THE INCREDIBLE SHRINKING LETTER: HOW FONT SIZE AFFECTS THE LEGIBILITY OF TEXT VIEWED IN BRIEF GLANCES

Jonathan Dobres[1], Bryan Reimer[1], Lauren Parikhal[1], Emily Wean[1], Nadine Chahine[2]
[1]Massachusetts Institute of Technology, AgeLab, Cambridge, MA, USA
[2]Monotype Imaging Inc., Woburn, MA, USA
jdobres@mit.edu, reimer@mit.edu, parikhal@mit.edu, ewean@mit.edu,
nadine.chahine@monotype.com

Summary: As in-vehicle interfaces have become miniature computers with user-facing LCD screens, the complexities of designing for them have increased tremendously. Given their safety-critical nature, designers must carefully consider every aspect of the vehicle's digital interface. Recent research has suggested that even the typeface used to display the interface's text can have significant impacts on driver behaviors such as total off-road glance time and secondary task completion time. Here we outline a psychophysical method for rapidly assessing the glance-based legibility of two different typefaces (a "humanist" and a "square grotesque") presented in two different sizes (3mm and 4mm). Consistent with previous research, we find that humanist type is more legible than square grotesque. We also find that text is empirically less legible at 3mm compared to 4mm, and that this effect is especially pronounced for the square grotesque typeface. Legibility thresholds were also found to increase linearly with age, more than doubling across the age range studied. We hypothesize that the square grotesque's intrinsic design characteristics cause it to scale poorly at small sizes and lose important details, especially in suboptimal display conditions.

INTRODUCTION

Recent rapid advances in mobile computing have brought a new class of interfaces into the modern vehicle. Where once the text presented inside the vehicle was static and of low information density (such as the speedometer, fuel gauge, and digital radio readout), today's in-vehicle interfaces present screens filled with text arranged in dynamic, ever changing layouts. These types of layouts are necessary to accommodate the features that users have come to expect; a single screen can therefore be used to display weather information, navigation directions, or a variety of infotainment services. As drivers perform an increasing number of non-driving tasks while underway in the vehicle, it is crucial that the interfaces used for these tasks be designed to optimize usability and minimize visual distraction.

Historically, research on legibility has been concerned with "embedded reading", utilizing metrics and tasks that replicate the traditional experience of reading lines or entire paragraphs of text (for review, see Legge & Bigelow, 2011). However, modern reading is increasingly done in brief glances, whether looking down at a smartphone or glancing at an in-vehicle display. Recent work examining the legibility of in-vehicle menus in a full cab driving simulator has shown that the typeface used to display menu text can have meaningful effects on safety-relevant driver behaviors such as secondary task completion time and total time spent glancing to the in-vehicle screen (Reimer et al., 2014). Subsequent work has shown that traditional psychophysical methods can be used to reveal similar glance legibility effects (Dobres, Chahine, Reimer, Gould,

Mehler, Pugh, et al., 2014b). These methods are advantageous because they provide significant research flexibility while minimizing costs and operational complexity, allowing for a wide variety of issues to be examined rapidly under controlled conditions, even generalizing to studies of the legibility of foreign characters (Dobres, Chahine, Reimer, Gould, & Mehler, 2014a).

The size of type strongly impacts its legibility (Legge, Pelli, Rubin, & Schleske, 1985). Although an intuitive statement, typographic size is a nuanced topic governed by a number of complicating factors, particularly in regards to how a given typeface might be rendered on a page or screen. For example, in traditional metal printing, the letterforms of smaller sizes of a typeface would be modified from the master design to accommodate the physical behavior of the ink (Carter, 1984). This practice has largely been abandoned in the digital era. As a result, in digital typography, the legibility of type at small sizes is mediated by the limits of the pixel grid. A small letter may have a total width of 6-8 pixels, and the letter's strokes may be a single pixel or less in thickness. Therefore, fonts are often smoothed to improve their appearance, but this can lead to blurring of the typeface (Chaparro, Shaikh, Chaparroa, & Merkle, 2010).

