

Personalized Font Recommendations: Combining ML and Typographic Guidelines to Optimize Readability

Tianyuan Cai
tcai@adobe.com
Adobe Inc.
San Francisco, California, USA

Shaun Wallace*
shaun_wallace@brown.edu
Brown University
Providence, Rhode Island, USA

Tina Rezvanian
rezvania@adobe.com
Adobe Inc.
San Francisco, California, USA

Jonathan Dobres
jdobres@gmail.com
Virtual Readability Lab
Orlando, Florida, USA

Bernard Kerr
bkerr@adobe.com
Adobe Inc.
San Francisco, California, USA

Samuel Berlow
samuelberlow@gmail.com
Typography for Good
Putney, Vermont, USA

Jeff Huang
dis@jeffhuang.com
Brown University
Providence, Rhode Island, USA

Ben D. Sawyer
sawyer@inhumanfactors.com
University of Central Florida
Orlando, Florida, USA

Zoya Bylinskii
bylinski@adobe.com
Creative Intelligence Lab
Adobe Research
Cambridge, Massachusetts, USA

ABSTRACT

The amount of text people need to read and understand grows daily. Software defaults, designers, or publishers often choose the fonts people read in. However, matching individuals with a faster font could help them cope with information overload. We collaborated with typographers to (1) select eight fonts designed for digital reading to systematically compare their effectiveness and to (2) understand how font and reader characteristics affect reading speed. We collected font preferences, reading speeds, and characteristics from 252 crowdsourced participants in a remote readability study. We use font and reader characteristics to train FontMART, a learning to rank model that automatically orders a set of eight fonts per participant by predicted reading speed. FontMART's fastest font prediction shows an average increase of 14–25 WPM compared to other font defaults, without hindering comprehension. This encouraging evidence provides motivation for adding our personalized font recommendation to future interactive systems.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Information systems** → **Personalization**.

KEYWORDS

readability, reading, typography, personalization

*Also with the Creative Intelligence Lab, Adobe Research, as an intern.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DIS '22, June 13–17, 2022, Virtual Event, Australia

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9358-4/22/06...\$15.00

<https://doi.org/10.1145/3532106.3533457>

ACM Reference Format:

Tianyuan Cai, Shaun Wallace, Tina Rezvanian, Jonathan Dobres, Bernard Kerr, Samuel Berlow, Jeff Huang, Ben D. Sawyer, and Zoya Bylinskii. 2022. Personalized Font Recommendations: Combining ML and Typographic Guidelines to Optimize Readability. In *Designing Interactive Systems Conference (DIS '22), June 13–17, 2022, Virtual Event, Australia*. ACM, New York, NY, USA, 25 pages. <https://doi.org/10.1145/3532106.3533457>

1 INTRODUCTION

The design of our reading experiences has historically been in the hands of the content producers (authors, publishers, designers, etc.), but as reading increasingly shifts to digital platforms, control over reading formats can be handed to the users. The multitude of device types, screen qualities, digital interfaces, and software settings available to the reader provide plentiful opportunities for personalization and customization. Readers today can interact with the fonts, sizes, contrast, and other settings on their devices to better fit their reading needs in the moment; for instance, a reader may increase the digital font size as an alternative to grabbing a pair of reading glasses.

Beyond font size manipulations, growing evidence suggests that personalizing reading formats can have significant impacts on reading performance [12, 90, 93]. For example, non-profit Readability Matters conducted a study showing that different text formats (a combination of font choices and spacing) increased accurate reading speed among adult readers by 20% or more [29]. Other recent work showed that, with the right font, individuals could potentially read 25% faster [90]. These works propose that individual readers need to be matched with their fastest font, just like an eye doctor might prescribe reading glasses unique to an individual. An open question is how to match an individual with their fastest fonts to help them efficiently digest large amounts of information?

With over 800,000 digital fonts available [30] and an evolving set of digital devices, finding individuals' fastest fonts to redesign their text presents a unique challenge [90]. Prior readability research has studied popular fonts such as Arial and Times New Roman [12, 73]

and individual font characteristics such as line and character spacing [2, 41, 73]. We collaborated with typographers to select eight fonts for a remote readability study and to identify various font and reader characteristics that might affect reading speed on digital devices. Our discussions and font selections focus on reading general body text in English. We performed structural coding [55, 60] on the typographer interviews to identify font and reader characteristics (i.e., age, font familiarity, x-height, etc.) that may affect reading. We also developed a web application, FontView, to extract and display the characteristics (i.e., weight, stroke contrast, etc.) per font.

Using the eight fonts identified with typographers (§3), we conducted a remote readability study to collect reading speeds, font preferences, and reader characteristics from 252 paid crowdworkers (§4). Using the font and reader characteristics as features and reading speeds as labels, we trained FontMART, a learning to rank model built off LambdaMART [21]. FontMART predicts the relative reading speeds for different readers and fonts, and we use it to rank and recommend the predicted fastest fonts to individual readers.

We ground this research in the following questions:

RQ1: How effective is machine learning at selecting a font to redesign text to increase a participant’s reading speed?

RQ2: Which reader and font characteristics identified by typographers are most predictive of individual reading speeds in a font?

To answer RQ1, we proposed three baselines to help evaluate FontMART’s font recommendation per participant (§5). Our results show that FontMART’s recommendation can provide average speed improvements of 14–25 WPM over the baselines (§6). To answer RQ2, we analyzed individual font recommendations and used Shapley values to understand how font and reader characteristics identified by typographers contribute to faster reading (§6) [54, 79]. We contextualized our empirical results with qualitative analyses of the typographer interviews. Finally, we discuss how our results can be used by typographers, designers, and publishers in selecting readable fonts; and how FontMART can be built into future interactive systems to benefit readers (§7).

Our contributions include: (1) FontView, a web application for extracting font characteristics that affect reading; (2) a crowdsourced reading dataset containing the results of 252 individuals reading in 8 different fonts, along with the relevant demographics and performance information; (3) a population analysis replicating prior results showing personalizing font choice can increase reading speed [90, 93]; (4) the FontMART model to provide personalized font recommendations; (5) quantitative and qualitative analyses of the font and reader characteristics that influence reading speed. We share FontView, FontMART, and our reading dataset¹.

2 RELATED WORK

Fonts are a vital element of modern digital reading interfaces. However, open research questions persist on the need to personalize font choice and how best to do so [47, 73, 80].

2.1 Interfaces for Font Selection

Selecting fonts involves making decisions based on many nuanced characteristics that are difficult for human eyes to distinguish [62]. Researchers have proposed various adaptive interfaces to simplify

font selection for different purposes. Acknowledging the difficulty of selection using the “long, alphabetically-sorted” font drop-down list, O’Donovan *et al.* provided alternatives that include interface organization with descriptive attributes and hierarchical grouping by perceptual similarity [62]. Jiang *et al.* proposed two font pairing algorithms that help narrow the selection for design tasks [45]. Other support for font selection includes indexing and recommending fonts based on their affective associations [48]. While previous font selection methods maintain the user’s freedom of choice, our approach can provide readers with an individualized font recommendation that is likely to increase their reading speed.

2.2 Recommender Systems for Interfaces

Recommender systems, many based on the principle of learning from other participants’ data to make automatic predictions for an individual, are now commonplace for predicting the most relevant video, music, and digital content for individual users [28, 88, 95]. Researchers have proposed recommender systems to help users modify and understand their interfaces. For instance, after observing the user’s limited understanding of their smartphone’s accessibility features, Wu *et al.* explored approaches to provide contextual recommendations on setting adjustments [99]. Son *et al.* proposed an interface augmentation that easily suggests harmonious color combinations for users not equipped to describe subtle differences in colors [82]. For readers, font characteristics can be difficult to distinguish, and their effects on reading speed unclear. Current font selection interfaces present an overwhelmingly long list of font choices with limited guidance on selection [62]. Thus, selecting fonts is a promising task for recommender systems. Such a system can be based on users providing personal information, including their preferences and performance.

Users’ personal information can help improve the relevance of items returned by recommender systems. Teevan *et al.* found that when performing a web search with the same query, results deemed relevant to some may be irrelevant to others due to individual differences in judgment [88]. They found that results personalized based on the user’s explicit and implicit input may improve the recommender system’s accuracy [88]. However, even when users are well-equipped to specify their search queries fully, they may not be initially motivated to do so [87]. In studying virtual assistants, Pal *et al.* found that when perceived benefits outweigh the cost of sharing personal information, participants were more willing to provide information to improve the accuracy of recommendations [64]. Another study on book recommendations found that participants will provide personal information when a perceived improvement in recommendation quality is tangible [85]. Similar to previous work, our personalized font recommender uses participants’ personal information to provide more accurate recommendations.

2.3 Multifactorial Influences on Reading

Prior work has studied the effect of font characteristics or font family on reading speed [5, 7, 9, 13, 15–18, 67, 70, 73, 97], often using well-studied fonts, such as Arial and Verdana [5, 12–14, 46, 58, 97, 98]. Recent work incorporated newer fonts designed for digital screens, such as Roboto and Open Sans [90]. Additionally, based on Kadner *et al.*’s recommendation to include typography

¹<https://github.com/TianyuanCai/FontMART>

experts in-the-loop [47], we collaborate with typographers to select fonts well-suited for digital reading to explore the effect of various font characteristics on reading performance.

Many font characteristics may influence reading outcomes. Font size is a common parameter manipulated in readability studies. Prior research shows that larger characters help improve digital readability [12, 18, 73]. Ohnishi *et al.* and Oderkerk *et al.* respectively found that fonts with heavier weight and wider characters better support letter recognition [61, 63], a proxy for the readability of a font [50, 51]. Other characteristics studied in the past include character width [3, 56], character spacing [11, 86], stroke contrast [10, 11] and page background [43]. Suffice to say, influences on reading are multifactorial. Thus, it is important to capture and consider a variety of factors when making font recommendations. To ensure the accurate extraction of multiple font characteristics, we developed FontView, a web application that traces the path data of the vector font graphics.

Research has also shown the importance of considering multiple font characteristics simultaneously. For instance, when considering character width in conjunction with font size, Arditi *et al.* found that fixed-width² characters facilitate faster reading for small font sizes, but variable-width characters work better for large font sizes [3].

Because font characteristics can have different effects on different reader populations, font and reader characteristics should be considered simultaneously. For instance, prior research found that larger character spacing and character width could improve reading outcomes among old and low vision readers [11, 100, 104]. Conversely, recent eye-tracking research on participants shows that condensed fonts with tight character spacing can increase reading speed, but character width manipulations had no effect [56]. Our work explores the interaction between reader and font characteristics qualitatively through typographer interviews and empirically through model explainability.

3 TYPOGRAPHIC CONSIDERATIONS

The study of font readability lies at the intersection of science and design. To capture design considerations, we interviewed five typographers. These interviews yielded: (1) a selection of fonts for our remote readability studies, (2) a set of font and reader characteristics most relevant to predicting reading speed, and (3) methods for extracting font characteristics in our FontView tool. We conducted semi-structured 45-minute interviews with each typographer: **P1**: Male, 30 years of experience, typographer and founder of a typography consultancy; **P2**: Female, 25 years of experience, typographer and Professor of Design; **P3**: Male, 20 years of experience, typographer at a large corporation; **P4**: Male, 18 years of experience, typographer at a large corporation; **P5**: Male, 12 years of experience, typographer at a large corporation.

3.1 Font Selection for Readability

There are hundreds of thousands of digital fonts available [30], and of those designed for digital reading, only a few are consistently available across devices and applications. The set of available fonts per device and application often differ, and will change over time as

²Fonts with fixed-width have no variation in character widths among their letters. Conversely, the character widths in variable-width fonts vary.

new fonts are developed. Our goal was to select a set of fonts spanning different attributes and anatomies. We limited our selection to eight fonts to facilitate a within-subjects design and maintain a palatable study length of 30 minutes on average. Based on prior discussions with typographers and reading experts, we chose the following three criteria to guide font selection:

- **Prevalence:** To support the applicability of the results in the wild, we select fonts commonly used for English reading.
- **Availability:** To ensure that our recommendations and insights are readily applicable to readers in the wild, we select fonts that are freely distributed or available on both Windows and macOS operating systems.
- **Diversity:** To explore how the variation in font characteristics influences readers differently, we select fonts that differ along various typographic dimensions.

Before the interviews, we worked with typographer P1 to select an initial group of eight fonts: Montserrat, Open Sans, Arial, Roboto, Merriweather, Georgia, Source Serif Pro, and Times. To facilitate the convergence of selections, we ask the other four typographers to independently validate the selection by P1. All typographers agreed that our initial selections satisfy our criteria, but P1 and P4 believed Montserrat to be less readable than other fonts because of its larger character width, taller x-height, and tighter character spacing. Both suggested Poppins as a more readable Geometric sans serif than Montserrat. Figure 1 features our final font selections.



Figure 1: Illustration of our font selections in the same nominal size. Merriweather and Poppins are the largest fonts in serif and sans serif families, respectively.

3.2 Font and Reader Characteristics

Another goal of our interview with typographers was to identify font and reader characteristics that could be features in a Machine Learning model predictive of reading speed. Such collaboration with domain experts to develop predictive models and recommendations has been successful in other domains such as personal informatics, health, and recommender systems [32, 44, 49, 92, 94]. Fonts have various qualities that improve or hinder reading performance. The effect of these qualities may vary by reader. Our interviews followed P1's suggestion to discuss digital readability in

general instead of preference or speed specifically. P1 explained that a typographer’s goal is often to design a digital font that is readable across many scenarios. These interviews focused on what factors typographers believe may influence the digital readability of body text. We performed structural coding [55, 60] on the interviews to analyze common patterns. We converged on a set of font and reader characteristics that typographers indicated would influence digital readability. We use these characteristics as features in our population analysis and development of FontMART in later sections, eliciting quantitative evidence on how these features affect digital reading speed. The Discussion section includes follow-up interviews with typographers on how such quantitative insights may contribute to their current workflow.

3.2.1 Font Characteristics. Readable fonts help readers differentiate glyphs (P1, P4, P5) and pace the delivery of information (P1). Among the range of font characteristics discussed, character spacing, x-height, weight, and grayscale are the ones most typographers agreed influence readability (see Appendix §F for definitions).

Serifs. Our typographers believed that sans serif and serif fonts would be equally readable (P1 – P5) so long as they “adhere to the currently accepted standards of [character] spacing, stroke contrast, weight, proportions” (P4). Additional classification exists for serif and sans serif families. Stroke contrast, or the contrast between the thickness of strokes, helps differentiate these classes and is considered important for readability (P4).³ P3 emphasized the importance of avoiding extreme stroke contrast. Based on conversations with typographers, we measured the stroke contrast using the thinnest and the thickest strokes of the character “o” (P1, P2).

Spacing. All typographers recommended avoiding fonts with narrow character spacing. Wide character spacing helps distinguish characters and limits the number of words per line (P1). Therefore, it is especially beneficial for those experiencing reading difficulties or reading on smaller screens (P1, P4). P5 recommended avoiding overly wide character spacing. Citing the advantage of wider character spacing, P1 and P4 supported the use of monospaced fonts for the study, referencing their popularity in programming and movie scripts (P1). However, others believed that monospaced fonts are uncommon for typical digital reading, and the results may not be widely applicable (P2, P3, P5). P3: “Monospaced typefaces feel so different from everything else. It’s almost a decorative style. From this point of view, I would be more interested in judging monospaced-ness, as a quality.”

