.

As in the preliminary study, familiarity was not predictive of font preference. Participants were familiar with their recommended font only 52% of the time. Pearson's Correlation shows only a small effect between font familiarity and Elo Rating per participant ($r = 0.18, p < 0.05$). The most preferred font, Noto Sans, was also among the least familiar fonts to participants.

Is people's preferred font their most effective font? We set this study up to explicitly consider effectiveness, measured via WPM and comprehension. However, since comprehension has been controlled for, here we look at reading speed only. Each participant read passages and answered comprehension questions in each of: their most preferred font, Noto Sans, Times, and two randomly selected fonts.

We found no consistent differences in WPM of different fonts that were stable across participants. Only Garamond showed significant increases ($p < 0.05$) in average WPM compared to other fonts (according to two-tailed t-tests of unequal variance with Bonferroni correction). No other fonts were consistently effective across participants. However, differences in font effectiveness did show up at the individual level.

Overall, participants read the fastest in their most preferred font 20% of the time, but they also read the slowest in their preferred font 19% of the time (out of five total fonts tested per participant), which works out to precisely chance level. Overall participants read in their preferred font at an average WPM. Participants read faster in their most preferred fonts than in Times 50% of the time, and faster than in Noto Sans 51% of the time. In other words, participants do no better or worse, on average, by reading in their preferred font. These findings run contrary to participants' beliefs: 73% of participants believed their most preferred font would be their most effective font to read in. Also, Times and Noto Sans, which were generally preferred fonts, were not consistently effective fonts across participants.

What gains in reading are achievable by font choice? At the individual level, there appear to be effective fonts for participants. Participants read 14% faster in their fastest font (313 WPM) compared to their most preferred font (274 WPM). Most impressively, participants read 32% faster (313 WPM) in their fastest font compared to their slowest font out of the five tested (238 WPM). Importantly, because we controlled for comprehension, and removed unreliable WPM measurements (outside of 100–650 WPM), differences in reading speed across fonts do not imply that participants were skimming the material, but rather that they were able to get through it faster while achieving comparable comprehension scores. Font size was also controlled for in this study. In other words, the impressive differences in reading speed that we observe here are due to font type alone (Fig. 10).

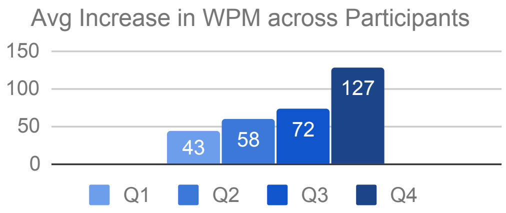


Figure 10. Participants are divided in quartiles based on average WPM. This figure shows that the fastest readers also had the biggest gains in reading speed from slowest to fastest font, indicating the potential for the type of font to have significant impacts on reading efficiency.

Take-aways

Here we summarize the take-aways for this study, together with what we learned from the other studies in this paper.

- Preference for fonts is personal. People differ in what they prefer.
- Familiarity drives neither preference nor effectiveness of fonts. A font need not be chosen for an application just because people may be used to seeing it.
- Preference ≠ effectiveness. People do not know what is good for them in terms of font choice for reading.
- Different fonts are effective for different people, leading us to believe that custom reading experiences can help people read more effectively.
- A single size does not fit all fonts. If an application has a few font options for the same piece of text, then each font needs to be adjusted in size according to the font's characteristics.

Limitations and future work

Participants: We recruited participants from the general population. Our participants are those who specifically opted-in to our reading studies, so we may have self-selected for the more effective readers. Even though we aimed for diversity, by running our experiments on students, professionals, and crowdworkers, there are populations we inevitably did not reach, and as a result, are not represented in our data. In particular, the majority of participants were in their 20s and 30s, and we specifically excluded participants who reported any learning or reading disabilities (due to a small sample). Our reading studies can be extended to other specialized populations to evaluate the generalizability of our results, and to identify where the most significant effects of font type on reading performance are achievable.

Metrics: The words per minute (WPM) calculation used in this study is a rough measure of effectiveness. Future studies can consider how people read by moving their eyes or can evaluate the fluency of reading by recording reading aloud.

Reading formats: In studying fonts, there are multiple factors to control for. This work concentrates on normalizing size across a wide range of popular font families. Future work could delve into fine-grained format differences beyond font, which include character spacing, kerning, and stroke width.

