The Effect of Font Weight and Rendering System on Glance-Based Text Legibility

Jonathan Dobres, Bryan Reimer
Massachusetts Institute of Technology, AgeLab
Cambridge, MA United States
jdobres@mit.edu, reimer@mit.edu

Nadine Chahine
Monotype Imaging, Inc.
Woburn, MA United States
nadine.chahine@monotype.com

ABSTRACT

In-vehicle user interfaces increasingly rely on digital text to display information to the driver. Led by Apple's iOS, thin, lightweight typography has become increasingly popular in cutting-edge HMI designs. The legibility trade-offs of lightweight typography are sparsely studied, particularly in the glance-like reading scenarios necessitated by driving. Previous research has shown that even relatively subtle differences in the design of the on-screen typeface can influence to-device glance time in a measurable and meaningful way. Here we investigate the relative legibility of four different weights (line thicknesses) of type under two different rendering systems (suboptimal rendering and optimal rendering). Results indicate that under suboptimal rendering, the lightest weight typeface renders poorly and is associated with markedly degraded legibility. Under optimal rendering, lighter weight typefaces show enhanced legibility compared to heavier typefaces. The reasons for this pattern of results, and its implications for design considerations in modern HMIs, are discussed.

Author Keywords

Automotive human machine interface; Distraction; Driver safety; Psychophysics; Font characteristics; Legibility; Typeface style.

ACM Classification Keywords

J.4. [**Social and Behavioral Sciences**]: Psychology; J.7 [**Computers in Other Systems**]: Real time; H5.m [**Information Interfaces and Presentation**]: Miscellaneous.

INTRODUCTION

As technological advances in mobile computing have been brought inside the vehicle, the complexity of in-vehicle interfaces has increased dramatically. Most new in-vehicle HMIs present a variety of textual information on digital displays, from live weather reports to integrated navigation systems. Static hardware buttons and simple gauges have gradually given way to larger in-vehicle screens (often touch controlled) that can accommodate the growing number of data streams that a driver might access.

These types of displays come with new design challenges. One must consider that as these displays become ever-richer sources of information, the driver's attention may be pulled toward the device with increasing frequency and/or duration, instead of the road. Therefore, any aspect of the interface's design that can be optimized to reduce visual demands, such as minimizing off-road glance time, can free attentional resources to devote to the roadway, a critical aspect of safe driving [30].

Previous research has shown that even something as subtle as the choice of typeface used for the in-vehicle display can significantly impact off-road glance time and task completion time [27]. That study compared two similar sans-serif typefaces: a humanist style typeface, and a square grotesque. In a fully simulated driving environment, drivers spent less time glancing at an in-vehicle display set in a humanist style typeface as compared to a square grotesque typeface, particularly among males. The differences governing the design of these typefaces, relatively subtle outside the world of typography, nevertheless had a significant impact on driver behavior. Follow-up research based on traditional psychophysical techniques that emulated glance-like reading behavior found similar differences between these same typefaces [7], as well as dramatic effects of font size on legibility thresholds.

Recent design trends have popularized the use of extremely lightweight fonts, perhaps most prominently in recent versions of Apple's iOS operating system. The effect of stroke weight (line thickness) on the usability and aesthetics of modern software is much debated in the popular press [16], but empirical research on the matter is sparse.

The earliest studies on the effects of font weight on text legibility indicated a kind of "goldilocks" relationship, with medium weight fonts demonstrating superior legibility compared to lightweight or bold ones [2,10,14,15], a finding supported by much later work [1]. However, some studies show a more linear relationship between legibility

Automotive'UI 16, October 24-26, 2016, Ann Arbor, MI, USA
 ISBN 978-1-4503-4533-0/16/10...$15.00
DOI: http://dx.doi.org/10.1145/3003715.3005454

and font weight [8]. The effect of increasing line weight on the clarity of letters and/or numerals tended to be subtle; an early study by Hughes demonstrated that increased character boldness significantly affected the accurate identification of some numerals, but not others [9].