Automotive OEMs are not obligated to use cutting-edge displays in their vehicles; they may use screens with a lower pixel-per-inch resolution, or they may use software architectures that do not support more advanced forms of font smoothing, especially given that suboptimal components are likely to come at a cheaper cost with little perceptible effect on the vehicle's overall quality. Moreover, current guidelines on the use of typography for in-vehicle displays suggest measuring the size of a typeface by the height of its capital 'H' (International Standards Organization, 2009), which ignores the fact that not all typefaces scale down to small sizes equally well, owing to the intrinsic qualities of the typeface's design (Legge & Bigelow, 2011).

These factors often go unconsidered when designing in-vehicle interfaces, yet have significant safety-relevant implications, as a less legible typeface can lead to greater off-road glance time, and thus, more time spent with eyes off the forward roadway. To examine the impact of digital type scaling on glance legibility, here we examine two typefaces: Monotype's Frutiger®, a "humanist" sans-serif; and Monotype's Eurostile®, a "square grotesque" typeface. Each typeface is displayed at 3mm and 4mm sizes, and legibility thresholds for each condition are measured.

METHODS

Participants

Thirty participants aged 36-75 took part in this study. All participants gave their written, informed consent to participate. Exclusion criteria included experience of a major medical illnesses 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.

Data from 5 participants were excluded due to a failure to use appropriate corrective lenses. Data from 6 participants were excluded due to an apparent failure to reach a stable threshold estimate

in the allotted time. One participant was excluded because he/she exhibited mean reaction times greater than 1.5s. This left a total of 18 participants, equally split between men and women. Men had a mean age of 54.1 years (SD = 14.3) and women had a mean age of 61.1 years (SD = 8.9). Age did not differ significantly between genders ($t_{(13.4)} = 1.24$, $p = .235$).

Task, Stimuli, and Apparatus

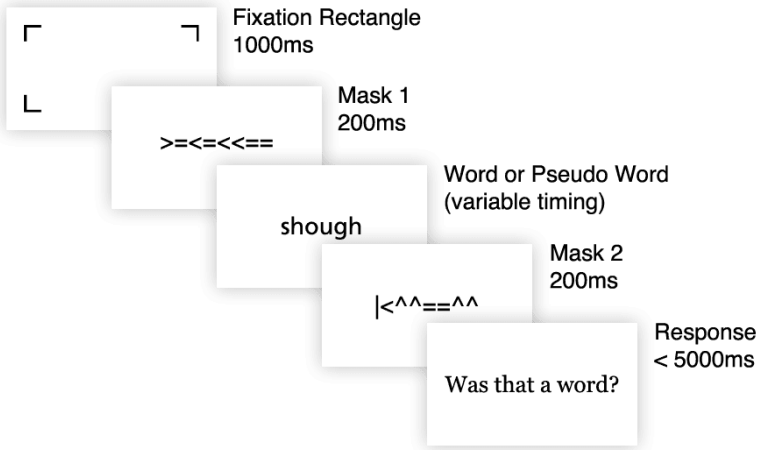

Figure 1. A schematic illustrating one trial of the word recognition task

The experimental task consisted of a 1-interval forced choice word recognition task. A single trial of this task is illustrated in Figure 1. Participants were asked to determine whether a briefly presented set of letters formed a word or pseudoword. The difficulty of the task was automatically adjusted based on the participant's performance accuracy, with the goal of keeping mean accuracy at approximately 79.4%. In this way, participants will arrive at different presentation time thresholds for each typeface. A more legible typeface should require less time on screen to be read accurately (a lower threshold).