CONCLUSION

This work presented the first large-scale reading study considering the effect of font on Interlude Reading performance, as measured by words per minute and comprehension. It was the first study to systematically pit typeface preference against effectiveness in such reading. On evidence from our study that font size may impact WPM, and that participants tended to prefer larger fonts when given a choice, we worked to normalize font size. Our strategy for normalization is an exciting innovation in its own right, and as such, that methodology may be useful to future researchers. Even with our size normalization, we found preference unrelated to effectiveness, and not predicted by familiarity. Indeed, preference does not drive effectiveness, despite participants believing that it does. We did find favorable typefaces, and here we have made recommendations for font families and corresponding font sizes. However, the implications of this work go beyond a list of recommendations.

There is, given our pattern of findings, an exciting opportunity to augment reading performance for adult readers. Indeed, our average reader could add 38 words a minute by merely adjusting their font, equivalent to an additional 3-4 pages an hour. Participants in our top quartile for the delta between best and worst font would add 93 WPM, or eight pages an hour. In both cases, average comprehension remains similar and high. In the context of interlude reading, this gain is perhaps best framed in the ability to consume more in limited windows. If a news article, journal, or forum post is roughly 700 words, requiring around two minutes for an average reader, that individual could read it 24% faster in their most effective font while still retaining normal levels of comprehension. The approximately 30 seconds saved in this case might be used to read comments or look at related posts. Social media companies, who thrive on frequent interaction, might find it in a typeface.

There is an enormous opportunity, as well, to engineer better reading. Font familiarity drove neither preference nor effectiveness, and preference was not a driver of performance, findings which together challenge conventional wisdom on the benefits of following prior conventions on font choice. Indeed, such conventions may be argued as relics of a time when the ability to choose a new font and re-flow a document was the labor of hours or weeks. Can we provide flexible options that help individual readers toward success? In that regard, the high variability seen in our studies represents both a challenge and an opportunity. Clearly, there is potential for improvement, but understanding how to help each individual, and potentially how to help each individual in a subset of different contexts in which they read, is presently an unsolved problem. Further research here should target the individuation of typeface to user and context.

While preference did not drive performance, that is no reason to believe that it is not a driver of use. Indeed, the ideas of design for pleasure and Hedonomics [23] suggest that design for pure utility is a poor strategy. Aesthetic considerations remain, and the present findings can be considered a mandate to match beautiful fonts with the people they can help, and who will likewise appreciate them. If a novel method could ascertain someone's most effective font, this work suggests users would believe this recommendation, read in the given font, and become a more effective reader.

The transformation of reading by digital devices is at the heart of our work and dictates the subsequent work necessary. Our definition and description of Interlude Reading, our large sample data collection, and our substantiated idea of font individuation supporting readers, each is made possible through digital media. The potential impacts on individual reading efficacy highlighted here point to a future in which machines help adult readers to reach for their full reading potential. We invite the present reader, and the multidisciplinary communities that will perform this work, to join us. Let us engineer better reading for everyone.

ACKNOWLEDGEMENTS

We have a lot of people to thank for giving our work the resources and context it needed to bloom: Drs. Benjamin Wolfe and Jonathan Dobres for formative discussions about reading that helped shape some of our initial ideas; Allen Ellison and Emily Seminerio of Adobe Inc. for providing us with helpful resources; Kathy Crowley and Marjorie Jordan of Readability Matters for helpful discussions of the study results; Dan Rhatigan and Tim Brown of Adobe Type for insightful discussions about typography; Tong Sun of Adobe Research for her support.

REFERENCES

[1] Aries Arditi. 2004. Adjustable typography: an approach to enhancing low vision text accessibility. *Ergonomics* 47, 5 (2004), 469–482.

[2] Justin Baer, Mark Kutner, John Sabatini, and Sheida White. 2009. Basic Reading Skills and the Literacy of America's Least Literate Adults: Results from the 2003 National Assessment of Adult Literacy (NAAL) Supplemental Studies. NCES 2009-481. *National Center for Education Statistics* (2009).

[3] Jayeeta Banerjee, Deepti Majumdar, Madhu Sudan Pal, and Dhurjati Majumdar. 2011. Readability, subjective preference and mental workload studies on young indian adults for selection of optimum font type and size during onscreen reading. *Al Ameen Journal of Medical Sciences* 4, 2 (2011), 131–143.