Research conducted on digital displays showed an advantage for heavier weight fonts, and that this advantage was more pronounced on LCD screens compared to CRT [29,31]. As in earlier research, a "goldilocks" relationship is apparent in contemporary assessments of font weight [3], with most of the relationship evident in a performance decrement for lighter-weight fonts.

Several of the more recent findings in this area raise a question as to whether font weight exerts its effects because of interactions with the pixel grid. The primary task of modern text rendering systems is to translate a glyph—a mathematically defined set of curves representing a character—into a bitmap that can be rendered on a finite pixel grid. Recent research suggests that this function does not proceed equally well for all fonts [7]. A recent study in Chinese found that font weight made a significant difference in legibility only for characters of higher visual complexity (i.e., more strokes), which may be tied to rendering limitations on the pixel grid [13]. A similar effect has been reported in another study on Chinese text legibility [6], and is implicated in some studies on font weight conducted in English [29,31].

Over the years, researchers have assessed legibility with a variety of techniques, including blink rate [14,15], optimal reading distance and/or text size assessment [1,8-10,28,29,31], and reading speed [3,13]. Crucially, all of these studies have assessed legibility using self-paced metrics, without any meaningful time constraint or intake pressure. Thus, one might question the applicability of this research to contexts in which information must be taken in quickly and under constraints, as in the behavior commonly observed during driving. A recent line of research has used psychophysical thresholding procedures to emulate glance-like reading [6,7], of the kind typically observed when reading from in-vehicle screens while driving. This more closely situates the assessment of legibility to a real-world behavior with notable safety implications.

To more directly examine these issues with empirical assessment methods, here we present a set of studies that examines the effect of font weight on glance legibility under two different rendering systems.

METHODS

Participants

This study was split into two phases, with each phase using a different participant sample. Phase I included a total of 48 participants between the ages of 35 and 75 (26 women and 22 men, no significant difference in age between genders [$t(46) = 0.47$, $p = 0.645$]). All participants gave their written, informed consent to participate. Exclusion criteria included experience of a major medical illness in the last six months, conditions that impair vision (other than typical nearsightedness or farsightedness), or a history of chronic or acute neurological problems. Participants were also required to be native English speakers. All participants had normal or corrected-to-normal vision (glasses or contact lenses) and were tested on site for near acuity using the Federal Aviation Administration's test for near acuity (Form 8500-1), and for far acuity using a Snellen eye chart.

Phase II included a total of 47 participants between the ages of 35 and 75 (24 women and 23 men, no significant difference in age between genders [$t(45) = 0.85$, $p = 0.848$]). Inclusion criteria were the same as in Phase I.

Apparatus

Monitor

An Asus monitor was used to display the experiment (27", 1920x1080 resolution, 60Hz refresh rate). Participants were asked to maintain a distance of approximately 70cm (27.5") from the display. Head position was left unconstrained, to allow for a reasonable amount of positional variability, as might be encountered in automotive contexts.

Rendering Systems & Experiment Software

Phase I utilized custom software written in Matlab and the Psychtoolbox 3 [5,22] running on a 2.5Ghz Intel Core i5 Mac Mini running Mac OS 10.9.1. Psychtoolbox is a cross-platform framework for experiment control, and its text rendering algorithms rely heavily on the capabilities of the underlying operating system. As a result, the quality of its rendering on Mac OS X tends to be "good enough"; while it employs basic text smoothing algorithms, it is not heavily optimized for precise typographic display.

Phase II utilized custom software written in PsychoPy2 [20]. PsychoPy is also a cross-platform experiment framework, but uses self-contained modules for the display of visual elements that are independent of the underlying operating system. PsychoPy's default text rendering pipeline was modified to utilize Monotype's iType rendering system for text display, which enables "continuous stroke modulation" (CSM)—a feature that allows for the contours of the typeface to be altered mathematically and optimized for the monitor being used [25]. This represents a best in class rendering engine for typographic display. The integration of Monotype's proprietary rendering engine necessitated a switch to a PC with a 3.60Ghz Intel Core i7 processor running Windows 7.