Character Size. Four typographers discussed the importance of character width and height for readability. In general, a tall x-height⁴ is good for reading (P1, P3, P4), although one typographer believed that this trend may be a matter of fashion (P2). P1 argued the need for a “healthy proportion of lowercase to uppercase [...]”

³For instance, serif fonts generally have more stroke contrast than sans serif fonts, i.e., their strokes vary more in thickness throughout their characters. Within serif fonts, neoclassical serifs are characterized by more dramatic stroke contrasts, while glyphic serifs often see minimum stroke contrast.

⁴x-height may refer to the average height of lowercase characters or height of the “x” character. In some cases, they are also used synonymously with proportion, the ratio of lowercase to uppercase heights. During the interviews, typographers referred to x-height as the general height of lowercase characters, and we applied the same measurement approach in our FontView tool.

and that proportion depends on the design.” When discussing font character width, typographers advised staying away from the extremes, “not too narrow or too expanded” (P5). Among our eight fonts, those with larger character widths also have taller x-heights.

Weight. Weight characterizes the relative thickness of a font’s strokes (P1) [35]. Although four typographers discussed the importance of weight, none discussed if less or more weight benefited a reader. P2 did caution against using extreme values.

Grayscale. P4 and P5 referred to grayscale as the proportion of opaque to transparent pixels when rendering alphabets. Both mentioned grayscale as a factor to consider due to its influence on eye movements. P4: “If you have, for instance, a book and see a page, and it has an even texture, that’s nice. But if it has a spotty texture, that’s something that can be distracting because your eyes will automatically go towards those heavier spots.” While we do not specifically measure the effect of grayscale on eye movements, we nonetheless include the proportion of displayed to empty pixels as a font characteristic in our work.

Other characteristics. We additionally controlled for the standard deviation of character width as a font characteristic because it has been shown to affect readability by prior research [3]. Typographers also mentioned several other characteristics that may influence the reading experience, such as counter and rhythm. However, there is a lack of consensus on how they can be measured. We include the typographers’ comments on them in Appendix §B.

3.2.2 Reader Characteristics. None of the typographers believed that there is a universally most readable font for all (P1 – P5). Therefore, we asked them to elaborate on the potential effects different font characteristics exhibit on different readers.

Reader Age. Typographers emphasized the importance of adapting font selections to reader age because the effects of character size and weight are expected to differ. Older adults tend to have weaker and more variable eyesight and may have difficulty identifying glyph details (P2, P4, P5). Some font characteristics, such as heavier weight, may benefit older readers while their effects on younger readers may be unclear (P2).

Reading Devices. The typographers interviewed believed that the “difference between the specific devices doesn’t really require a font change” (P2, P4, P5). The biggest effect of the device is on the visual size of the fonts (P2, P4, P5). Specifically, reading on mobile phones may expose readers to smaller fonts than reading on desktop devices (P5). Readers compensate by bringing the device closer to their eyes (P4). Other device-related effects include the rendering algorithm (P1), screen resolution (P4), and font hinting⁵ (P1, P4, P5). However, typographers agreed that these factors may be challenging to control (P1, P5).

Font Familiarity. Past familiarity with serif or sans serif fonts may improve reading performance in the respective fonts (P1, P3, P4, P5). Familiarity stems from an individual’s handwriting (P4) and typical reading material (P1, P3, P5). P4: “If they learn longhand writing in school, they will probably be able to read serifs [better].”

⁵Font hinting refers to the programming instructions made to a font’s outline to help the letters fit on the pixel grid for digital display.

But if they don't, if they only learn block letters, it will probably be easier to read sans serifs." Print publications, such as books and newspapers, predominantly use serif fonts. P5: "If you're more like a book person, you might be more comfortable with serif typefaces."

3.3 FontView: Extracting Font Characteristics

From our interview with typographers and past work, we learned that the metrics contained in font files, the OS/2 tables, often do not accurately reflect font characteristics (P3, P5, [8]). Therefore, we developed a web application, FontView, to quantify the typographical features discussed during the interviews. A web application is necessary because many font metrics are not easily discoverable, nontrivial to measure by hand, and non-replicable if not measured in a consistent computational way. For instance, the average character spacing requires many repeated measurements covering multiple character combinations, and stroke contrast require exact tracing of the font's vector path.

We measured font characteristics by programmatically tracing the path data of the vector font graphics with `OpenType.js`⁶, an open-source font parser that allows in-browser access to letterforms. An alternative method is to extract these characteristics from rasterized images. However, this method can result in incompatible measurements on different operating systems (P5).

FontView can analyze any English font to compute each font characteristic in pixels. We validated FontView's calculation with typographers (P1 – P5) and include the detailed approach in Appendix §F. Figure 2 provides an example of how the font characteristics are quantified, and Figure 3 shows their corresponding values for each of our eight fonts. In this work, we used the default font size (16px) of modern browsers such as Chrome and Firefox [12, 27, 57].

4 CROWDSOURCED READABILITY STUDY

4.1 Study Design Considerations

Many past readability studies from the Human-Computer Interaction community have adopted in-person study design methods similar to the foundational work by Boyarski *et al.* [18]. These studies provided initial evidence on the relationships between reading speed and conditions such as font family and font size. They often included 20–100 participants, who were asked to read passages with different fonts in lab settings [12, 13, 26, 39].

We conducted our study remotely by recruiting paid crowdworkers from Amazon's Mechanical Turk (MTurk). We model our study design on several recent studies that have successfully recruited paid crowdworkers and volunteers to conduct remote readability tests focusing on reading speed [53, 90, 93]. In our study, participants used their device of choice in their natural, everyday reading environment. Two factors mainly influenced these design choices. First, the COVID-19 pandemic made it challenging to run a readability study in person. Second, a remote study allowed us to recruit enough participants to train a personalized font recommender.

We used eighth-grade level reading passages and comprehension questions from past work⁷. See Appendix §D for the description of reading passages used. Prior research used these passages for

similar remote readability tests with paid crowdworkers. Their results showed that the passages' topic, and the readers' familiarity and interest did not affect reading speed results [90, 93]. There was also no relationship found between font and comprehension. The authors stated that this might allow participants to read as fast as possible while retaining comprehension because the material is easy enough to comprehend at this eighth-grade level [90, 91].

4.2 Study Methods

Our remote study was conducted by adapting an open-source web application from prior work [90, 91, 93]. The reading interface has a fixed width of 420px across all devices. Each reading passage is split in four consecutive sections, or screens, to ensure readers do not have to scroll to read the text. Prior work has suggested using multiple screens to display text, to capture more robust performance measurements [90]. Fonts are loaded from our web server to ensure a consistent user experience. The web application monitors and resets the participant's browser zoom levels to ensure consistent zoom level and font size across participants. The web application collects participants' reading speed measurements remotely. Prior research has successfully collected this type of data using JavaScript [33, 69], validating that the response times collected in web browsers are reliable [34, 78].

Our study features a pre-survey, study overview instructions, a practice round⁸, the main study, and a post-survey. The main study includes an instruction screen and a font preference test, followed by another instruction screen and reading speed and comprehension tests per font. The main study features all eight fonts for a within-subjects study design, following the recommendations of Wallace *et al.* [90]. Figure 4 illustrates the study design.

4.2.1 Pre-Survey. The pre-survey asks participants to self-report their demographics (age, education, native language), reading experience (frequency, type of content, device of choice), reading speed (7-point Likert scale), vision (normal, corrected), disabilities (learning, medical, and reading-related), current substance influence (drugs, medications, alcohol), and current reading environment (lighting, time of day). Pre-survey questions are in Appendix §C.1.

4.2.2 Study Overview Instructions. After completing the pre-survey, participants see an instruction screen that overviews the entire study. The main study consists of several parts. The first part is a warm-up, followed by eight short rounds of readings and comprehension questions. Participants are instructed to read as quickly as possible during the reading sections without reading aloud or re-reading, but be prepared to answer comprehension questions about the reading. Participants may take breaks during the instruction screens before each reading round.

4.2.3 Font Preference Test. Instructions inform participants they will be asked to choose what font would be easiest to read in from several pairings by toggling between choices and then selecting their preference. They can take a break if they need to.

⁶<https://opentype.js.org/>

⁷Passages and questions available at <https://github.com/virtual-readability-lab/tochi-paper-materials-towards-individuated-reading/>

⁸The practice round uses four fonts: Comic Sans, Raleway, Lexend Deca, and Oswald.



Figure 2: Rather than relying on inconsistent and inaccurate font files, we extracted characteristics for all our study fonts in the same way. Visualized here are some of those font characteristics, using the font Source Serif Pro as an example. For example, grayscale was calculated as the proportion of space filled by pixels of lowercase alphabets. The letters shown are a subset of those used to quantify font characteristics. See Appendix §F for more details on the calculations.

METRIC	FONT								
	Weight	Grayscale	Stroke Contrast	Character Spacing	Standard Deviation of Character Width	Descender Length	Ascender Length	Character Width	x-Height
Georgia	0.22	0.46	0.24	0.42	2.08	3.68	4.50	7.54	7.70
Arial	0.19	0.42	0.08	1.20	2.27	3.41	3.25	6.64	8.30
Times	0.13	0.46	0.49	0.53	1.86	3.60	3.79	6.80	7.20
Open Sans	0.17	0.39	0.11	1.55	2.35	3.96	3.66	6.83	8.56
Merriweather	0.27	0.39	0.60	0.79	2.47	4.20	4.31	8.25	8.88
Source Serif Pro	0.13	0.44	0.27	0.71	2.07	4.02	4.32	7.44	7.60
Poppins	0.17	0.44	0.16	1.53	2.79	4.30	3.18	7.46	8.77
Roboto	0.19	0.45	0.04	1.45	2.23	3.43	3.63	6.54	8.45
Min	0.13	0.39	0.04	0.42	1.86	3.41	3.18	6.54	7.20
Max	0.27	0.46	0.60	1.55	2.79	4.30	4.50	8.25	8.88

Figure 3: Heatmap of quantified font characteristics. Each column represents a font characteristic. The changing shades of the heatmap highlight the diversity among our selection of eight fonts. All characteristics are measured in pixels (px) except for weight, grayscale, and stroke contrast. Only 2 decimal places are shown for visualization purposes.

A participant makes pairwise comparisons among the eight fonts during the preference test. A font is eliminated from further comparisons when the participant picks against it twice. Using a double-elimination tournament provides a method to find a participant's

most preferred font while limiting the total number of pairwise comparisons to a maximum of $(N \times 2) - 1$. To compare a pair of fonts, participants freely toggle between a text sample in each font before selecting the preferred font. In the end, participants complete repeat comparisons of 6 pairs of fonts as a measure of consistency.

4.2.4 Reading Speed and Comprehension Test. Before each round of reading, instructions inform participants to read four sections as quickly as possible, without reading aloud or re-reading, and being prepared to answer comprehension questions after reading. Breaks can be taken between reading rounds, during instruction screens.

Our study includes eight rounds of speed and comprehension tests, each with a randomly assigned font and reading passage. This design is similar to past reading studies[4, 53, 90]. Each reading passage is split across four consecutive sections (individual screens with 34–47 words each). A reading speed measurement is recorded after each section. To compute reading speed in words per minute (WPM), we recorded the time elapsed between when the text is first shown on the screen to when the participant clicks to continue to the next reading section. The button to proceed is initially disabled for 2 seconds to prevent accidental clicks. Each passage is followed by three multiple-choice comprehension questions and a mini questionnaire. Comprehension questions help motivate the participant to read the passage carefully. The mini questionnaire asks the participant about their reading technique, familiarity, and interest in the topic presented, using a 5-point Likert scale. The mini questionnaire is included in Appendix §C.3.

4.2.5 Post-Survey. The post-survey asks the participant to indicate their familiarity with each font using a 5-point Likert scale, their experience with the toggle interface, their perceived effects of the font change, and their willingness to use the font recommendations. Post-survey questions are in Appendix §C.2.

4.3 Data Collected

We recruited 500 participants on MTurk. Participants were required to have completed at least 100 tasks on MTurk with a minimum

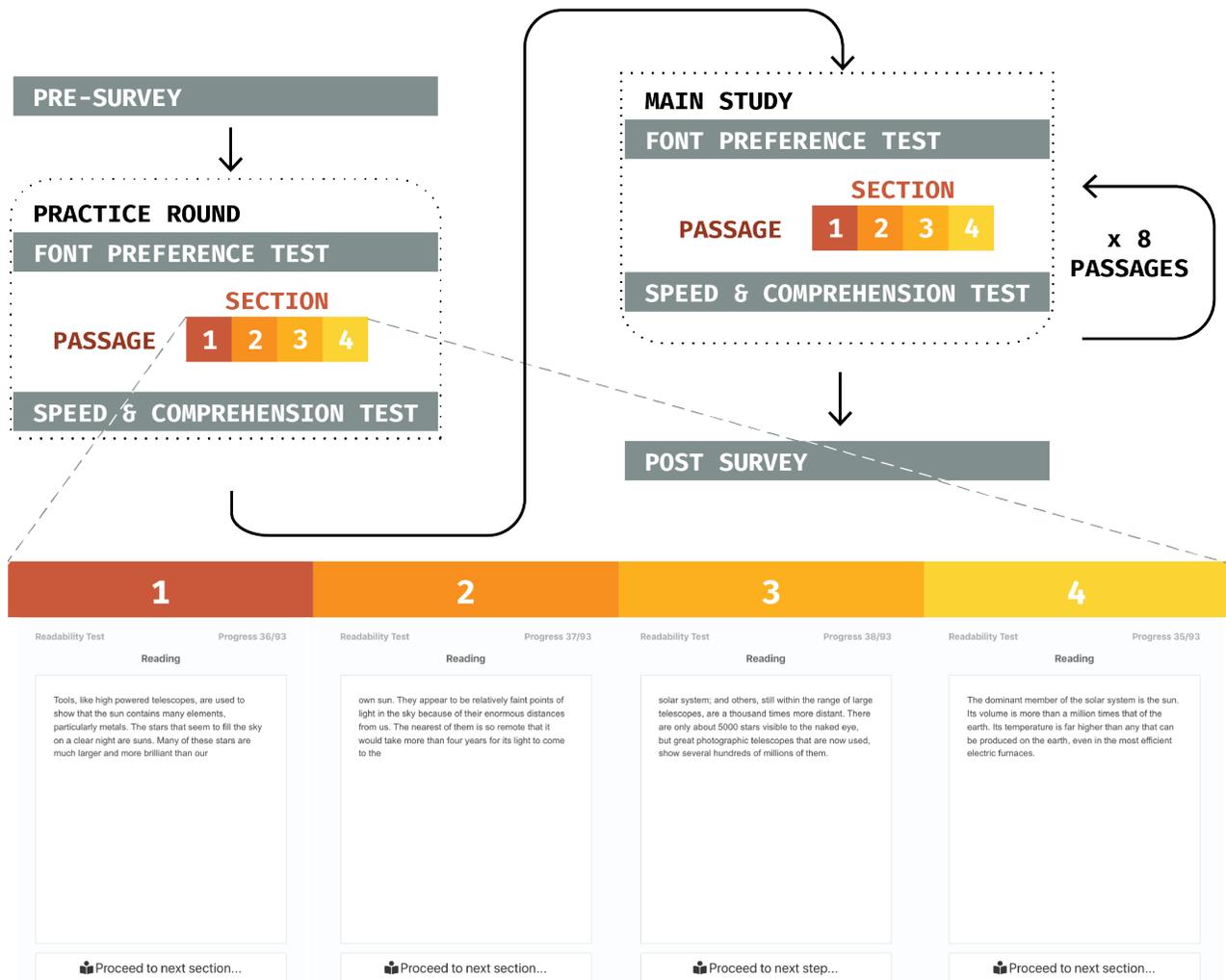


Figure 4: Illustration of the remote readability study. The top figure shows our study design. The bottom figure shows screenshots of four sections of a passage shown on consecutive screens during the study, presented in the study’s reading interface. Passages were split into sections to obtain multiple speed measurements per font. The practice round and main study start with the font preference test, followed by an instruction screen, and then the participant initiates the reading speed and comprehension test. Using each of the eight study fonts, participants read a passage, completed a three-question comprehension test and a mini questionnaire. The passage order and font presentation were randomized across participants.

approval rate of 95–99%. They were compensated US\$3.75. After applying our data removal methods in §4.3.1 we kept data from 252 (50.4%) participants. We collected over 5,000 font preference pairwise comparisons and over 8,000 reading speed measurements from these participants. We used this data for the population analysis (§4.4), evaluating the baseline approaches for font recommendation (§5), and training FontMART (§6).