Figure 2. Samples of the typefaces used, scaled to identical capital heights. Figure rendered in Adobe Photoshop CS5

Stimuli drew from the same pool of words and pseudowords as used in (Dobres, Chahine, Reimer, Gould, Mehler, Pugh, et al., 2014b). These were 6-letter words (or 6-letter pseudoword strings) generated from an online orthographic database (Medler & Binder, 2005). Stimuli were always displayed in white text (RGB: 255, 255, 255) on a black (RGB: 0, 0, 0) background. Four typographic conditions were tested: Monotype's Frutiger typeface in 3mm and 4mm sizes, and Monotype's Eurostile typeface in 3mm and 4mm sizes. Typefaces were scaled to a target size

based on the height of the typeface's capital 'H'. Sizes were selected based upon an ad hoc review of in-vehicle displays from a 2010 Infinity EX35, 2010 Lincoln MKS, 2014 Chevrolet Impala, and 2014 Mercedes CLA250. The 4mm size is at the top end of the ISO 15008 standard's acceptable range for typeface sizes (International Standards Organization, 2009), while the 3mm size is representative of smaller typographic sizes, but still remains well above the minimum height described in the standard, and it is also larger than the smallest sizes observed in production vehicles.

The experiment was divided into 4 blocks (1 per typeface/size configuration), and each block 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 ($X^2_{(3)} = 0.67$, $p = .881$, Friedman test of block order).

Data were collected on a 2.5Gz Intel Core i5 Mac Mini running Mac OS X 10.9.1. Stimuli were displayed using Matlab and Psychtoolbox 3 (Brainard, 1997; Pelli, 1997). A high refresh rate Asus monitor was used to display the experiment (27", 1920x1080 resolution, 109.9Hz refresh rate). Participants were asked to maintain a distance of approximately 27.5" from the display.

Data Analysis

Thresholds were obtained for each condition by calculating the median stimulus duration (presentation time) of each condition's final 20 trials. In addition to stimulus duration values, reaction times and response accuracies were also recorded for each trial, and corresponding metrics were computed: for performance accuracy, by averaging values for the last 20 trials; for reaction time, by averaging all but the first 20 trials. Data were analyzed in a repeated-measures design (typeface and size as within-subjects factors). All statistics were computed and visualized using R (R Core Team, 2014).

RESULTS

Performance Accuracy

When using adaptive staircase procedures, the goal is to vary task difficulty in accordance with participant responses and thus hold performance accuracy constant. Therefore, while we expect each condition tested to produce a different stimulus duration threshold, response accuracy across conditions should be similar. Overall response accuracy in this experiment did not differ from the theoretical calibration point of 79.4% ($t_{(17)} = -0.15$, $p = 0.884$). This also holds true when the four conditions are tested separately (all $p > 0.225$). Taken together, these results indicate that the adaptive staircase calibration procedure was able to converge on a stable estimate of stimulus duration threshold within the allotted trials.

Reaction Time

A number of reaction time effects are evident in these data. Reaction times did not differentiate typefaces or display sizes. However, reaction times were significantly slower for incorrect responses compared to correct responses (610ms and 492ms, respectively, $F_{(1, 17)} = 29.8$, $p <$

0.001), and were also slower when responding to pseudoword stimuli compared to word stimuli (548ms and 483ms, respectively, $F_{(1, 17)} = 19.0$, $p < 0.001$). These differences are consistent with data from other similar studies (Dobres, Chahine, Reimer, Gould, & Mehler, 2014a; Dobres, Chahine, Reimer, Gould, Mehler, Pugh, et al., 2014b), and support the idea that participants may have needed more time to reach a "decision boundary" when dealing with stimuli that were ultimately misjudged, or were composed of unfamiliar pseudowords (Ratcliff & McKoon, 2008).