Aside from the different text rendering pipelines used in each phase of the experiment and the different participant samples, all aspects of the experiment were kept constant between the two phases. Both Psychtoolbox and PsychoPy are capable of frame-level stimulus display and event timing. Any differences that might arise from the use of the two different experiment control platforms would be too minuscule to be meaningful.

Task

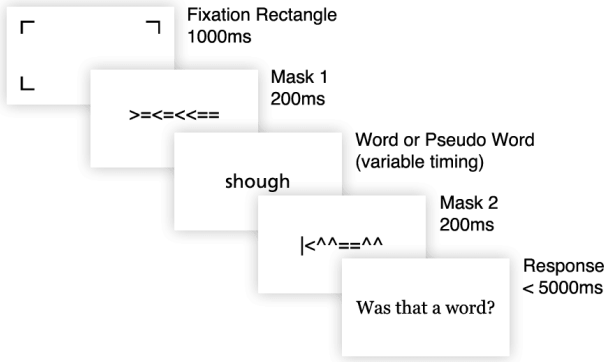

Figure 1: A schematic illustration one trial of the lexical decision task.

Participants completed a yes/no lexical decision task [18] in which they had to decide whether a briefly presented stimulus was either a word or a non-word (a pronounceable combination of letters that does not form a word). A schematic of a task trial is shown in Figure 1.

Stimuli drew from the same pool of words and pseudowords as used in [7]. These were 6-letter words (or 6-letter pseudoword strings) generated from an online orthographic database [17].

The amount of time that the word (or non-word) is presented on screen is automatically adjusted depending on the respondent's accuracy, with presentation time eventually converging on a value corresponding to approximately 79.4% accuracy [11,12]. Thus, more legible conditions should result in a lower threshold presentation time. For details on this method and its implementation in this experiment, see [7].

Conditions Tested

Figure 2: Samples of the 4 weights of Frutiger used in this study, from top to bottom: Light, Regular, Medium, Bold. Note that the experiment scaled text to a nominal capital letter height of 4mm (13 pixels), which is significantly lower resolution than the samples shown here (~40 pixels). Figure rendered in Adobe Photoshop CS6.

Thresholds were assessed for variants of the Frutiger typeface in four different weights, as illustrated in Figure 2. Stimuli were displayed in white text (RGB: 255, 255, 255) against a black background (RGB: 0, 0, 0). Typefaces were scaled to a target size of 4mm (19.6 arcmin) based on the height of the typeface's capital 'H', though all stimuli were displayed in lowercase letters.

The experiment employed a complete within-subjects design, with all participants undergoing threshold assessment for all four conditions. Each threshold assessment consisted of 100 trials, equally split between randomly interleaved word and pseudoword trials. Block order was counterbalanced across participants. Counterbalancing was effective, in that typeface and block order were unrelated (Phase I: $X^2_{(3)} = 0.00$, $p = 1.00$; Phase II: $X^2_{(3)} = 6.19$, $p = .103$, Friedman tests of block order).

Data Analysis

Thresholds were obtained for each condition by calculating the median stimulus duration (presentation time) of each condition's final 20 trials. Participant responses were also saved for secondary analyses of performance accuracy. Data were analyzed under a linear mixed effects model. Rendering system was treated as a categorical between-subjects factor, and font weight was treated as a linear within-subjects factor, with numerical values established by measuring the average stroke width in pixels of each font's capital 'H' at a large size (these values were 7 for Light, 10 for Regular, 12 for Medium, and 14 for Bold). All statistics were computed and visualized using R [26].

RESULTS

Performance Accuracy

When using adaptive threshold procedures, task difficulty varies in accordance with participant performance to converge on a target accuracy level (in this case, 79.4% accuracy). Therefore, accuracy should not be significantly different between the conditions tested. In Phase I, accuracy did not differ significantly between font weights ($F(3, 141) = 0.16$, $p = 0.922$), nor did any condition differ significantly from the calibrated accuracy level (all $p > 0.447$, t-tests against a *mu* of 79.4%). Likewise, in Phase II, accuracy did not differ significantly between font weights ($F(3, 138) = 0.28$, $p = 0.842$), nor did any condition tested show an accuracy significantly different from the intended calibration point (all $p > 0.158$, t-tests). Thus, the calibration procedure worked as intended for the samples tested.