4.3.1 Data Removal Methods. We only include participants whose behavior is indicative of normal reading. We removed participants if they met one of the following exclusion criteria: (1) did not submit

pre- or post-survey, (2) did not self-report being “very comfortable” reading in English, (3) self-reported being diagnosed with any reading or learning disability, medical or neurological condition, (4) self-reported being under the influence of any drugs, medications, or alcohol, (5) had an average reading comprehension score below 0.66⁹, and (6) nonsensical responses in pre- or post-survey, see Appendix §E. Similar to prior remote readability research [90],

⁹The reading comprehension score is calculated as the number of correctly answered comprehension questions out of the total of 24 questions. The cut-off was determined using the elbow heuristics [77, 89].

we removed individual speed measurements outside the range of 100–650 WPM (based on Carver *et al.*'s recommendations [24, 25]) to ensure participants were not memorizing or skimming the text.

4.3.2 Participants. Of the 252 participants (34.1% female) that remained after data removal, age ranged from 18 to 71 years (average = 34.2, standard deviation = 10.7): 3 were younger than 20, 103 were in their 20s, 86 in their 30s, 32 in their 40s, and 28 were at least 50. Participants took on average 35 minutes to complete this study. Participants completed the study on their chosen devices: 39.7% used a desktop, 59.9% used a laptop, one participant used a tablet, and no participants used mobile phones¹⁰.

4.4 Population Analysis

If the influence of fonts on reading speed is unmoderated by individual differences between readers, then the same font can be recommended to all. A between-subjects effect would be visible in the results of a parametric linear model. We used Linear Mixed-Effects models (LMEs) to measure whether individualized effects of fonts exist for reading speed.

In this approach, data were first aggregated to average reading speed measurements in WPM per participant and font. We then constructed an LME model to predict reading speed, with font and age as fixed effects, and participant ID (identifying each participant) and passage ID (identifying each of the eight passages) as crossed random effects. Age is well known to influence reading speed [23] and is included as a covariate. The inclusion of random effects produces a hierarchical model by creating separate intercepts for each value of participant ID and passage ID, reflecting that participants read at different speeds and that each passage may have a different mean reading speed.

Our LME model indicates that age significantly affects reading speed (reading speed slows by 1.6 WPM per year; $\chi^2 = 8.4$; $p < 0.01$). The model does not demonstrate a significant effect of the font, indicating that fonts lack a uniform effect across participants. One could interpret these results to mean that font has no impact on reading speed under the present study's conditions. However, font is well known to moderate reading speed under many similar conditions [5, 7, 9, 13, 15–18, 67, 73, 97]). Therefore, the lack of statistical significance most likely suggests that *a font may affect participants differently depending on their individual characteristics.*

We constructed an alternative model by replacing the individual fonts with serif vs. sans serif indicators as the fixed effect. This model found no significant difference in reading speed between serif and sans serif fonts (290.8 WPM vs. 289.5 WPM, respectively). This result is consistent with the expectations (P1–P5) established from the earlier typographer interviews in §3.2.1.

5 FONT RECOMMENDATION BASELINES

Font choices vary across platforms and software, providing readers with an evolving set of choices. How does a reader find their fastest font? They could hypothetically try reading in many different fonts in hopes of finding their fastest. However, this would require a considerable time investment. Before presenting our personalized

recommender model, we consider straightforward alternatives that could simplify the font selection process for readers.

5.1 Preferred Font

We first consider the alternative of having participants select their preferred fonts, by completing pairwise comparisons as part of the Font Preference Test of the study (§4.2.3). To derive a participant's preference per font from the pairwise comparisons, we computed an Elo rating [40] per font per participant. Since participants do not see every possible comparison, Elo ratings provide a method to determine preference in this scenario, by accounting for the strength of each font per pairwise comparison [42] and mitigating cold-start problems common in recommender systems [101]. A higher Elo rating reflects more wins in the tournament and characterizes the participant's level of preference for that font.¹¹

Do participants read the fastest in their preferred fonts? In our study, 76/252 (30.2%) participants read the fastest in their preferred font. However, the most preferred fonts were on average 71.2 WPM slower than the empirically fastest fonts as measured by reading tests. As a frame of reference, the average difference between a participant's fastest and slowest font is 159.04 WPM.

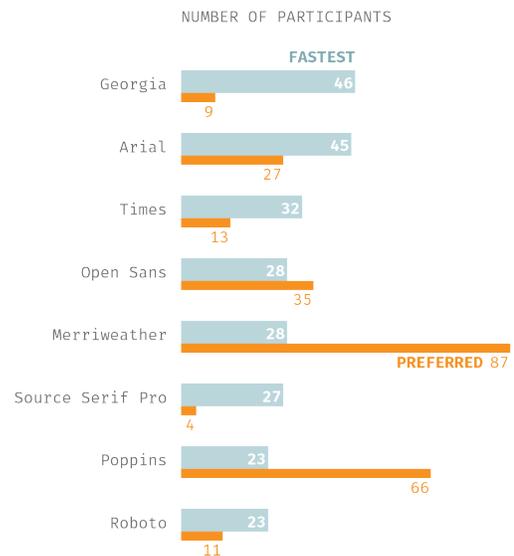


Figure 5: Distribution of the fastest and preferred fonts. Each bar represents the number of participants who preferred or read the fastest in a particular font. Many participants prefer to read in Merriweather and Poppins, the largest fonts in serif and sans serif families, respectively. On the other hand, the distribution of the fastest fonts is more uniform, with Georgia and Arial showing a slight advantage.

¹⁰Our web server detects a participant's device from their http request object using the library: <https://www.npmjs.com/package/express-device>

¹¹We computed Elo ratings using a starting value of 1500 and a higher than standard K value of 64 to account for fewer pairwise comparisons.

Which font characteristics do people prefer? People’s preferred fonts are often different from the fastest ones (Figure 5), leaving an opportunity for font recommendation approaches to offer improvements. Participants tend to prefer fonts with taller x-heights, a font characteristic that stands out when visually comparing fonts. 87/252 (34.5%) of participants preferred to read in Merriweather and 66/252 (26.2%) in Poppins, two fonts with the largest x-height in the serif and sans serif families, respectively. On the other hand, the fastest font varied more across participants. Georgia was the fastest font for 46/252 (18.3%) participants, and Arial for 45/252 (17.9%) participants, followed by Times and Open Sans. Prior remote readability work similarly found that preference was not predictive of reading speed, providing further evidence for the need to personalize font recommendations to the individual reader [90].

5.2 Random Font

A reader may choose the default font selections of designers, publishers, or software. While we cannot perfectly replicate this experience, we approximate it with the eight fonts we selected in collaboration with typographers. For each reader, we compare their performance in a randomly chosen font, from our selection of eight, to understand whether a randomly selected font outperforms the font recommended by FontMART.

5.3 One-size-fits-all Font

None of the typographers we interviewed believed there is one font that is fastest for all readers. Ideally, if this were true, a personalized font recommender would not be necessary. Prior research agrees with this idea [47, 93]. Our within-subjects study design allows us to compare FontMART’s recommended font with the other seven fonts each participant reads in. If any single font from our study could consistently outperform FontMART’s recommendation, this would indicate that there is no benefit from the consideration of reader characteristics during font recommendation. In such scenario, domain experts could hypothetically choose the fastest font(s) for everyone. We denote this baseline as the “one-size-fits-all font” — the single font that a hypothetical domain expert might choose to speed up all readers.

6 FontMART: A PERSONALIZED FONT RECOMMENDER MODEL

Discussions with typographers and prior work show that matching participants with specific fonts can help them read faster [93]. To provide personalized font recommendations, we introduce FontMART, a learning to rank model that is a derivative of LambdaMART [21]. FontMART learns to associate fonts with reader characteristics and can thus be used to order fonts by predicted reading speed for the individual reader. Instead of participants engaging in lengthy reading speed tests to find their fastest fonts, FontMART can directly use a reader’s characteristics (i.e., age, self-reported reading frequency, self-reported reading speed, and font familiarity) to predict their faster fonts. The self-reported measures have been used in previous studies [90], and are characteristics that are quick and easy to collect in the wild.

6.1 Model Design

Our goal is to rank a set of fonts by predicted speed for an individual participant. Among the available learning to rank models, we chose LambdaMART for its state-of-the-art performance, interpretability, and use in other Information Retrieval applications [21]. LambdaMART learns to associate documents with queries by training on relevance labels. LambdaMART takes as input (document, query) pairs and predicts relevance scores. These predicted scores can be used to rank a set of documents for a given query by relevance. LambdaMART is composed of a sequence of weak tree learners and optimizes the model with ranking metrics, such as normalized discounted cumulative gain (NDCG).

We adapt this modeling approach to the problem of font recommendation. Specifically, our FontMART model¹² takes as input (reader, font) pairs and outputs scores representing the relative speed a participant is predicted to achieve in a given font. We do not predict reading speeds directly due to noise and individual differences (see below). The input to our model includes the reader and font characteristics listed in Table 1 and detailed in Appendix §F. We use 50-fold cross-validation for model evaluation due to small sample size. During cross-validation, observations from each participant are contained in a single fold to avoid cross-contamination.

Feature Group	Feature Name
Font Characteristics*	Weight
	Stroke Contrast
	Ascender Length
	Descender Length
	x-Height
	Character Width
	Standard Deviation of Character Width
	Character Spacing
	Grayscale
Reader Characteristics	Age
	Self-reported Reading Speed
	Self-reported Reading Frequency
	Font Familiarity

Table 1: Features used as input to the FontMART model. See Appendix §F for detailed descriptions. Some features collected during the pre-survey are used as filters or removed due to high correlation with others. *Note that for the purposes of our study, the font characteristics are fixed and constant across all readers, so that the model makes a prediction given a reader’s characteristics and the font characteristics of the eight fonts we selected for this study.

Converting participant reading speeds to training labels. Participants’ reading speeds vary significantly, so training a model to predict the reading speeds directly would be impractical. Instead, we convert the raw reading speeds from our crowdsourced readability study to labels that can be compared across participants. We experiment with two possible labeling schemes: graded and binary

¹²The Python package LightGBM provides an implementation of LambdaMART: <https://github.com/microsoft/LightGBM>

labels. We create graded labels¹³ by dividing a participant’s speed in a particular font by their maximum speed across the fonts tested and rescaling so that each font is assigned an integer label from 0 to 10. This achieves the effect of remapping all participants’ reading speeds to the same range. The graded labeling scheme provides the model with granular information about the relationship between fonts. However, uncontrollable environmental distractions during the remote readability tests may introduce noise at the individual participant level [83]. On the other hand, using binary labels may mitigate the model’s tendency to overfit the noise on this relatively small dataset. Therefore, we also use the following binary labeling scheme¹⁴: fonts with WPM within 10% of the participant’s best WPM are labeled with a 1, and the rest of the fonts are labeled with a 0. This threshold was experimentally selected. See Table 4 (Appendix §H) for a demonstration of the labeling scheme.

Trained with either labeling scheme, learning to rank models predict the relative reading speeds for different fonts and readers, and we use them to rank fonts for individual readers and then recommend them their predicted fastest fonts. If multiple fonts receive the same prediction, we use random tie-breaking to select a single recommendation.

6.2 Evaluation

There are no established computational baselines that our ranking metrics can compare to, so we have included our cross-validated results in rank-based measures in Table 6 (Appendix §I) in the interest of providing benchmarks for future research. We also compare our model’s predictions of the fastest font to our previously introduced baselines: (1) the participant’s preferred font (§5.1), (2) a randomly chosen font (§5.2), and (3) a one-size-fits-all font (§5.3). Compared to the three baselines, our model provides speed improvements on average. FontMART exhibits more substantial improvements when trained using binary labels. Figure 6 shows that FontMART with binary labels provides an average improvement of +25.6 WPM relative to a participant’s preferred font. When evaluated against the one-size-fits-all font baselines, the improvements from the best defaults, Arial and Georgia, are +14.8 and +14.4 WPM, respectively. This result shows that the consideration of reader characteristics alongside font characteristics is indeed important for making font recommendations to improve reading speed. Importantly, these improvements speed up a participant’s reading with minimal impact on their comprehension (recall that we removed participants with an average comprehension score below 0.66%). As a practical point of reference, an improvement of 10 WPM translates to around 600 additional words per hour, or one additional single-spaced letter page with a 1-inch margin and 16px font.

6.2.1 Recommended Fonts that Increase Reading Speed. To understand the effect of personalization, we first examine the general distribution of the top-recommended fonts in Figure 7. Arial and Georgia are among the two most recommended fonts by the model. This observation is unsurprising considering that Arial (45/252 participants) and Georgia (46/252 participants) frequently occur as the empirically fastest fonts in the crowdsourced readability data used

for training our model. Without personalization, these two fonts would be the best one-size-fits-all candidates, as shown in Figure 6.

Slicing the recommendations by quartiles of age reveals that participants 40 and older receive recommendations for Georgia more than the younger participants, and the latter receive more recommendations for Arial and Poppins. This trend is similarly observed in the distribution of the empirically fastest fonts (Figure 8).

6.2.2 Factors Affecting Reading Speed. To provide personalized recommendations, FontMART predicts relative reading speed based on a combination of reader and font characteristics. Our results show that font characteristics that help some readers read faster may not work for others. Figure 9 shows the feature importance of the font and reader characteristics that support FontMART’s predictions. The higher values for reader characteristics reflects the effectiveness of personalization over one-size-fits-all solutions. The font preference data was not included in the training data because our experiment did not find it improving the model performance. See Appendix §J for detailed explanation.

Shapley value is a helpful tool to understand which combination of font and reader characteristics makes a font faster to read for a participant. Specifically, a Shapley value reflects the direction and magnitude of influence the given reader and font characteristics have on the predicted relative speed of the font [54, 79].