[4] Michael Bernard, Chia Hui Liao, and Melissa Mills. 2001. The effects of font type and size on the legibility and reading time of online text by older adults. In *CHI'01 extended abstracts on Human factors in computing systems*. ACM, 175–176.

[5] Michael Bernard, Bonnie Lida, Shannon Riley, Telia Hackler, and Karen Janzen. 2002. A comparison of popular online fonts: Which size and type is best. *Usability news* 4, 1 (2002), 2002.

[6] Michael Bernard and Melissa Mills. 2000. So, what size and type of font should I use on my website. *Usability news* 2, 2 (2000), 1–5.

[7] Michael L Bernard, Barbara S Chaparro, Melissa M Mills, and Charles G Halcomb. 2003. Comparing the effects of text size and format on the readibility of computer-displayed Times New Roman and Arial text. *International Journal of Human-Computer Studies* 59, 6 (2003), 823–835.

[8] Sanjiv K Bhatia, Ashok Samal, Nithin Rajan, and Marc T Kiviniemi. 2011. Effect of font size, italics, and colour count on web usability. *International journal of computational vision and robotics* 2, 2 (2011).

[9] Dan Boyarski, Christine Neuwirth, Jodi Forlizzi, and Susan Harkness Regli. 1998. A study of fonts designed for screen display. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co., 87–94.

[10] Ivan Burmistrov, Tatiana Zlokazova, Iuliia Ishmuratova, and Maria Semenova. 2016. Legibility of light and ultra-light fonts: Eyetracking study. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. ACM, 110.

[11] Ronald P Carver. 1990. *Reading rate: A review of research and theory.* Academic Press.

[12] Ronald P Carver. 1992. Reading rate: Theory, research, and practical implications. *Journal of Reading* 36, 2 (1992), 84–95.

[13] Maneerut Chatrangsan and Helen Petrie. 2019. The effect of typeface and font size on reading text on a tablet computer for older and younger people. In *Proceedings of the 16th Web For All 2019 Personalization-Personalizing the Web*. ACM, 5.

[14] Jan Constantin. 2013. Typographic Design Patterns And Current Practices (2013 Edition) âĂŤ Smashing Magazine. `https://www.smashingmagazine.com/2013/05/typographic-design-patterns-practices-case-study-2013/`. (May 2013). (Accessed on 09/19/2019).

[15] Kathy Crowley and Marjorie Jordan. 2019a. Base Font Effect on Reading Performance - Readability Matters. `https://readabilitymatters.org/articles/font-effect`. (June 2019). (Accessed on 09/19/2019).

[16] Kathy Crowley and Marjorie Jordan. 2019b. Readability Formats Offer Instantaneous Change - Readability Matters. `https://readabilitymatters.org/articles/instantaneous-change`. (Jan 2019). (Accessed on 09/20/2019).

[17] Bill Davis. 2004. Fonts on the front page. A study of typefaces on the front pages of AmericanâĂŹs top newspapers. *Chicago: Ascender Corporation* (2004).

[18] Jonathan Dobres, Nadine Chahine, Bryan Reimer, David Gould, Bruce Mehler, and Joseph F Coughlin. 2016a. Utilising psychophysical techniques to investigate the effects of age, typeface design, size and display polarity on glance legibility. *Ergonomics* 59, 10 (2016), 1377–1391.

[19] Jonathan Dobres, Bryan Reimer, and Nadine Chahine. 2016b. The effect of font weight and rendering system on glance-based text legibility. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 91–96.

[20] Arpad E Elo. 1978. *The rating of chessplayers, past and present*. Arco Pub.

[21] Shengbo Guo and Scott Sanner. 2010. Real-time multiattribute bayesian preference elicitation with pairwise comparison queries. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 289–296.

[22] Severin Hacker and Luis Von Ahn. 2009. Matchin: eliciting user preferences with an online game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1207–1216.

[23] Peter A Hancock, Aaron A Pepe, and Lauren L Murphy. 2005. Hedonomics: The power of positive and pleasurable ergonomics. *Ergonomics in design* 13, 1 (2005), 8–14.

[24] Patrick T Harker. 1987. Incomplete pairwise comparisons in the analytic hierarchy process. *Mathematical modelling* 9, 11 (1987), 837–848.

[25] Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. TrueSkill: a Bayesian skill rating system. In *Advances in neural information processing systems*. 569–576.

[26] Mark Kutner, Elizabeth Greenberg, and Justin Baer. 2006. A First Look at the Literacy of America's Adults in the 21st Century. NCES 2006-470. *National Center for Education Statistics* (2006).