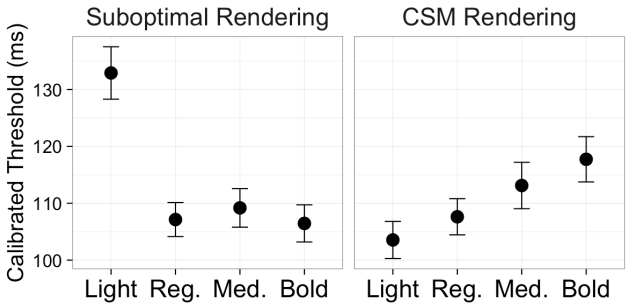

Figure 3: Estimated mean thresholds (±1 mean-adjusted standard error) for each condition tested in Phase I (left) and Phase II (right).

Reading Time Thresholds

In a model that considers both phases of the experiment together, no main effect of rendering system on reading time threshold is found ($F(1, 295) = 0.24$, $p = 0.627$), nor is there a main effect of font weight ($F(1, 52) = 1.27$, $p = 0.265$). However, there is a significant interaction between rendering system and font weight ($F(16, 316) = 15.74$, $p < 0.001$), as is overtly apparent in Figure 3.

Decomposing each phase of the experiment into separate models, we find that in Phase I (suboptimal rendering), there is a significant effect of font weight ($F(1, 47) = 14.13$, $p < 0.001$). Posthoc testing shows that this effect is primarily driven by the greatly elevated threshold observed for the light-weight font, whereas no significant differences are observed between other weights (light vs. regular, medium, or bold, all $p < 0.001$; all other comparisons, $p > 0.596$, t-tests).

In Phase II (CSM rendering), we find a different pattern. Once again, a significant main effect of font weight is evident ($F(1, 46) = 6.30$, $p = 0.016$). Posthoc testing indicates a significant difference between the two extreme weights ($t(46) = 2.29$, $p = 0.027$), while all comparisons that differ by more than one weight (i.e., Light vs. Medium) trend toward significant differences ($p < 0.10$).

The regular-weight conditions employed in this study correspond precisely to conditions examined as part of two earlier experiments [7]. Comparing all four samples, there is no evidence to suggest that the reading time threshold estimated for regular-weight Frutiger differed between experiments ($F(3, 171) = 0.164$, $p = 0.920$), giving a good indication of the assessment method's reliability.

DISCUSSION

We found that the effect of font weight on legibility thresholds was strongly mediated by the rendering engine used to draw the glyphs. Under suboptimal or "good enough" rendering, we found that lightweight text degraded considerably compared to other fonts. At the same time, there were no significant differences in legibility thresholds between regular, medium, and bold weight fonts. Conversely, under "best in class" CSM rendering, we found evidence of a linear relationship between font weight and legibility, with lighter-weight fonts having superior legibility thresholds compared to heavier weights.

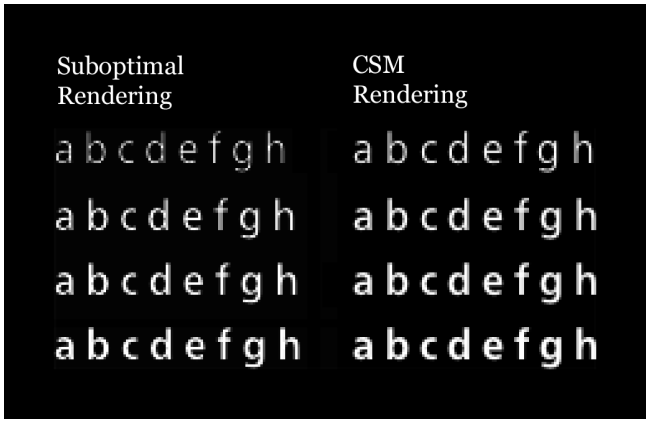

Figure 4: Text samples for light, regular, medium, and bold weight fonts as rendered in Matlab (left) and PsychoPy (right) at a nominal height of 16 pixels, enlarged to show detail.