We focus the Shapley values analysis on the font characteristics with a large influence on prediction scores when considered alongside reader characteristics, including font x-height, weight, descender length, and familiarity, as shown in Figure 10. We also note that the findings from Shapley values are specific to our selection of eight fonts. Future work is needed to understand the results’ generalizability to fonts varying within a wider range. Within our font selections, we find that relatively shorter x-height and heavier weight often make a font relatively faster to read per participant, as shown in blue. When considered in combination with reader characteristics, we find that heavier weight especially benefits older readers, as shown by the larger magnitude of Shapley values. On the other hand, weight alone has minimal impact on the reading speed of those below the age of 28. When all fonts are rendered at 16px, shorter x-height (Georgia and Arial), i.e., a greater difference between lowercase and uppercase characters heights, especially benefits participants below 21, and those self-reported to be slower or less frequent readers, see Figure 11. Although older participants similarly benefit from shorter x-height, overly short x-height should be avoided for those above 48 (Times and Source Serif Pro). While longer descenders work well for readers below 22, others tend to benefit more from fonts with shorter descender lengths. No consistent trend exists for the effect of font familiarity. “Moderate familiarity” is associated with improved reading speed, while “extreme familiarity” is correlated with the opposite effect. We do not have enough evidence to make claims about familiarity’s effect on font and reading speed because of the limited way in which we measured familiarity (Section §8).

6.2.3 Participant Attitudes Towards Font Recommendations. Our study found that only 76/252 (30.2%) participants selected their fastest font solely based on preference among the eight study fonts. The disparity between the preferred and fastest fonts is consistent with prior work [90]. Given this disparity, would readers be

¹³Given participant i , $10 \times \text{current speed}_i / \text{maximum speed}_i$.

¹⁴Given participant i , $\mathbb{1}_{[S_{\text{font}_i} / S_{\text{max font}_i} > 0.9]}$.

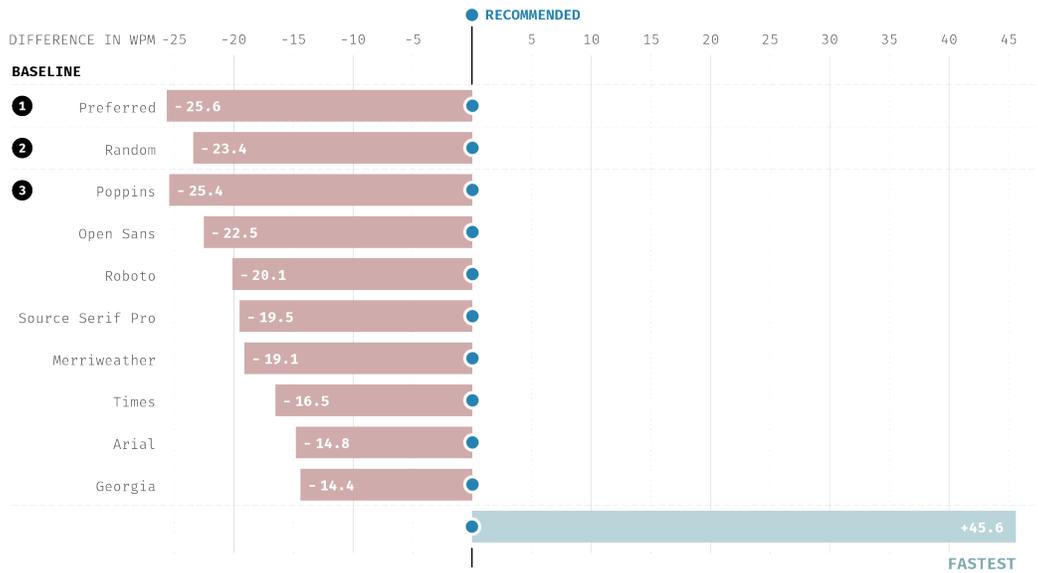


Figure 6: Reading speed of various baselines relative to the fonts recommended by FontMART, where each bar represents the potential improvement gap. The participants’ preferred fonts perform at a similar level as a randomly selected font. The gap in reading speed between using the empirically fastest and the preferred font is 71.2 WPM. FontMART’s recommended font bridges 25.4 WPM of that gap. While FontMART provides reading speed improvements compared to the various baselines, our results show that future improvements are possible to continue to bridge the gap in helping readers find their fastest fonts.

open to using a personalized recommender to find a faster font? In the post-survey, 227/252 (90.1%) of participants believe that changing font characteristics could help them read faster, and 217/252 (86.1%) would trust computer applications for such tasks. While these results are promising, it is important to note that our paid crowdworkers chose to take a readability study, and may be more motivated to improve their reading.

7 DISCUSSION

7.1 Supporting Faster Reading with Quantitative and Qualitative Evidence

Interpreting our quantitative and qualitative results may help typographers and designers refine their designs to optimize reading speed for specific populations. This section’s discussion is derived from our initial interviews with typographers in Section §3.2 and quantitative results from Section §6.2.2.

Weight: While four of the typographers in our interviews (P1 – P4, §3.2) discussed the importance of a font’s weight (i.e., the relative thickness of a font’s strokes), the discussions were brief and no specific recommendations were made. P3 mentioned that weight is “unexplored territory”. Our quantitative results show that among our eight fonts, those with heavier weight were associated with faster reading speeds, consistent with findings from past research [38, 63, 81]. While none of our fonts had overly thick strokes per P2’s recommendations, the fonts with the thinnest strokes,

Times and Source Serif Pro, had the most negative effects on reading speed. For participants above age 35, our results show that increased/decreased weight leads to more considerable changes in reading speed. In contrast, weight had minimal effect on readers under the age of 30.

x-Height: The typographers we interviewed believed that tall x-height is essential to improving readability (P1, P3, P4). Other typographers have also designed fonts with taller x-heights to increase readability. For example, Georgia and Verdana were designed to have a taller x-height than Times New Roman [18]. In our results, different reader characteristics matched better with different x-heights. For participants above age 35, our results agree with typographers’ views that fonts with shorter x-heights (Times and Source Serif Pro) negatively impact reading speed. However, participants self-reporting lower reading frequency and speed also benefited from fonts with shorter x-height (Times, Source Serif Pro, Georgia, and Arial). To expand on our results, future work could focus on fonts with more extreme x-height values while controlling the other characteristics.

Character Spacing: While typographers we interviewed agreed that narrower character spacing would result in worse readability (P1 – P5), we found it to support faster reading, such as in the case of Georgia. Past research showed that narrower character spacing might increase reading speed, indicated by longer fixation durations and fewer saccades [3, 56]. In addition, Tai *et al.* discovered that the reader’s recognition of *high-frequency* words may be unaffected

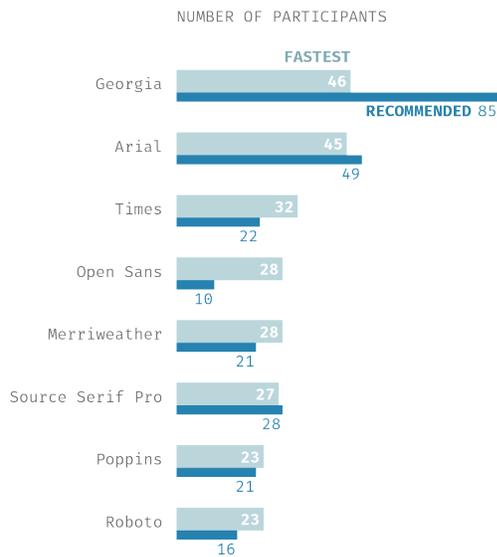


Figure 7: Distribution of the predicted and fastest fonts. Each bar represents the number of participants who received recommendations or perform the fastest in a particular font, out of a total of 252 participants. We observe that the distributions of the recommended and empirically fastest fonts are similar. Our recommender suggests Georgia more often than the observed trend.

by manipulating character spacing even when it is adjusted to be narrower than the default [86]. Therefore, our use of the eighth-grade reading material may not be difficult enough for narrow character spacing to impact reading speed negatively. It may also be the case that the differences in character spacing across our fonts were not significant enough to observe the negative effects of lower spacing.

In Summary: As illustrated, model explainability provides an opportunity to facilitate collaborations between typographers and a recommender like FontMART. Our use of the Shapley value visualization represents a step in that direction. Typographers may use our results (Figure 10, 11) to quantify how their designs could affect someone’s reading speed. In the section to follow, we interview typographers to explore such a possibility.

7.2 Supporting Interactions between a Font Recommender, Typographers, and Readers

Starting with a selection of readable fonts, FontMART can detect nuanced associations between reading speed and reader and font characteristics. This section discusses how recommendation systems like FontMART may serve as the bridge that improves the interface between typographers and everyday readers to enhance font design for readability.

7.2.1 Typographers: text designers. With a focus on readability, typographers may use our results (Figure 10, 11) to identify which

factors to consider during typeface design and quantify how their designs could affect someone’s reading speed. Following the study, we interviewed Typographers P1 and P4 again to understand how model explainability may help support typographers.

Both typographers agreed that the features FontMART uses for font recommendation (Figure 9) are ones they commonly consider important in the typeface design process (P1, P4). The typographers also agreed that the large effect of age compared to other characteristics matches their experience. Also, that x-height, character spacing, and stroke contrast are three font characteristics commonly considered during the font design process (P4). FontMART can identify relevant characteristics that could help typographers during typeface design. Exploring the interplay among these characteristics is an equally important feature for typographers (P1).

Typographers think that the insights from Shapley value visualization (Figure 11) can provide good recommendations to typographers when designing fonts for specific populations (P1, P4). For instance, when designing for the older readers, the typographers can reference the visualization to find the best range of font weight for this population. Both expressed interest in expanding the generalizability of the existing insights to more extreme fonts. For instance, while in the study’s selection of eight readable fonts, they agree based on their experience that fonts with heavier weights are easier to read for older readers, another font with excessive weight may disrupt this trend (P1, P4). In addition, P1 expressed interest in simultaneously exploring the effect of varying multiple font characteristics. Constructing a larger dataset with greater font variety is necessary future work.

Based on our results, typographers suggested the inclusion of other features to explore their impact on readers. For instance, P1 suggested the inclusion of counter and rhythm in future analyses and models (See Appendix §B). P4 discussed the need to explore the trade-offs between readability recommendations and a typographer’s goals for aesthetics and style (P4).

With a focus on readability, typographers could use our results (Figure 10, 11) to quantify how their designs might affect someone’s reading speed. For example, they may optimize a font’s weight and x-height to help older readers, who are more likely to experience vision loss. Such optimizations may also benefit readers who temporarily experience degraded visual conditions, such as those reading while walking or in a moving vehicle.

7.2.2 Readers: text consumers. Readers can benefit from using FontMART in the future by selecting fonts with characteristics that match their reader characteristics. For example, older readers might benefit from fonts with weights optimized for them. Furthermore, the additional data they provide after trying the recommended fonts could help improve model performance and support more accurate recommendations for others in the future.

To support such interaction, reader openness to novel interface experiences is key to adopting a personalized font recommender. To dig a bit deeper into reader sentiments, we interviewed 5 university students who completed our readability study¹⁵ about their willingness to improve their reading experience by manipulating fonts.

¹⁵We recruited these 5 university students separately from the crowdworkers who completed the readability study, and do not include their results in our main study.

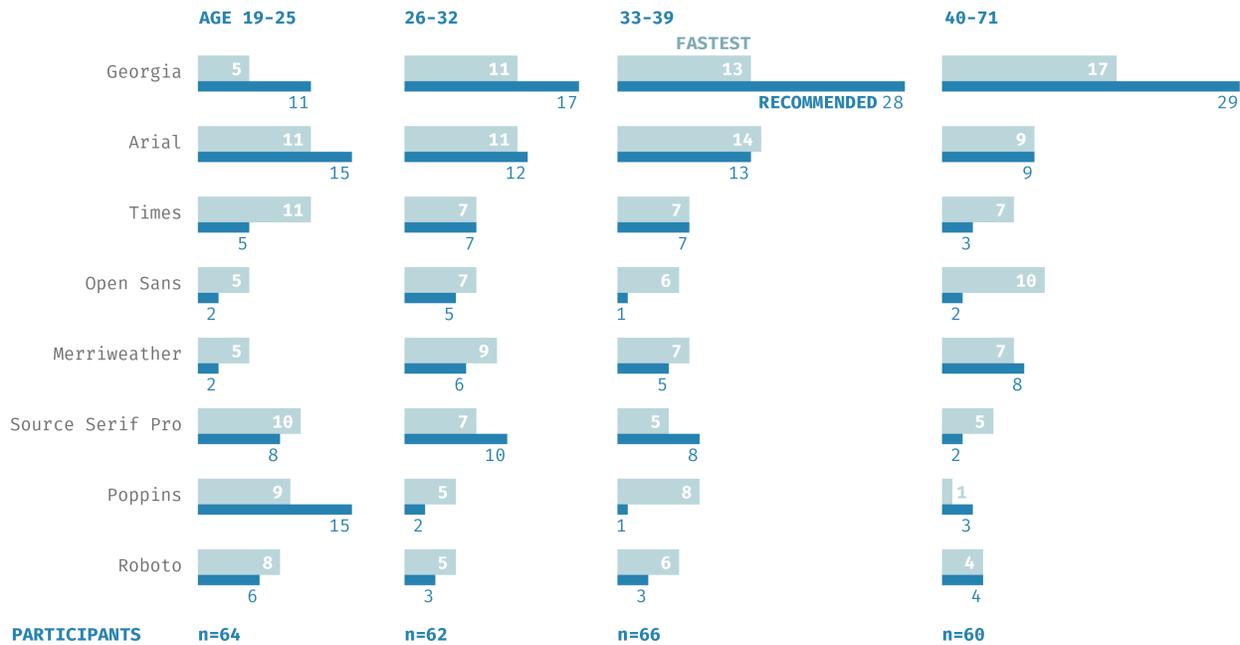


Figure 8: Distribution of the recommended fonts by age quartiles. Each bar represents the number of participants who received recommendations or perform the fastest in a particular font, out of a total of 252 participants. Georgia is recommended more often for the participants above the age of 40, and Arial and Poppins more often for the younger participants, especially those below 25. The numbers of Arial recommendations are similar across age groups. The font recommendations are varied otherwise. The distribution of the font recommendations is similar to that of the empirically fastest fonts. Each quartile contains a similar number of participants.

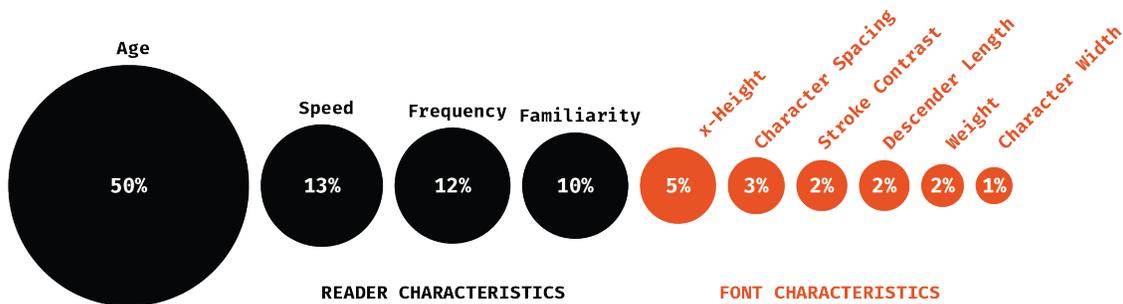


Figure 9: FontMART feature importance. The size of the circle is proportional to the relative importance of each reader and font characteristic to the model’s performance. Resonating with typographers’ views (P2, P4, P5, §3.2.2), a participant’s age is the most important factor for FontMART’s decision-making. Diversity of font characteristics is important, in contrast to the preference test results where participants predominantly preferred fonts with taller x-heights. Unimportant features are omitted from this visualization.