[27] J Stephen Mansfield, Gordon E Legge, and Mark C Bane. 1996. Psychophysics of reading. XV: Font effects in normal and low vision. *Investigative Ophthalmology & Visual Science* 37, 8 (1996), 1492–1501.

[28] Gail McKoon and Roger Ratcliff. 2016. Adults with poor reading skills: How lexical knowledge interacts with scores on standardized reading comprehension tests. *Cognition* 146 (2016), 453–469.

[29] Brett Miller, Peggy McCardle, and Ricardo Hernandez. 2010. Advances and remaining challenges in adult literacy research. *Journal of learning Disabilities* 43, 2 (2010), 101–107.

[30] Aliaksei Miniukovich, Antonella De Angeli, Simone Sulpizio, and Paola Venuti. 2017. Design guidelines for web readability. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. ACM, 285–296.

[31] Sahand Negahban, Sewoong Oh, and Devavrat Shah. 2016. Rank centrality: Ranking from pairwise comparisons. *Operations Research* 65, 1 (2016), 266–287.

[32] Arlene C Neuman, Harry Levitt, Russell Mills, and Teresa Schwander. 1987. An evaluation of three adaptive hearing aid selection strategies. *The Journal of the Acoustical Society of America* 82, 6 (1987), 1967–1976.

[33] Peter O'Donovan, Jānis Lībeks, Aseem Agarwala, and Aaron Hertzmann. 2014. Exploratory font selection using crowdsourced attributes. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 92.

[34] Radek Pelánek. 2016. Applications of the Elo rating system in adaptive educational systems. *Computers & Education* 98 (2016), 169–179.

[35] Eustace Christopher Poulton. 1965. Letter differentiation and rate of comprehension in reading. *Journal of Applied Psychology* 49, 5 (1965), 358.

[36] Li Qian, Jinyang Gao, and HV Jagadish. 2015. Learning user preferences by adaptive pairwise comparison. *Proceedings of the VLDB Endowment* 8, 11 (2015), 1322–1333.

[37] Bryan Reimer, Bruce Mehler, Jonathan Dobres, Joseph F Coughlin, Steve Matteson, David Gould, Nadine Chahine, and Vladimir Levantovsky. 2014. Assessing the impact of typeface design in a text-rich automotive user interface. *Ergonomics* 57, 11 (2014), 1643–1658.

[38] Luz Rello, Martin Pielot, and Mari-Carmen Marcos. 2016. Make it big!: The effect of font size and line spacing on online readability. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3637–3648.

[39] Alexander I Rudnicky and Paul A Kolers. 1984. Size and case of type as stimuli in reading. *Journal of Experimental Psychology: Human Perception and Performance* 10, 2 (1984), 231.

[40] Ben D Sawyer, Jonathan Dobres, Nadine Chahine, and Bryan Reimer. 2017. The Cost of Cool: Typographic Style Legibility in Reading at a Glance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 61. SAGE Publications Sage CA: Los Angeles, CA, 833–837.

[41] Bonnie Shaver-Troup, Kathy Crowley, and Marjorie Jordan. 2017. Optimizing Reading Performance by Manipulating the Shape, Size, and Spacing of Text to Match the Individual's Visual Processing Capacity. *Collective Impact Project, revReading* (2017).

[42] Bonnie Shaver-Troup, Kathy Crowley, and Marjorie Jordan. 2019. Optimizing Reading Performance by Manipulating the Shape, Size, and Spacing of Text to Match the Individual's Visual Processing Capacity - Readability Matters. `https://readabilitymatters.org/articles/optimizing-reading`. (Jan 2019). (Accessed on 09/20/2019).

[43] William Thorn. 2009. International adult literacy and basic skills surveys in the OECD region. (2009).

[44] Miles Albert Tinker. 1963. *Legibility of print.* Technical Report.

[45] Jingwei Xu, Yuan Yao, Hanghang Tong, Xianping Tao, and Jian Lu. 2015. Ice-breaking: mitigating cold-start recommendation problem by rating comparison. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*. 3981–3987.

[46] Jinfeng Yi, Rong Jin, Shaili Jain, and Anil Jain. 2013. Inferring users' preferences from crowdsourced pairwise comparisons: A matrix completion approach. In *First AAAI Conference on Human Computation and Crowdsourcing*. 207–215.