The legibility thresholds obtained from suboptimal rendering make intuitive sense. As Figure 4 shows, the lightest-weight font is rendered very poorly under this system, resulting in a markedly elevated reading time threshold. At the same time, the more subtle differences between the three heavier fonts are lost, resulting in threshold estimates that are significantly lower than the lightest weight font, but not significantly different among each other.

The results obtained from "best in class" rendering demonstrate that under more finely tuned rendering pipelines, many of the subtle differences between fonts can be preserved. This produces a linear trend in the data, with lighter weight fonts performing better than heavier ones. The superior performance of the lighter-weight fonts contradicts much of the previous literature on font weight and legibility, which typically show enhanced legibility for heavier fonts. This difference may arise from the assessment methods used; the present study assesses legibility in a way that simulates glance-like, time-pressured reading, whereas all other research in this area has employed self-paced methods of assessment. The brief presentation times employed here may in turn expose a crowding effect. Although crowding is traditionally conceptualized as a phenomenon of peripheral vision [4,21], there is evidence that crowding can occur even in the fovea, and is ameliorated by increased inter-character spacing [19,23,24]. We speculate that lighter-weight fonts may have larger inter-character spaces by definition, and thus be less affected by foveal crowding. It should be noted that at least one previous study using the present stimulus paradigm showed a legibility advantage for heavier fonts [6], but given that a suboptimal renderer was used in that work (the same as the one used here), those results are not inconsistent with the present findings.

Taken together, these findings show the complex ways in which a user-facing design element and an underlying technology can interact, and highlight the unique task that HMI designers face when attempting to balance the needs of a brand or set of style guidelines with the practicalities of a given technology or behavioral context. Particularly at a time when centralized software systems are being deployed across a diversity of device types, care must be taken to ensure that the underlying hardware can support the visual appearance of the software as designed.

ACKNOWLEDGMENTS

Support for this work was provided by the US DOT's Region I New England University Transportation Center at MIT, the Toyota Class Action Settlement Safety Research and Education Program, and Monotype Imaging, Inc. The views and conclusions being expressed are those of the authors, and have not been sponsored, approved, or endorsed by Toyota or plaintiffs' class counsel. Monotype provided additional support in the form of typeface files, rendering engine customization, and typographic expertise.