Overall, participants were interested in using applications to improve their reading speed, and they showed a willingness to modify their reading interface. Unprompted, two participants voiced their willingness to choose efficiency over preference if shown evidence

of better performance. One participant said: “I’m all about how to do things in the most efficient way to get the best outcome. So for the reading, if there’s a quick way to reformat things to make me more efficient, like hitting a few buttons and changing everything

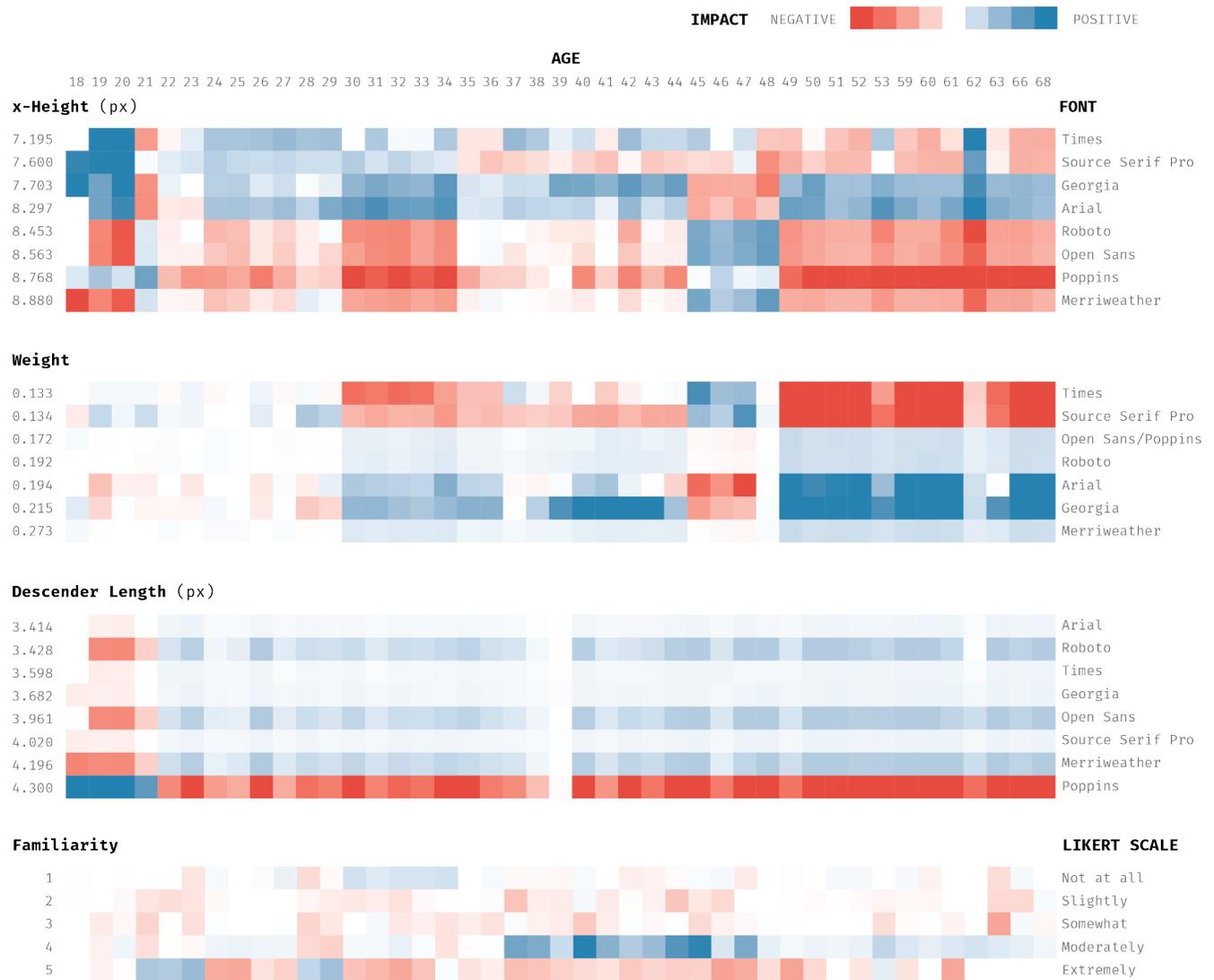


Figure 10: Shapley value of font characteristics by participant age. Across our eight fonts, those with shorter x-height and heavier weight are more often associated with faster reading speed, but the effects vary by age. Blue indicates a positive impact on the relative speed of the font for the participant, and red indicates a negative impact. Note that Open Sans and Poppins have similar weight values. The age distribution is sparser for participants 40 and older, resulting in noisier results. When multiple observations exist for a feature combination (a cell in the heatmap), the Shapley values are averaged.

perfectly, I would totally do it.” Another participant noted: “Cognitively, I wouldn’t enjoy reading in Open Sans, but knowing that it’s my fastest font, and given that I care more about my comprehension instead of looking good, I would be like, ok, give me whatever that’s fastest for me.” Both qualitative and quantitative evidence lends confidence to the idea of deploying a recommendation system in the wild. Qualitative evidence shows readers are open to using new reading fonts, and quantitative evidence shows it is possible to increase their reading speed (§4.4).

7.3 In-the-Wild Interactions with Font Recommenders to Improve Reading

The effectiveness of FontMART and the insights it provides could improve the design of reading experiences in the wild by automatically tailoring interfaces to match an individual reader’s needs. It may support readers in the form of a browser extension. By collecting the reader’s demographic information, the extension can update a web page’s body text with the model’s recommendation.

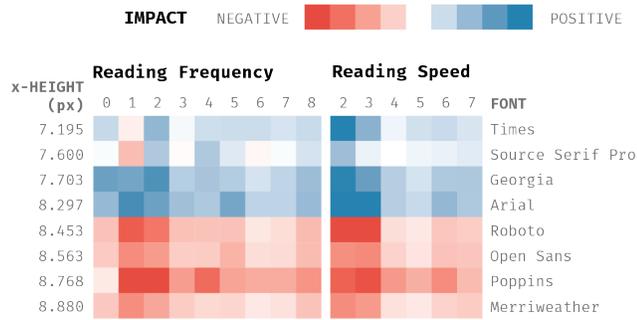


Figure 11: Shapley value of font characteristics by participant’s self-reported reading frequency and reading speed. Participants with lower reading frequency and self-reported reading speed tend to benefit more from fonts with shorter x-height. Blue indicates a positive impact on the relative speed of the font for the participant, and red a negative impact. The values are averaged when multiple observations exist for a feature combination (a cell in the heatmap).

With permission, the model may also refine its recommendation as it gathers more measures of the reader’s reading speed.

7.3.1 Font Recommendations by Age. If FontMART was deployed in the wild, collecting readers’ detailed demographic information could be challenging. In this scenario, just knowing a reader’s age could go a long way towards providing effective font recommendations. Based on our results, we provide the following general recommendation by age group:

Participants over 40 may benefit from using Georgia: We found that heavier weight and short x-height make Georgia especially suitable for participants 40 and older. Similar to typographers’ views (P2, P4, P5, §3.2.2), our model found strong interaction effects between age and weight. In addition to the empirical validation provided by the model, the typographers and past research confirm that the anatomy of Georgia is designed for digital reading (P1, [58]) and that serifs may help older participants more easily distinguish characters [1, 59, 76]. Our recommendation for this age group may be an overgeneralization because of the skewed age distribution of the participants. Future studies may offer more targeted recommendations for readers in different age groups over 40.

Arial has shown good performance across age groups: Based on our results, Arial could be a reliable default font choice for reading material developed for all ages. Similar to Georgia, Arial’s relatively short x-height and heavier weight correlate with faster reading based on our model. Other than the general suitability and availability of Arial [73], no other clear winner emerged across age groups, implying the importance of personalizing the recommendation based on reader information. This result supports typographers’ views (P1 – P5) that there is not a universally most readable font for all, see Section §3.2.2.

Poppins shows value among young participants: Poppins show effectiveness among the participants below 25 but is rarely the

fastest for those above 40. Analyzing the Shapley values indicates that Poppins’s effectiveness is mainly correlated with its longer descender and tall x-height, especially for the 12/252 (4.8%) participants below 22. Because of this age group’s limited number of participants, our observations about Poppins require further study. Interestingly, prior research has identified Verdana as another font that is legible and preferred among younger populations [5, 80]. Using FontView, we compared the font characteristics of Verdana and Poppins. If we added Verdana to our font selections, its weight and x-height are most similar to Poppins. However, its ascender and descender lengths differ.

7.3.2 Personalized Font Recommendations. This paper provides further evidence that personalized font recommendations can increase reading speed for participants. Existing browsers offer “reading mode” extensions that help remove distracting page elements and offer consistent formatting [53]. However, these reader modes could benefit from adding the font selection feature. FontMART could augment a reading mode by collecting reader information to provide personalized font recommendations. This section discusses the challenges and possible solutions when implementing such a reading extension, supported by a tool like FontMART.

Improving recommendation with personal information: For a personalized recommender to work in the wild, users will need to trust the recommendations provided by the system [31]. Our initial results show that participants (1) are willing to trust font recommendations from a computer application and (2) are open to adopting new fonts into their digital reading experiences (§6.2.3). To improve the quality of recommendations, readers will need to voluntarily provide additional demographic information. Past research has shown that users will voluntarily participate in self-experimentation through online platforms [69]. Also, they will provide personal information when the perceived benefits outweigh the costs [64, 85]. A multi-load interface might create a more frictionless experience to collect personal information compared to a front-load interface [22]. In contrast to our front-load process, a multi-load interface asks for information at multiple points instead of at one point. This approach is promising because age is the most important reader characteristic for cohort-level recommendations. Less critical information can be collected after the reader has experienced the benefit of more efficient reading to continue to tweak the recommendation.

Putting readers in the loop: The typographers we interviewed discussed the negative effect of age on vision and reading (P2, P4, P5, §3.2.2). One typographer mentioned that as readers age, vision might change daily (P2, §3.2.2). Findings from prior research support this observation [65, 66, 96]. To account for changing reading characteristics, such as age, a timely update of personal information can ensure the model’s recommendation adapts to the reader’s evolving needs. While the reader’s evolving needs may warrant new recommendations, their past recommendations may provide valuable insights for future recommendations. Therefore, reinforcement learning approaches for font recommendation may be an alternative capable of providing more accurate recommendations on an individual level over time [84].

Balancing reader’s objectives: Faster reading speed may not be a reader’s only objective. In this work, reading speed varied more than comprehension due to the use of eighth-grade level passages and the removal of participants with poor comprehension, allowing us the singular focus on speed. On the other hand, the use of more difficult reading passages may lead to higher variance in comprehension. Section §8 discusses opportunities to evaluate speed and comprehension tradeoff when personalizing font recommendation. While FontMART may be extensible to other reading objectives, affording readers agency over font selection may ensure their final font choice better satisfies their needs and incentivizes continued engagement with the tool. Reader intervention is also important because different optimizations may work against each other. For instance, learning from less readable fonts has been associated with better retention in and outside of lab settings [37]. And while preferred fonts were not participants’ fastest fonts, it remains up to future work to investigate if they may otherwise prove beneficial to readers engaging with the reading longer-term.

8 LIMITATIONS AND FUTURE WORK

While this work provides some initial promising results on the effects of reader and font characteristics on reading speeds, conclusions are drawn based on a population of 252 crowdsourced participants reading in 8 typographer-selected fonts. We release all our study materials, tools, model, and data, in hopes that our work can be replicated and extended to a larger and more diverse population of readers, reading materials, languages, and fonts. Below we outline some of these limitations. Our work accounts for a subset of circumstances readers find themselves in. See Appendix §A for additional opportunities for future research.

Readers: The age distribution of our participants is not representative of the general population. While we only have 60/252 (23.8%) participants above 40, they make up 47.8% of the U.S. population [20]. Future studies may consider recruiting participants by specific age groups. Other reader characteristics, such as dyslexia, are harder to diagnose but important to consider [53, 57, 71]. Because of the design of our opt-in reading tests and our aggressive filtering of data to exclude slow reading speeds and low comprehension scores, we may be leaving out of consideration well-intentioned readers that struggle to read. Future work should expand on the populations recruited for such reading studies to incorporate the needs of diverse readers into recommender models like FontMart.

Reading Materials: To add control to our study, we used reading passages in English normed to an 8th-grade reading level [8, 90]. Prior work shows these 8th-grade reading passages are easy enough to allow the reader to read faster without reducing comprehension [90]. Future studies may consider optimizing for multiple objectives, such as the speed and comprehension trade-off [8, 91] by presenting passages with multiple levels of difficulty. Future work is also needed to evaluate the model’s effectiveness for other types of reading activities, such as long-form reading or skimming.

Language: Participants’ first language was English, and the typographers interviewed are from North America (P1, P3) and Europe (P2, P4, P5) with expertise in designing fonts for reading in

English. Thus, our results and model currently focus on fonts, passages, and reading in English. There are still many additional font characteristics to evaluate in different languages [102, 103]. For example, the effect of color combinations and optical sizing on reading speed may vary across languages [72, 75].

Fonts: The fonts in our selection are deemed readable by the typographers we interviewed. Additional training data on more extreme fonts may be necessary for expanding the model’s applicability. For instance, we currently find that heavier weight helps support faster reading, but the same result may not hold when bold fonts, such as Arial Bold, are included. In some instances, the font characteristics we use are too coarse. For example, average character spacing describes spacing at a lower fidelity by omitting the effect of kerning on readability [6, 36]. We believe that continued collaborations with typographers and readers can help grow the list of characteristics worth exploring.

9 CONCLUSION

We presented the FontMART model, a derivative of LambdaMART, that provides a new type of prediction: which fonts are most likely to improve the reading speed of individual readers for general body text. To train FontMART, we conducted a remote readability study with 252 paid crowdworkers to gather their preferences and reading speeds using eight fonts with different font characteristics.

FontMART can provide reading speed improvements compared to various baselines: providing an average improvement of +25.6 WPM over a participant’s preferred font, +14.8 WPM over Arial and +14.4 WPM over Georgia, the best font defaults from our study. Our results also show there is still a gap between the predicted and fastest fonts (Figure §6). Future research could recruit and incentivise unpaid participants [19, 31, 68, 69, 74] as well as evaluate more diverse font types, in order to broaden our understanding of the relationships between additional fonts and readers.

FontMART leverages reader and font characteristics identified through qualitative interviews with typographers. By comparing FontMART’s predictions with typographers’ domain expertise, we reinforce and uncover relationships between font and reader characteristics (i.e., reader age and x-height) to provide individual and cohort-level font recommendations. We found that specific fonts are likely more effective for different age groups, such as Poppins for younger and Georgia for older participants. By interviewing typographers to develop features for our Machine Learning model, our collaborative effort can help readers and designers alike. We hope future work can build on our findings and tools to personalize the design of reading interfaces and help improve the accessibility of digital information for all.