Stimulus Duration Thresholds

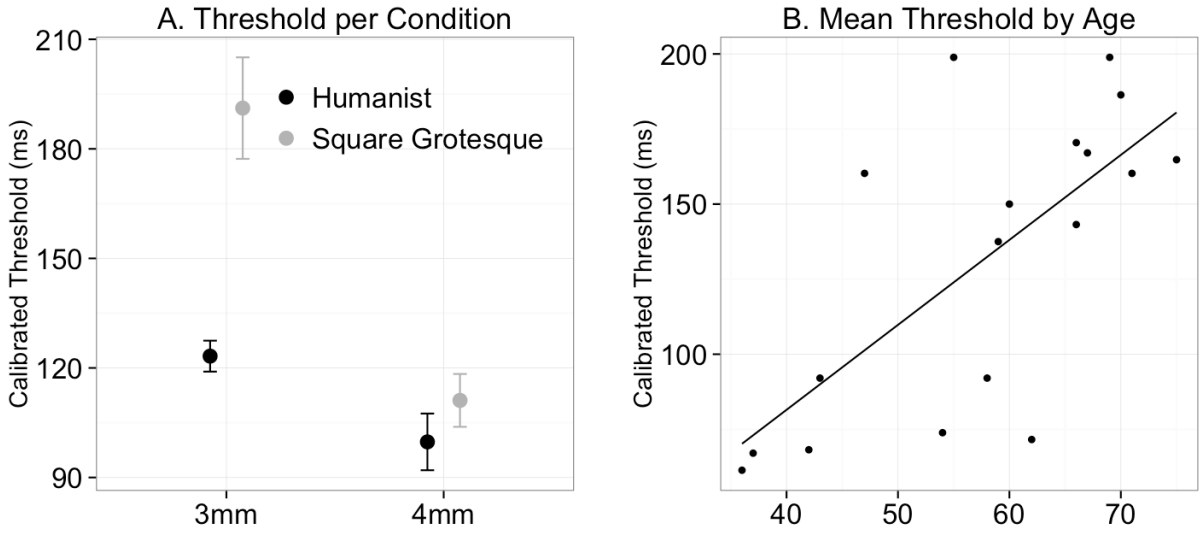

Figure 3. A) Mean stimulus duration thresholds for each condition. Error bars represent ±1 mean-adjusted standard error. B) Mean threshold per participant plotted against that participant's age. Solid line represents a linear regression through the data

Stimulus duration thresholds are shown in FigureA. Thresholds were significantly higher for the square grotesque typeface compared to the humanist across size conditions ($F_{(1, 17)} = 16.05$, $p < 0.001$). Thresholds were also significantly elevated in 3mm conditions compared to 4mm ($F_{(1, 17)} = 15.91$, $p < 0.001$). As is evident in FigureA, typeface and display size interacted significantly ($F_{(1, 17)} = 13.94$, $p = 0.002$). Legibility thresholds for the square grotesque typeface were strongly affected by the switch to a smaller display size, while the humanist typeface exhibits a less pronounced (though still significant) threshold increase.

As an added check on the data, thresholds from the 4mm conditions were compared to data from a condition in a previous study (Dobres, Chahine, Reimer, Gould, Mehler, Pugh, et al., 2014b) that utilized the same font size, contrast polarity, and typefaces. Threshold estimates were not significantly different between studies ($F_{(1, 64)} = 0.98$, $p = 0.325$). Thresholds between these studies differed by just 12.9ms. This difference is less than the 16.7ms monitor refresh rate used in the earlier study (the smallest increment by which thresholds could be adjusted in that study).

We find that thresholds significantly increase with age ($F_{(1, 16)} = 15.10$, $p = 0.001$). As shown in FigureB, thresholds become substantially elevated as age increases, particularly after the age of 65. Under this model, we would expect the mean 79.4% legibility threshold for a 40 year-old to be approximately 81ms, compared to 166ms for a 70 year-old (an increase of 105%).

DISCUSSION

Figure 3. Samples of typefaces as displayed in actual screen pixels at 4mm (13 pixel capital height) and 3mm sizes (10 pixel capital height) for humanist (top 2 rows) and square grotesque (bottom 2 rows). Image taken directly from the Psychtoolbox frame buffer, zoomed to show rendering artifacts

Consistent with previous work using these typefaces, the humanist typeface was more legible than the square grotesque typeface across all conditions (Dobres, Chahine, Reimer, Gould, Mehler, Pugh, et al., 2014b; Reimer et al., 2014). There was also a pronounced effect of size, with the 3mm conditions producing greatly elevated thresholds compared to 4mm, particularly for the square grotesque typeface. As illustrated in Figure 3 (zoomed for detail), the humanist typeface (top 2 rows) scales more cleanly at the 3mm size, and its letterforms remain largely intact. In contrast, the square grotesque typeface (bottom 2 rows) degrades considerably at the smaller size. This is especially noticeable in the 'i' and 'j' characters, which lose distinguishing features in the square grotesque typeface at 3mm. The square grotesque's muddled 'a' and 'g' characters at 3mm are also notable in contrast to the humanist's stronger letterforms at 3mm.