REFERENCES

1. A Arditi, R Cagenello, and B Jacobs. 1995. Letter stroke width, spacing, and legibility. *OSA Technical Digest Series* 1: 324–327.
2. C Berger. 1944. II. Stroke-width, form and horizontal spacing of numerals as determinants of the threshold of recognition. *Journal of Applied Psychology* 28, 3: 208–231. http://doi.org/10.1037/h0054045
3. Jean-Baptiste Bernard, Girish Kumar, Jasmine Junge, and Susana T L Chung. 2013. The effect of letter-stroke boldness on reading speed in central and peripheral vision. *Vision research* 84, C: 33–42. http://doi.org/10.1016/j.visres.2013.03.005
4. H Bouma. 1970. Interaction effects in parafoveal letter recognition. *Nature* 226, 5241: 177–178.
5. David H Brainard. 1997. The Psychophysics Toolbox. *Spatial vision* 10, 4: 433–436. http://doi.org/10.1163/156856897X00357
6. Jonathan Dobres, Nadine Chahine, Bryan Reimer, David Gould, and Nan Zhao. 2016. The effects of Chinese typeface design, stroke weight, and contrast polarity on glance based legibility. *Displays* 41, C: 42–49. http://doi.org/10.1016/j.displa.2015.12.001
7. Jonathan Dobres, Nadine Chahine, Bryan Reimer, David Gould, Bruce Mehler, and Joseph F Coughlin. 2016. Utilising psychophysical techniques to investigate the effects of age, typeface design, size and display polarity on glance legibility. *Ergonomics* 59: 1–15. http://doi.org/10.1080/00140139.2015.1137637
8. P R Hind, B H Tritt, and E R Hoffmann. 1976. The Effects of Level of Illumination, Stroke-Width, Visual Angle and Contrast on the Legibility of Numerals of Various Fonts. 46–55.
9. C L Hughes. 1961. Variability of stroke within digits. *Journal of Applied Psychology* 45, 6: 364–368. http://doi.org/10.1037/h0048997
10. J E Kuntz and R B Sleight. 1950. Legibility of numerals: The optimal ratio of height to width of stroke. *The American Journal of Psychology* 63, 4: 567. http://doi.org/10.2307/1418871
11. Marjorie R Leek. 2001. Adaptive procedures in psychophysical research. *Perception & psychophysics* 63, 8: 1279–1292.
12. H Levitt. 1971. Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America* 49, 2B: 467–477. http://doi.org/doi:10.1121/1.1912375
13. Na Liu, Ruifeng Yu, and Yunhong Zhang. 2016. Effects of Font Size, Stroke Width, and Character Complexity on the Legibility of Chinese Characters. *Human Factors and Ergonomics in Manufacturing & Service Industries* 26, 3: 381–392. http://doi.org/10.1002/hfm.20663
14. M Luckiesh and F K Moss. 1939. The visibility and readability of printed matter. *Journal of Applied Psychology* 23, 6: 645–659. http://doi.org/10.1037/h0055273
15. M Luckiesh and F K Moss. 1940. Boldness as a factor in type-design and typography. *Journal of Applied Psychology*.
16. Sean Madden. 2013. Design Lessons From iOS 7. *Harvard Business Review*. Retrieved May 23, 2016 from https://hbr.org/2013/09/design-lessons-from-ios-7
17. D A Medler and J R Binder (eds.). 2005. *MCWord*. Retrieved December 13, 2013 from http://www.neuro.mcw.edu/mcword/
18. D E Meyer and R W Schvaneveldt. 1971. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology* 90, 2: 227–234.
19. Veronica Montani, Andrea Facoetti, and Marco Zorzi. 2014. The effect of decreased interletter spacing on orthographic processing. *Psychonomic Bulletin & Review* 22, 3: 824–832. http://doi.org/10.3758/s13423-014-0728-9

20. Jonathan W Peirce. 2008. Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics* 2: 1–8. http://doi.org/10.3389/neuro.11.010.2008

21. D G Pelli, K A Tillman, J Freeman, M Su, T D Berger, and N J Majaj. 2007. Crowding and eccentricity determine reading rate. *Journal of Vision* 7, 2: 1–36. http://doi.org/10.1167/7.2.20

22. D G Pelli. 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial vision* 10, 4: 437–442.

23. Manuel Perea and Pablo Gomez. 2012. Increasing interletter spacing facilitates encoding of words. *Psychonomic Bulletin & Review* 19, 2: 332–338. http://doi.org/10.3758/s13423-011-0214-6

24. Manuel Perea, Carmen Moret-Tatay, and Pablo Gomez. 2011. The effects of interletter spacing in visual-word recognition. *Acta psychologica* 137, 3: 345–351. http://doi.org/10.1016/j.actpsy.2011.04.003

25. Ronald N Perry and Sarah F Frisken. 2002. *A New Framework for Representing, Rendering, Editing, and Animating Type*. Mitsubishi Electric Research Laboratories.

26. R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from http://www.R-project.org/

27. Bryan Reimer, Bruce Mehler, Jonathan Dobres, et al. 2014. Assessing the impact of typeface design in a text-rich automotive user interface. *Ergonomics* 57, 11: 1643–1658. http://doi.org/10.1080/00140139.2014.940000

28. Barbara Elizabeth Roethlein. 1912. The Relative Legibility of Different Faces of Printing Types. *The American Journal of Psychology* 23, 1: 1. http://doi.org/10.2307/1413112

29. 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: 797–815.

30. M. Sodhi, B. Reimer, and I. Llamazares. 2002. Glance analysis of driver eye movements to evaluate distraction. *Behavior Research Methods* 34, 4: 529–538.

31. Venkiteshwar Subbaram. 2004. Effect of display and text parameters on reading performance. 1–275.