ACKNOWLEDGMENTS

We thank Christine Dierk, Qisheng Li, Jing Qian, Aleena Niklaus, Dave Miller, and Yi-le Zhang for their valuable feedback and edits. We thank Frank Griebßhammer, Sofie Beier, Sam Berlow, Tim Brown, and Bram Stein for sharing their perspectives on typography. We thank reviewers for their time and suggestions.

REFERENCES

- [1] David Amdur. 2006. *Typographic design in the digital studio*. Thomson Delmar Learning.
- [2] Aries Ardití and Jianna Cho. 2005. Serifs and font legibility. *Vision Research* 45, 23 (Nov. 2005), 2926–2933. <https://doi.org/10.1016/j.visres.2005.06.013>
- [3] Aries Ardití, Kenneth Knoblauch, and Ilana Grunwald. 1990. Reading with fixed and variable character pitch. *Journal of the Optical Society of America A* 7, 10 (Oct. 1990), 2011. <https://doi.org/10.1364/josaa.7.002011>
- [4] Jayeeta Banerjee and Moushum Bhattacharyya. 2011. Selection of the optimum font type and size interface for on screen continuous reading by young adults: an ergonomic approach. *Journal of human ergology* 40, 1-2 (2011), 47–62. <https://doi.org/10.11183/jhe.40.47>
- [5] Jayeeta Banerjee, Deepthi 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.
- [6] Radosław Bednarski and Maria Pietruszka. 2013. The computer-aided estimate of the text readability on the web pages. In *Advances in Intelligent Systems and Computing*, Aleksander Zgrzywa, Kazimierz Choroś, and Andrzej Siemiński (Eds.). Vol. 183 AISC. Springer Berlin Heidelberg, Berlin, Heidelberg, 211–220. https://doi.org/10.1007/978-3-642-32335-5_20
- [7] Sofie Beier. 2009. *Typeface Legibility : Towards defining familiarity*. Ph.D. Dissertation.
- [8] Sofie Beier, Sam Berlow, Esat Boucaud, Zoya Bylinskii, Tianyuan Cai, Jenae Cohn, Kathy Crowley, Stephanie L Day, Tilman Dingler, Jonathan Dobres, Jennifer Healey, Rajiv Jain, Marjorie Jordan, Bernard Kerr, Qisheng Li, Dave B Miller, Susanne Nobles, Alexandra Papoutsaki, Jing Qian, Tina Rezvanian, Shelley Rodrigo, Ben D Sawyer, Shannon M Sheppard, Bram Stein, Rick Treitman, Jen Vanek, Shaun Wallace, and Benjamin Wolfe. 2021. Readability Research: An Interdisciplinary Approach. *arXiv:2107.09615 [cs]* (July 2021).
- [9] Sofie Beier and Kevin Larson. 2013. How does typeface familiarity affect reading performance and reader preference? *Information Design Journal* 20, 1 (2013), 16–31. <https://doi.org/10.1075/idj.20.1.02bei>
- [10] Sofie Beier and Chiron A.T. Oederkerk. 2021. High letter stroke contrast impairs letter recognition of bold fonts. *Applied Ergonomics* 97 (2021), 103499. <https://doi.org/10.1016/j.apergo.2021.103499>
- [11] Sofie Beier, Chiron A. T. Oederkerk, Birte Bay, and Michael Larsen. 2021. Increased letter spacing and greater letter width improve reading acuity in low vision readers. *Information Design Journal* 26, 1 (2021), 73–88. <https://doi.org/10.1075/idj.19033.bei>
- [12] 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. *Conference on Human Factors in Computing Systems - Proceedings* (2001), 175–176. <https://doi.org/10.1145/634067.634173>
- [13] 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), 8.
- [14] Michael Bernard and Melissa Mills. 2000. So, What Size and Type of Font Should I Use on My Website? *Usability News* 2, 2 (2000), [online].
- [15] Michael L. Bernard, Barbara S. Chaparro, Melissa M. Mills, and Charles G. Halcomb. 2003. Comparing the effects of text size and format on the readability of computer-displayed Times New Roman and arial text. *International Journal of Human Computer Studies* 59, 6 (2003), 823–835. [https://doi.org/10.1016/S1071-5819\(03\)00121-6](https://doi.org/10.1016/S1071-5819(03)00121-6)
- [16] David Beymer, Daniel Russell, and Peter Orton. 2008. An eye tracking study of how font size and type influence online reading. , 15–18 pages. <https://doi.org/10.14236/ewic/hci2008.23>
- [17] Sanjiv K. Bhatia, Ashok Samal, Nithin Rajan, and Marc T. Kiviniemi. 2011. Effect of font size, italics, and colour count on web usability. *International Journal of Computational Vision and Robotics* 2, 2 (2011), 156–179. <https://doi.org/10.1504/IJCVR.2011.042271>
- [18] Dan Boyarski, Christine Neuwirth, Jodi Forlizzi, and Susan Harkness Regli. 1998. Study of fonts designed for screen display. In *Conference on Human Factors in Computing Systems - Proceedings*. ACM Press, Los Angeles, California, United States, 87–94. <https://doi.org/10.1145/274644.274658>
- [19] Erin Brady, Meredith Ringel Morris, and Jeffrey P. Bigham. 2015. Gauging receptiveness to social microvolunteering. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2015-April. ACM, New York, NY, USA, 1055–1064. <https://doi.org/10.1145/2702123.2702329>
- [20] U S C Bureau. 2019. Age and sex composition in the United States. 2010. *Tech. Rep* (2019).
- [21] Christopher J C Burges. 2014. From ranknet to lambdarank to lambdamart : An overview From RankNet to LambdaRank to LambdaMART : An Overview. *Learning* 11, May (2014), 81.
- [22] Victor S. Bursztyn, Jennifer Healey, Eunye Koh, Nedim Lipka, and Larry Birnbaum. 2021. Developing a Conversational Recommendation System for Navigating Limited Options. In *Conference on Human Factors in Computing Systems - Proceedings*. ACM, Yokohama Japan, 1–6. <https://doi.org/10.1145/3411763.3451596>
- [23] Aurélie Calabrèse, Allen M.Y. Cheong, Sing Hang Cheung, Yingchen He, Mi Young Kwon, J. Stephen Mansfield, Ahalya Subramanian, Deyue Yu, and Gordon E. Legge. 2016. Baseline MNREAD measures for normally sighted subjects from childhood to old age. *Investigative Ophthalmology and Visual Science* 57, 8 (2016), 3836–3843. <https://doi.org/10.1167/iovs.16-19580>
- [24] Ronald P Carver. 1990. *Reading rate: A review of research and theory*. Academic Press.
- [25] Ronald P Carver. 1992. Reading rate: Theory, research, and practical implications. *Journal of Reading* 36, 2 (1992), 84–95.
- [26] Barbara S. Chaparro, A. Dawn Shaikh, and Alex Chaparro. 2006. The legibility of cleartype fonts. *Proceedings of the Human Factors and Ergonomics Society* (2006), 1829–1833. <https://doi.org/10.1177/154193120605001724>
- [27] Jan Constantin. 2013. *Typographic Design Patterns And Current Practices* (2013 Edition) – Smashing Magazine. *Smashing Magazine* (2013).
- [28] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *RecSys 2016 - Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, Boston Massachusetts USA, 191–198. <https://doi.org/10.1145/2959100.2959190>
- [29] Kathy Crowley and Marjorie Jordan. 2020. Readability formats Deliver instantaneous change. <https://readabilitymatters.org/articles/instantaneous-change> (Accessed Aug 7, 2021).
- [30] Alexandru Cuibaris. 2021. What Font is | WhatFontis.com. <http://www.whatfontis.com/>
- [31] Nediya Daskalova, Eindra Kyi, Kevin Ouyang, Arthur Borem, Sally Chen, Sung Hyun Park, Nicole Nugent, and Jef Huang. 2021. Self-e: Smartphone-supported guidance for customizable self-experimentation. In *Conference on Human Factors in Computing Systems - Proceedings*. ACM, New York, NY, USA, 1–13. <https://doi.org/10.1145/3411764.3445100>
- [32] Nediya Daskalova, Danaë Metaxa-Kakavouli, Adrienne Tran, Nicole Nugent, Julie Boergers, John McGeary, and Jeff Huang. 2016. SleepCoacher: A personalized automated self-experimentation system for sleep recommendations. In *UIST 2016 - Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 347–358. <https://doi.org/10.1145/2984511.2984534>
- [33] Joshua R. de Leeuw. 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods* 47, 1 (2015), 1–12. <https://doi.org/10.3758/s13428-014-0458-y>
- [34] Joshua R. de Leeuw and Benjamin A. Motz. 2016. Psychophysics in a web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods* 48, 1 (2016), 1–12. <https://doi.org/10.3758/s13428-015-0567-2>
- [35] Material Design. 2018. Understanding typography. <https://material.io/design/typography/understanding-typography.html#type-properties>
- [36] Steve Devan. 1987. Desktop Publishing on the Macintosh: A Software Perspective. *Educational Technology* 27, 8 (1987), 12–14.
- [37] Connor Diemand-Yauman, Daniel M. Oppenheimer, and Erika B. Vaughan. 2011. Fortune favors the: Effects of disfluency on educational outcomes. *Cognition* 118, 1 (Jan. 2011), 111–115. <https://doi.org/10.1016/j.cognition.2010.09.012>
- [38] Jonathan Dobres, Bryan Reimer, and Nadine Chahine. 2016. The effect of font weight and rendering system on glance-based text legibility. In *AutomotiveUI 2016 - 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Proceedings*. ACM, Ann Arbor MI USA, 91–96. <https://doi.org/10.1145/3003715.3005454>
- [39] Berrin Dogusoy, Filiz Cicek, and Kursat Cagiltay. 2016. How serif and sans serif typefaces influence reading on screen: An eye tracking study. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 9747. Springer, 578–586. https://doi.org/10.1007/978-3-319-40355-7_55
- [40] Arpad E Elo. 1978. *The rating of chessplayers, past and present*. Arco Pub.
- [41] Michael Gasser, Julie Boeke, Mary Haffeman, and Rowena Tan. 2005. The influence of font type on information recall. *North American Journal of Psychology* 7, 2 (2005), 181–188.
- [42] Severin Hacker and Luis Von Ahn. 2009. Matchin: Eliciting user preferences with an online game. In *Conference on Human Factors in Computing Systems - Proceedings*. ACM, New York, NY, USA, 1207–1216. <https://doi.org/10.1145/1518701.1518882>
- [43] Richard H. Hall and Patrick Hanna. 2004. The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour and Information Technology* 23, 3 (2004), 183–195. <https://doi.org/10.1080/01449290410001669932>
- [44] R. L. Hu, J. Granderson, D. M. Auslander, and A. Agogino. 2019. Design of machine learning models with domain experts for automated sensor selection for energy fault detection. *Applied Energy* 235 (2019), 117–128. <https://doi.org/10.1016/j.apenergy.2018.10.107>