The Psychtoolbox employs a grayscale font smoothing algorithm, and while it is suboptimal, it is still fairly common in many software architectures, particularly those used in industries where display quality is not a priority. The degradation of quality shown in Figure 3 demonstrates that seemingly subtle typographic aesthetics, such as the humanist's more open letterforms and varied shapes, can translate not only to greater *intrinsic* legibility, but can also strongly affect how the typeface interacts with *extrinsic* factors like the quality of display media and the rendering algorithm.

These results suggest the typographic choices can significantly impact legibility, and moreover, that one's sensitivity to legibility effects increases with age. Since older drivers drive frequently and are also the top buyers of new vehicles (Naughton, 2013), the legibility effects demonstrated in this research should be a point of concern for HMI designers. It should be clear that "minimum legible size" is not an acceptable metric for the majority of new car buyers, and thus the aesthetics of the interface must be balanced with the needs of an aging population. One approach to this may be to develop interfaces with enhanced accessibility settings that allow the driver to adjust the scaling of onscreen elements. It should be noted that this and other related techniques present a promising avenue for the investigation of a wide variety of design features, but does not necessarily address all aspects of the impact of design on legibility.

ACKNOWLEDGMENTS

Support for this work was provided by the US DOT's Region I New England University Transportation Center at MIT and the Toyota Class Action Settlement Safety Research and Education Program. 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 and typographic expertise.

REFERENCES

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. doi:10.1163/156856897X00357

Carter, H. (1984). Optical Scale in Type Founding. *Printing Historical Society Bulletin*, 144–148.

Chaparro, B. S., Shaikh, A. D., Chaparroa, A., & Merkle, E. C. (2010). Comparing the legibility of six ClearType typefaces to Verdana and Times New Roman. *Information Design Journal*, *18*(1), 36–49. doi:10.1075/idj.18.1.04cha

Dobres, J., Chahine, N., Reimer, B., Gould, D., & Mehler, B. (2014a). A Pilot Study Measuring the Relative Legibility of Five Simplified Chinese Typefaces Using Psychophysical Methods. Presented at the 2014 International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Seattle, WA.

Dobres, J., Chahine, N., Reimer, B., Gould, D., Mehler, B., Pugh, B., & Arredondo, S. (2014b). An Art Meets Science: Subtle Typeface Design Characteristics Affect Word Legibility in Brief Glances. *Vision Science Society Annual Meeting*. St. Pete Beach, FL.

International Standards Organization. (2009). *Ergonomic aspects of transport information and control systems* (No. 15008). Geneva, Switzerland.

Legge, G. E., & Bigelow, C. A. (2011). Does print size matter for reading? A review of findings from vision science and typography. *Journal of Vision*, *11*(5). doi:10.1167/11.5.8

Legge, G. E., Pelli, D. G., Rubin, G. S., & Schleske, M. M. (1985). Psychophysics of reading--I. Normal vision. *Vision Research*, *25*(2), 239–252.

Medler, D. A., & Binder, J. R. (Eds.). (2005). *MCWord*. Retrieved December 13, 2013, from http://www.neuro.mcw.edu/mcword/

Naughton, K. (2013, August 5). Boomers Replace Their Children as No. 1 Market for Autos. *Bloomberg*. Retrieved November 7, 2014, from http://www.bloomberg.com/news/2013-08-05/automania-strikes-boomers-supplanting-kids-as-buyers.html

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from http://www.R-project.org/

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. doi:10.1162/neco.2008.12-06-420

Reimer, B., Mehler, B., Dobres, J., Coughlin, J. F., Matteson, S., Gould, D., et al. (2014). Assessing the impact of typeface design in a text-rich automotive user interface. *Ergonomics*, 1–16. doi:10.1080/00140139.2014.940000