- [45] Shuhui Jiang, Zhaowen Wang, Aaron Hertzmann, Hailin Jin, and Yun Fu. 2020. Visual Font Pairing. *IEEE Transactions on Multimedia* 22, 8 (2020), 2086–2097. <https://doi.org/10.1109/TMM.2019.2952266> arXiv:1811.08015
- [46] Sheree Josephson. 2008. Keeping your readers' eyes on the screen: An eye-tracking study comparing sans serif and serif typefaces. *Visual Communication Quarterly* 15, 1-2 (April 2008), 67–79. <https://doi.org/10.1080/15551390801914595>
- [47] Florian Kadner, Yannik Keller, and Constantin A. Rothkopf. 2021. Adaptofont: Increasing individuals' reading speed with a generative font model and bayesian optimization. In *Conference on Human Factors in Computing Systems - Proceedings*. ACM, New York, NY, USA, 1–11. <https://doi.org/10.1145/3411764.3445140> arXiv:2104.10741
- [48] Tugba Kulahcioglu and Gerard De Melo. 2019. Fontlex: A typographical lexicon based on affective associations. *LREC 2018 - 11th International Conference on Language Resources and Evaluation* (2019), 62–69.
- [49] Pawan Kumar and Manmohan Sharma. 2021. Data, Machine Learning, and Human Domain Experts: None Is Better than Their Collaboration. *International Journal of Human-Computer Interaction* (2021), 1–14. <https://doi.org/10.1080/10447318.2021.2002040>
- [50] Kevin Larson. 2004. The Science of Word Recognition. *Microsoft Typography* July 2004 (2004), 1–14.
- [51] Kevin Larson and Matthew Carter. 2016. Sitka: A collaboration between type design and science. In *Digital Fonts and Reading*. Vol. 1. World Scientific, 37–53. https://doi.org/10.1142/9789814759540_0003
- [52] Qisheng Li, Sung Jun Joo, Jason D. Yeatman, and Katharina Reinecke. 2020. Controlling for Participants' Viewing Distance in Large-Scale, Psychophysical Online Experiments Using a Virtual Chinrest. *Scientific Reports* 10, 1 (2020), 1–11. <https://doi.org/10.1038/s41598-019-57204-1>
- [53] Qisheng Li, Meredith Ringel Morris, Adam Fourney, Kevin Larson, and Katharina Reinecke. 2019. The impact of web browser reader views on reading speed and user experience. In *Conference on Human Factors in Computing Systems - Proceedings*. ACM, Glasgow Scotland UK, 1–12. <https://doi.org/10.1145/3290605.3300754>
- [54] Scott M. Lundberg and Su In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, Vol. 2017-December. 4766–4775. arXiv:1705.07874
- [55] K M MacQueen, E McLellan-Lemal, K Bartholow, and B Milstein. 2008. Team-based Codebook Development: Structure, Process, and Agreement In Guest G & MacQueen KM (Eds.), *Handbook for Team-Based Qualitative Research*. Lanham, New York, Toronto, Plymouth: Altamira Press.[Google Scholar] (2008).
- [56] Katsumi Minakata and Sofie Beier. 2021. The effect of font width on eye movements during reading. *Applied Ergonomics* 97 (2021), 103523. <https://doi.org/10.1016/j.apergo.2021.103523>
- [57] Aliaksei Miniukovich, Antonella De Angeli, Simone Sulpizio, and Paola Venuti. 2017. Design guidelines for web readability. In *DIS 2017 - Proceedings of the 2017 ACM Conference on Designing Interactive Systems*. ACM, Edinburgh United Kingdom, 285–296. <https://doi.org/10.1145/3064663.3064711>
- [58] Ahmad Zamzuri Mohamad Ali, Rahani Wahid, Khairulnuar Samsudin, and Muhammad Zaffwan Idris. 2013. Reading on the computer screen: Does font type has effects on Web text readability? *International Education Studies* 6, 3 (2013), 26–35. <https://doi.org/10.5539/ies.v6n3p26>
- [59] S Morrison and J Noyes. 2003. A Comparison of Two Computer Fonts: Serif versus Ornate Sans-serif. *Usability News* (2003), 1–4.
- [60] Emily Namey, Greg Guest, Lucy Thairu, and Laura Johnson. 2008. Data reduction techniques for large qualitative data sets approaches to data analysis. *Handbook for team-based qualitative research* 2, 1 (2008), 137–161.
- [61] Chiron Oederkerk, Katsumi Minakata, and Sofie Beier. 2020. Fonts of wider letter shapes improve legibility. *Journal of Vision* 20, 11 (2020), 1285. <https://doi.org/10.1167/jov.20.11.1285>
- [62] Peter O'Donovan, Janis Libeks, Aseem Agarwala, and Aaron Hertzmann. 2014. Exploratory font selection using crowdsourced attributes. *ACM Transactions on Graphics* 33, 4 (July 2014), 1–9. <https://doi.org/10.1145/2601097.2601110>
- [63] Madoka Ohnishi and Koichi Oda. 2021. The effect of character stroke width on legibility: The relationship between duty ratio and contrast threshold. *Vision Research* 185 (2021), 1–8. <https://doi.org/10.1016/j.visres.2021.03.006>
- [64] Debajyoti Pal, Chonlameth Arpikanondt, and Mohammad Abdur Razzaque. 2020. Personal Information Disclosure via Voice Assistants: The Personalization-Privacy Paradox. *SN Computer Science* 1, 5 (2020), 1–17. <https://doi.org/10.1007/s42979-020-00287-9>
- [65] Allen L. Pelletier, Ledy Rojas-Roldan, and Janis Coffin. 2016. Vision loss in older adults. *American Family Physician* 94, 3 (2016), 219–226.
- [66] Anne Marie Piper, Robin Brewer, and Raymundo Cornejo. 2017. Technology learning and use among older adults with late-life vision impairments. *Universal Access in the Information Society* 16, 3 (2017), 699–711. <https://doi.org/10.1007/s10209-016-0500-1>
- [67] E. C. Poulton. 1965. Letter differentiation and rate of comprehension in reading. *Journal of Applied Psychology* 49, 5 (1965), 358–362. <https://doi.org/10.1037/h0022461>
- [68] Sam Ransbotham and Gerald C. Kane. 2011. Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in Wikipedia. *MIS Quarterly: Management Information Systems* 35, 3 (2011), 613–627. <https://doi.org/10.2307/23042799>
- [69] Katharina Reinecke and Krzysztof Z. Gajos. 2015. Labin the wild: Conducting large-scale online experiments with uncompensated samples. In *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*. ACM, Vancouver BC Canada, 1364–1378. <https://doi.org/10.1145/2675133.2675246>
- [70] Luz Rello and Ricardo Baeza-Yates. 2013. Good fonts for dyslexia. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, *ASSETS 2013*. 1–8. <https://doi.org/10.1145/2513383.2513447>
- [71] Luz Rello, Ricardo Baeza-Yates, Abdullah Ali, Jeffrey P. Bigham, and Miquel Serra. 2020. Predicting risk of dyslexia with an online gamified test. *PLoS ONE* 15, 12 December (2020), e0241687. <https://doi.org/10.1371/journal.pone.0241687> arXiv:1906.03168
- [72] Luz Rello and Jeffrey P. Bigham. 2017. Good background colors for readers: A study of people with and without dyslexia. In *ASSETS 2017 - Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. 72–80. <https://doi.org/10.1145/3132525.3132546>
- [73] Luz Rello, Martin Piolot, and Mari Carmen Marcos. 2016. Make it big! The effect of font size and line spacing on online readability. In *Conference on Human Factors in Computing Systems - Proceedings*. ACM, San Jose California USA, 3637–3648. <https://doi.org/10.1145/2858036.2858204>
- [74] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. In *Fifth International AAAI Conference on Weblogs and Social Media*. AAAI, Menlo Park, CA, USA, 321–328.
- [75] Elisabete Rolo. 2021. Type to Be Seen and Type to Be Read. In *Lecture Notes in Networks and Systems*, Vol. 220. Springer, New York, NY, USA, 334–341. https://doi.org/10.1007/978-3-030-74605-6_42
- [76] Nicolas P. Rougier and Behdad Eshahbod. 2018. *Digital typography*. Wiley. 1–29 pages. <https://doi.org/10.1145/3214834.3214837>
- [77] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*. IEEE, 166–171.
- [78] Thomas W. Schubert, Carla Murteira, Elizabeth C. Collins, and Diniz Lopes. 2013. ScriptingRT: A Software Library for Collecting Response Latencies in Online Studies of Cognition. *PLoS ONE* 8, 6 (2013), e67769. <https://doi.org/10.1371/journal.pone.0067769>
- [79] Lloyd S Shapley. 2016. *17. A value for n-person games*. Princeton University Press.
- [80] James E. Sheedy, Manoj V. Subbaram, Aaron B. Zimmerman, and John R. Hayes. 2005. Text legibility and the letter superiority effect. *Human Factors* 47, 4 (Dec. 2005), 797–815. <https://doi.org/10.1518/001872005775570998>
- [81] Janan Al Awar Smither and Curt C. Braun. 1994. Readability of prescription drug labels by older and younger adults. *Journal of Clinical Psychology in Medical Settings* 1, 2 (June 1994), 149–159. <https://doi.org/10.1007/BF01999743>
- [82] KyoungHee Son, Seo Young Oh, Yongkwan Kim, Hayan Choi, Seok-Hyung Bae, and Ganguk Hwang. 2015. Color Sommelier: Interactive Color Recommendation System Based on Community-Generated Color Palettes. In *Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology*. 95–96.
- [83] Namrata Srivastava, Rajiv Jain, Jennifer Healey, Zoya Bylinskii, and Tilman Dingler. 2021. Mitigating the Effects of Reading Interruptions by Providing Reviews and Previews. In *Conference on Human Factors in Computing Systems - Proceedings*. ACM, New York, NY, USA, 1–6. <https://doi.org/10.1145/3411763.3451610>
- [84] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.
- [85] K. Swearingen and R. Sinha. 2001. Beyond Algorithms : An HCI Perspective on Recommender Systems. *ACM SIGIR 2001 Workshop on Recommender Systems (2001)* (2001), 1–11.
- [86] Yu-Chi Tai, Shun-nan Yang, John Hayes, and James Sheedy. 2012. *Effect of Character Spacing on Text Legibility*. Technical Report. Vision Performance Institute, Pacific University, Oregon, USA. 24 pages.
- [87] Jaime Teevan, Christine Alvarado, Mark S. Ackerman, and David R. Karger. 2004. The perfect search engine is not enough: A study of orienteering behavior in directed search. *Conference on Human Factors in Computing Systems - Proceedings* 6, 1 (2004), 415–422.
- [88] Jaime Teevan, Susan T. Dumais, and Eric Horvitz. 2005. Personalizing search via automated analysis of interests and activities. *SIGIR 2005 - Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2005), 449–456. <https://doi.org/10.1145/1076034.1076111>
- [89] Robert L. Thorndike. 1953. Who belongs in the family? *Psychometrika* 18, 4 (1953), 267–276. <https://doi.org/10.1007/BF02289263>

- [90] Shaun Wallace, Zoya Bylinskii, Jonathan Dobres, Bernard Kerr, Sam Berlow, Rick Treitman, Nirmal Kumawat, Kathleen Arpin, Dave B Miller, Jeff Huang, and Others. 2022. Towards Individualized Reading Experiences: Different Fonts Increase Reading Speed for Different Individuals. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–56.
- [91] Shaun Wallace, Jonathan Dobres, and Ben D. Sawyer. 2021. Considering the Speed and Comprehension Trade-Off in Reading Mediated by Typography. *Journal of Vision* 21, 9 (2021), 2249. <https://doi.org/10.1167/jov.21.9.2249>
- [92] Shaun Wallace, David Sasson, and Hua Guo. 2017. Visualizing self-tracked mobile sensor and self-reflection data to help sleep clinicians infer patterns. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. Part F127655. 2194–2200. <https://doi.org/10.1145/3027063.3053138>
- [93] Shaun Wallace, Rick Treitman, Nirmal Kumawat, Kathleen Arpin, Jeff Huang, Ben Sawyer, and Zoya Bylinskii. 2020. Individual Differences in Font Preference & Effectiveness as Applied to Interlude Reading in the Digital Age. *Journal of Vision* 20, 11 (2020), 412. <https://doi.org/10.1167/jov.20.11.412>
- [94] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24. <https://doi.org/10.1145/3359313> arXiv:1909.02309
- [95] Xinxin Wang, David Rosenblum, and Ye Wang. 2012. A daily, activity-aware, mobile music recommender system. In *MM 2012 - Proceedings of the 20th ACM International Conference on Multimedia*. ACM Press, Nara, Japan, 1313–1314. <https://doi.org/10.1145/2393347.2396459>
- [96] Gale R. Watson. 2001. Low vision in the geriatric population: Rehabilitation and management. *Journal of the American Geriatrics Society* 49, 3 (2001), 317–330. <https://doi.org/10.1046/j.1532-5415.2001.4930317.x>
- [97] Arnold Wilkins, Roanna Cleave, Nicola Grayson, and Louise Wilson. 2009. Typography for children may be inappropriately designed. *Journal of Research in Reading* 32, 4 (Nov. 2009), 402–412. <https://doi.org/10.1111/j.1467-9817.2009.01402.x>
- [98] Wilson R F. 2001. HTML E-mail: text font readability study. Results of a survey conducted. *Consultant* February (2001), 5.
- [99] Jason Wu, Gabriel Reyes, Sam C. White, Xiaoyi Zhang, and Jeffrey P. Bigham. 2020. Towards Recommending Accessibility Features on Mobile Devices. In *ASSETS 2020 - 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '20)*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3373625.3418007>
- [100] Ying Zi Xiong, Ethan A. Lorsche, John Stephen Mansfield, Charles Bigelow, and Gordon E. Legge. 2018. Fonts designed for macular degeneration: Impact on reading. *Investigative Ophthalmology and Visual Science* 59, 10 (2018), 4182–4189. <https://doi.org/10.1167/iovs.18-24334>
- [101] Jingwei Xu, Yuan Yao, Hanghang Tong, Xianping Tao, and Jian Lu. 2015. Ice-Breaking: Mitigating cold-start recommendation problem by rating comparison. In *IJCAI International Joint Conference on Artificial Intelligence*, Vol. 2015-January. IJCAI, California, USA, 3981–3987.
- [102] Shangshang Zhu, Xinyu Su, and Yenan Dong. 2021. Effects of the Font Size and Line Spacing of Simplified Chinese Characters on Smartphone Readability. *Interacting with Computers* 33, 2 (2021), 177–187. <https://doi.org/10.1093/iwc/iwab020>
- [103] Xiaofeng Zhu and Diego Klabjan. 2020. Listwise learning to rank by exploring unique ratings. In *WSDM 2020 - Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM '20)*. Association for Computing Machinery, Houston, TX, USA, 798–806. <https://doi.org/10.1145/3336191.3371814> arXiv:2001.01828
- [104] Abdelwahab Zramdini and Rolf Ingold. 1998. Optical font recognition using typographical features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 8 (1998), 877–882. <https://doi.org/10.1109/34.709616>

A ADDITIONAL LIMITATIONS

Uncontrolled Visual Font Size: One limitation of conducting remote readability studies is the variation in physical size and screen resolution across devices, as well as the difference in viewing distances. As a result, the absolute dimensions of a 16px font may vary between devices. This helps explain why prior research used smaller fonts, such as 10 point size [12, 18], because the pixel densities of physical monitors were lower. Our methods trade internal validity for applied validity by studying readers in their everyday reading environments. Future work may consider adding more control using tools such as virtual chinrest [52].

Generalizability to Mobile Devices: While we developed our study interface for all device types, our participants primarily read using desktops and laptops. Consequently, FontMART's ability to make recommendations on mobile devices is unclear. Future studies can focus on mobile reading to determine if recommendations would systematically differ in this case. Conducting similar studies on mobile devices may introduce several confounding factors. For instance, mobile devices allow the participants to easily vary their viewing distance because a mobile device is rarely in a fixed position. Thus, the visual font size could vary more per participant. It is also more likely that the reader could be in motion, such as jogging or riding in a car, and be more frequently interrupted in their reading. Each of these scenarios presents interesting constraints for future studies to understand.

Environmental Distractions: When participating in remote studies, it is not uncommon for participants to encounter distractions such as notifications on their computer screen, phone calls, or caring for family members [83]. While our study design does not directly control for external distractions, readers could take breaks before and after reading passages. We did explore various approaches of encoding the reading speeds to prevent the model from over-focusing on slight differences in reading speeds. In our study, models that used binary labeling generally performed better than graded labeling. This result suggests that external factors may be contributing to noisier data. Future remote readability studies may consider incorporating these factors in the modeling stages.

Self-assessed Reader Characteristics: The current version of FontMART uses reader characteristics that are easy to gather from participants, such as age or self-reported reading speed. Inclusion of additional characteristics and more robust collection approaches may further improve model performances. For instance, self-assessment of vision may be conducted through short remote vision tests rather than question with binary answers. Our methods to measure font familiarity may be sub-optimal, thus explaining its inconsistent effect on reading speed in our results (§6.2.2). It relies on participants' varying abilities to recognize fonts and judge their familiarity based on examples [1, 59, 62, 76]. Future research could develop more reliable methods to measure font familiarity over time and account for participants' prior habits and experiences with fonts.

B OTHER CHARACTERISTICS OF FONTS

The typographers also proposed other design considerations, such as the font's counter and rhythm. Not all characteristics are considered in our study because of a lack of consensus on how to best measure them.

Counter. Typographers interviewed define counter as the glyph area that is partially or entirely enclosed by strokes. P1 discussed the importance of having a large counter to help distinguish characters.

“If you don't have a large enough letter, those character forms will fill in, [and] it is very difficult to differentiate the letters.” (P1)

P3 also pointed out the relationship between counters and the character spacing. “If you compare the letter space volume to the counter space volume, a certain ratio may turn out to be more pleasing for reading. (P3)”

Rhythm. The rhythmic quality of the font is developed from stroke contrast and thickness. Typographers P3 and P4 mentioned the rhythm and texture of a typeface as qualities that may influence reading performance. “When you have [stroke] contrast, there is an opportunity for a rhythm to develop in where the heaviness falls. (P3)” However, no sufficient details or consensus exist on how to quantify this quality.

“If you were to pour some water into that area [between letters]. Out of this amount of space, you could rank fonts in terms of their evenness of these volumes of water, and you will easily see that well-made typefaces have an evenness that oddly made typefaces don't have.” (P3)

C STUDY QUESTIONS

Questions marked * are mandatory.

C.1 Pre-Survey

- (1) What is your age? (in years) *
Age selection dropdown
- (2) What is your gender?
Text answer
- (3) What is/are your native language(s) *
 - English
 - Other:
- (4) What other languages do you speak?
Text answer. Leave blank if you only speak English.
- (5) What is your highest attained education level? *
 - Less than high school
 - High school/GED
 - Some college
 - Associate's degree (2-years of college)
 - Bachelor's Degree (4-years of college)
 - Master's degree
 - Doctoral degree
 - Professional degree
 - Prefer not to say
- (6) Please describe your current occupation:
Text answer
- (7) Do you feel comfortable with reading articles written in English? *
 - Not comfortable
 - Somewhat comfortable
 - Very comfortable
- (8) How would you rate your speed as a reader? *
Likert Scale 1 (Very Slow) – 7 (Very Fast)
- (9) Do you feel your reading speed could be improved?
Likert Scale 1 (Not at all) – 7 (Extremely)
- (10) How would you rate your proficiency as a reader? *
Likert Scale 1 (Very Poor) – 7 (Excellent)
- (11) Do you read to young children under the age of 6? *
 - Yes
 - No
 - Maybe
- (12) Have you ever been diagnosed with a reading or learning disability (e.g., dyslexia)? If yes, which one and how long ago?
(If you prefer not to answer, you can leave this blank. If you choose to answer, this question will NOT be used to disqualify you from the study or be used against you in any way. Note: Learning disabilities are common. In fact, one in five children in the U.S. has learning and attention issues such as dyslexia and ADHD.)
Text answer
- (13) Have you ever been diagnosed with any medical and neurological conditions (macular degeneration, diabetes, ADD, memory disorders, LPD, dyspraxia, etc...) If yes, which one/s and how long ago?
(If you prefer not to answer, you can leave this blank. If you choose to answer, this question will NOT be used to disqualify you from the study or be used against you in any way.)
Text answer
- (14) Are you currently under the influence of any drugs, medications, alcohol, or other stimulants (e.g., caffeine, nicotine) that may affect reading/attention? If yes, which?
(If you prefer not to answer, you can leave this blank. If you choose to answer, this question will NOT be used to disqualify you from the study or be used against you in any way.)
Text answer
- (15) Do you have normal or corrected vision? *
 - No
 - Yes
- (16) If your vision is corrected, how was it corrected (glasses, lenses, surgery, etc.)?
Text answer
- (17) What device/s do you read on for leisure or personal interest? *
Text answer
- (18) What device/s do you read on for work or study? *
Text answer
- (19) What do you read for leisure or personal interest? *
Text answer
- (20) What do you read for work or study? *
Text answer

- (21) How often do you read English-written articles for leisure or personal interest? *
- Less than once a month
 - Once a month
 - Once a week
 - 2-3 times a week
 - Everyday
- (22) How often do you read English-written articles for work or study? *
- Less than once a month
 - Once a month
 - Once a week
 - 2-3 times a week
 - Everyday
- (23) What are the names of font(s) you commonly use, and/or familiar with? *
- Text answer
- (24) Current Reading Environment
- Text answer
- (25) Which device are you using right now to participate in this study? *
- Laptop
 - Desktop
 - Tablet
 - Phone
 - Kindle or other e-reader
- (26) Please describe your current surroundings. For example, are you indoors/outside, by a window, under natural or artificial light, is the room light/dark, is the room small/large? *
- Text answer

C.2 Post-Survey

C.2.1 Font Familiarity. We asked the participants about their familiarity with all eight fonts used in our study: Times, Source Serif Pro, Georgia, Merriweather, Roboto, Arial, Open Sans, Poppins.

- (1) How familiar are you with the Font [X] **?
- The following text is rendered in Font [X]:
- “How familiar are you with the Font [X] Here is some fun text to read in the above font to help you better assess your familiarity. Thanks for taking this readability test!”
- Not at all
 - Slightly
 - Moderately
 - Very
 - Extremely

C.2.2 Other Questions.

- (1) Would you use your fastest word spacing for reading if you had a choice? * (Note that this is a question asked as an exploration for a future study.)
- Yes
 - No
 - Maybe
- (2) Do you feel changing font attributes (such as size, character spacing, word spacing, etc..) could help you read faster? *
- Likert Scale 1 (Not at all) – 7 (Extremely)

- (3) Would you like an application to change font attributes (such as size, character spacing, word spacing, etc..) for reading if you had the choice? *
- Likert Scale 1 (Not at all) – 7 (Extremely)
- (4) Can you comment on any strategies you used to complete this study?
- Text answer
- (5) Are there any general rules you used to help you read faster or remember more information?
- Text answer
- (6) Do you have any other comments about the study, did you find anything confusing? (For example, was it easy/difficult, did you experience any issues, did you know what to do, etc...)
- Text answer

C.3 Mini Questionnaire after Each Passage

All questions are mandatory.

- (1) How familiar were you with the topic from the previous reading passage?
- Not at all
 - Slightly
 - Moderately
 - Very
 - Extremely
- (2) How interesting was the previous reading passage?
- Not at all
 - Slightly
 - Moderately
 - Very
 - Extremely
- (3) In the previous passage, do you feel the font allowed you to read faster than normal?
- Strongly disagree
 - Disagree
 - Neither agree nor disagree
 - Agree
 - Strongly agree

D DESCRIPTION OF READING PASSAGES

As reported in the open-source materials, the readability specialist selected the passages from ebooks in Project Gutenberg¹⁶. The topics per passage vary, covering topics such as history of science, biography, and botany. The reading specialist reduced the length per passage to 160–178 words and made minor adjustments to vocabulary and sentence structure to norm them to an eighth-grade level (Lexile range¹⁷: 800–1200, Flesch score¹⁸: 61 – 80).

E NONSENSICAL SURVEY RESPONSES

Nonsensical responses most frequently occur in mandatory questions requiring text answers in the pre-survey. All text survey responses are reviewed, and participants with nonsensical survey

¹⁶<https://www.gutenberg.org>

¹⁷<https://hub.lexile.com/analyzer>

¹⁸<https://www.readabilityformulas.com/free-readability-formula-tests.php>

responses are filtered from our study. Here are several examples of the responses received:

- (1) What devices do you read on for leisure or personal interest?
 - Try using words that might appear on the page you're looking for. For example, "cake recipes" instead of "how to make a cake."
 - skills, qualities or values in action
 - They Exercise Physical exercise is important for both physical and mental health.
 - Relate the hobby or interest directly to the company
- (2) Are you currently under the influence of any drugs, medications, alcohol, or other stimulants (e.g., caffeine, nicotine) that may affect reading/attention? If yes, which?
 - make any question in your survey required so that respondents must
 - Neurological disorders are medically defined as disorders that affect the brain as well as the nerves found throughout the human body and the spinal cord
 - If you prefer not to answer, you can leave this blank. If you choose to answer, this question will NOT be used to disqualify you from the study or be used against you in any way. Note: Learning disabilities are common. In fact, one in five children in the U.S. has learning and attention issues such as dyslexia and ADHD.

F FEATURES COLLECTED

See Table 2.

G RESULTS OF LINEAR MIXED EFFECT MODELS

See Table 3.

H EXAMPLE OF GRADED AND BINARY TRAINING LABELS

See Table 4.

I EFFECT OF FONT RECOMMENDERS BY OUTCOME LABELS

See Table 5 and Table 6.

J DOES PREFERENCE INFORMATION CONTRIBUTE TO BETTER PREDICTIONS?

Font preference may reveal additional valuable information about a participant, for prediction purposes. Still, in a limited dataset, it may also provide noise that could negatively affect model performance. When we experimented by including the preference information in the input of the model trained with binary labels, the resulting model did not outperform the original model. Detailed results are included in Table 5 and Table 6. Future work can consider data from an even larger population of participants to determine whether information about font preference can improve the predictions of the fastest font further.

K IS THERE A LEARNING EFFECT DURING THE STUDY?

Is there a learning effect during the study? Recall that each passage is split into four sections, distributed across four consecutive screens (Figure 4, §4.2). We observe that the average reading speed increases from sections 1 to 3 irrespective of the passage order, and sections 3 and 4 have similar speeds, as shown in Figure 12. The speed increase may reflect participants adjusting to the new font. In the font recommendation approach that follows, we assume that participant reading speeds increase at a similar rate across fonts. Because of the similarity in the average speeds between sections 3 and 4, we also assume that participants reach their peak speed by section 3. To wash out all these effects, we therefore average over all four sections to obtain a reading speed for a participant in a particular font.

Feature Group	Feature Name	Description
Font Characteristics	Weight	The width of the thickest stroke divided by the character’s width. Calculated with letter “o”.
	Stroke Contrast	$1 - r$, where r is the ratio of the thinnest to thickest stroke of the letter “o”. This value is larger for fonts with greater stroke contrast.
	Ascender Length	The vertical distance from the top of the letter with an ascender to the top of the letter “x”. The ascender length is calculated by averaging the results from letters “b”, “d”, “f”, “h”, “i”, “j”, “k”, “l”, “t”.
	Descender Length	The vertical distance from the bottom of the letter with a descender to the bottom of the letter “x”. The descender length is calculated by averaging the results from letters “g”, “q”, “p”, “y”, “j”.
	x-Height	The average height of the lowercase letters, excluding those with ascenders and descenders.
	Character Width	The average width of all lowercase characters.
	Standard Deviation of Character Width	The standard deviation of the average width of all lowercase characters.
	Character Spacing	The average horizontal space between the bounding boxes of all lowercase characters.
	Grayscale	The proportion of pixels shown in the bounding box when all lowercase characters are visualized. This is the only characteristics measured on the rasterized images of the rendered letters.
Reader Characteristics	Age	Participant’s age in years (0 – 99+).
	Self-reported Reading Speed	Participants’s self-reported reading speed on a 7-point Likert scale. Collected as a part of the pre-survey.
	Self-reported Reading Frequency	We asked participants about their frequency of reading English-written articles for work or leisure, and we assigned each frequency an ordinal integer value of 0 through 4, from the least frequent to the most. We then sum the values to obtain an overall measure of the reading frequency with a value ranging from 0 to 8. Collected as a part of the pre-survey.
	Font Familiarity	We presented the same text rendered in all eight fonts, and we asked participants to rate their familiarity with each font on a 5-point Likert scale. Collected as a part of the post-survey.

Table 2: Description of all features collected and analyzed in the scope of this study.

	Chisq	d.f.	Pr(>Chisq)
Font	9.122	7	0.244
Age	8.408	1	0.004 **

Table 3: Results of a linear mixed-effect model predicting the participant’s average reading speed for each font tested. Participant ID (uniquely identifying each participant) and passage ID (uniquely identifying each of the eight passages) are incorporated as crossed random effects.

Data for Participant 47:					
Passage ID	Font	Preference	Speed (WPM)	Graded Label	Binary Label
23	Open Sans	3	318.33	5	0
22	Source Serif Pro	4	351.75	5	0
28	Times	8	388.25	6	0
20	Poppins	2	452.25	7	0
26	Georgia	5	490.66	7	0
27	Roboto	6	502.33	8	0
21	Merriweather	1	559.00	8	0
25	Arial	7	626.00	10	1

Table 4: Example of Graded and Binary Training Labels. All rows belong to the same randomly drawn participant in our study. This participant has only one non-zero binary label because their reading speed measurements in other fonts are not within the 90% of their fastest, Arial.

Baseline Category	Font	Model 1: Graded Label with Preference	Model 2: Binary Label with Preference	Model 3: Binary Label without Preference
One-size-fits-all	Arial	4.143	-5.733	-14.832
	Georgia	2.942	-7.967	-14.419
	Merriweather	-2.520	-13.658	-19.146
	Open Sans	-5.012	-15.348	-22.465
	Poppins	-9.230	-19.250	-25.449
	Roboto	-1.903	-11.924	-20.068
	Source Serif Pro	-1.335	-11.181	-19.531
	Times	0.535	-8.890	-16.450
Random	One of Eight Fonts	3.004	-9.080	-23.413
Preferred	One of Eight Fonts	-5.699	-16.831	-25.592

Table 5: Differences between the speed of baselines and the speed of the recommended fonts, from the recommender models trained on different outcome labels.

	Model 1: Graded Label with Preference	Model 2: Binary Label with Preference	Model 3: Binary Label without Preference
Position	NDCG	MAP	MAP
1	0.386	0.513	0.539
2	0.488	0.413	0.448
3	0.540	0.407	0.439
4	0.587	0.429	0.458
5	0.624	0.460	0.489
6	0.667	0.499	0.522
7	0.706	0.537	0.562
8	0.752	0.574	0.597

Table 6: Ranking metrics for font recommender models trained on different outcome labels. The model performance is measured with ranking metrics using cross validation. We show our results in NDCG for the model trained with graded labels and MAP for the model trained with binary labels. Normalized discounted cumulative gain (NDCG) and mean average precision (MAP) help measure the model’s ability to rank the faster font higher in the list of eight fonts when trained with graded and binary labels, respectively.

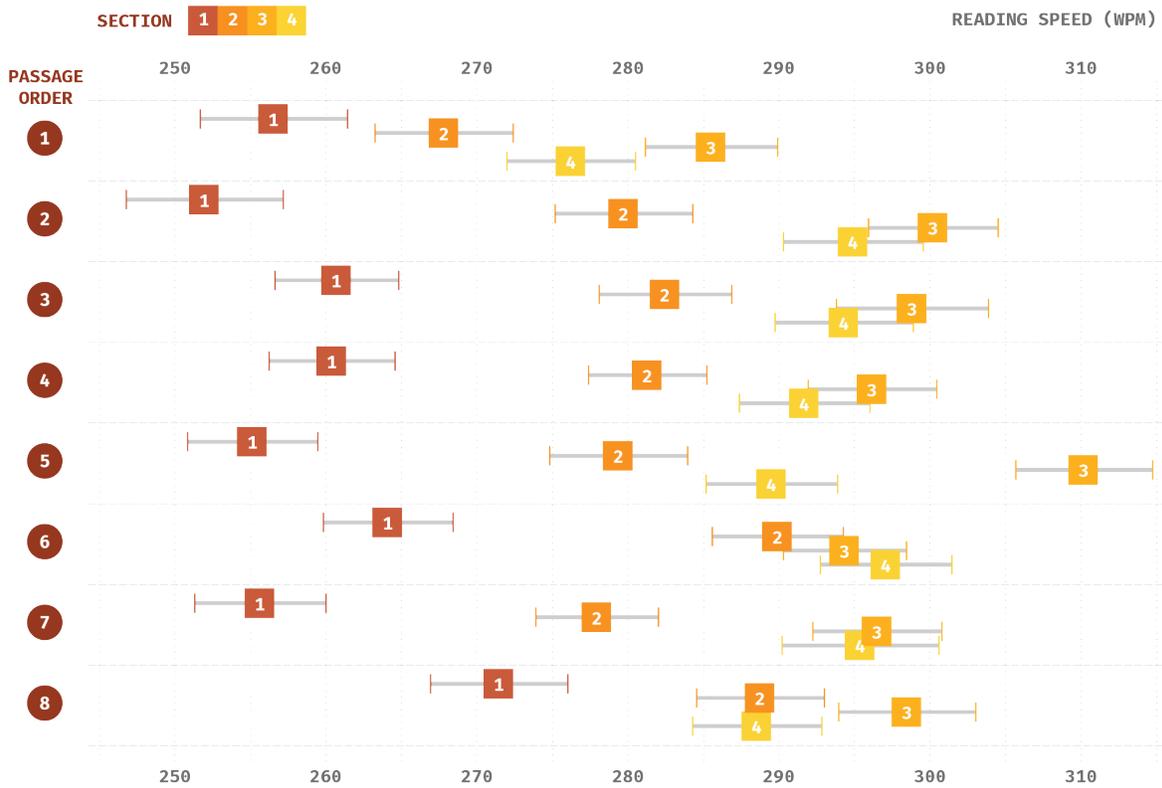


Figure 12: Average reading speed by section and passage, across participants. Each row represents the order in which a passage was presented in the study, and each data point shows the average reading speed across 252 participants for one of the four sections of a passage. Average reading speed increases from sections 1 through 3, suggesting a learning effect. On the other hand, the interval between sections 3 and 4 mostly overlaps, suggesting that reading speed peaks by the third section. Note that the order of font presentation is randomized and not linked to the reading order of passages. Error bars represent ± 1 within-subjects standard error.