



Life Tomorrow



White Paper 2014-3

Revealing Differences in Legibility Between Typefaces Using Psychophysical Techniques: Implications for Glance Time and Cognitive Processing

Jonathan Dobres, Nadine Chahine, Bryan Reimer, Bruce Mehler, & Joseph Coughlin

This study adapts classical vision science techniques to an investigation of the relative legibility of two different typefaces across two different polarity (color) conditions. Participants performed a simple yes/no lexical decision task, with task difficulty controlled by an adaptive staircase in four typeface/polarity conditions. Stimulus duration thresholds (minimum time on screen needed for accurate reading) were sensitive to differences between both typefaces and polarities, with a humanist style typeface showing a legibility advantage compared to a square grotesque style typeface, and positive polarity text (black on white) showing a legibility advantage compared to negative polarity text (white on black). Legibility thresholds were found to increase with age. As expected, reaction time measures were not sensitive to differences in typeface or polarity, but they did reveal cognitive processing differences between correct and incorrect responses, as well as differences in processing words and pseudowords. There was also some evidence that switching to a new typeface and/or polarity may incur a quantifiable “task switching cost”. This study is broadly consistent with MIT’s previous effort to assess the impact of typestyle on interface demand in a simulated driving environment. We believe that this simplified, resource efficient methodology of assessing legibility differences can be adapted to investigate a wide array of questions relevant to typographic and graphic design in automotive as well as other interfaces.

Introduction

The MIT AgeLab, in collaboration with Monotype Imaging, has previously investigated the effects of typeface on the demand of human-machine interactions during a simulated in-vehicle point-of-interest (POI) menu selection task. Reimer et al. (in press) report on the results of two studies of positive polarity (black-on-white) text. Participants performed the menu selection task while driving a fixed-based driving simulator. The in-vehicle device’s menu options were set in Frutiger, a “humanist” typeface, and Eurostile,

a “square grotesque” typeface. The results of the studies indicated that participants were able to perform the menu selection task more quickly and more accurately when menus were set in humanist type as compared to square grotesque, and that this effect was more pronounced for men as compared to women. A follow-up study attempted to replicate these findings for negative polarity (white-on-black) text and showed similar trends, though the effect sizes were smaller and did not reach statistical significance. A further study that used a brighter monitor setting to display menu items found a significant effect of typeface for both men and women, though the benefits of humanist type were once again more pronounced among men (9.1% less glance time compared to square grotesque for the men, 3.3% for women).

These initial results indicate that empirical methods can be applied to the investigation of legibility differences based upon the aesthetic characteristics of typefaces. However, given the number of possible characteristics (aesthetic and other), scenarios, and languages that could be tested, conducting tests in a full driving simulator environment would require a prohibitively large investment of resources and time. Moreover, the results of a pilot exploration of negative polarity (white on black) text suggested that, although differences between typefaces appear in the simulator, this approach may be somewhat underpowered in a modestly sized sample, and thus unable to detect more subtle differences in typeface characteristics. Although a fully simulated environment provides excellent face validity, it forces the investigator (be he/she a scientist, engineer, designer, etc.) to examine visual design characteristics in the complex context of the driving task, which places constant demands on visual attention, risk management, hand-eye coordination, and situational awareness. Significant design effects might be “swallowed” by the larger influences of ingrained driving behaviors and the demands of the driving environment, and thus be missed in the simulator. Lastly, a methodology that is specifically bound to a driving simulator calls into question whether its findings are valid outside of the vehicle or roadway environment. Therefore, a methodology that more directly examines the influence of typeface design characteristics on visual behavior is needed, ideally one that is both versatile and cost effective.

Psychophysics: Measuring Perception

“Psychophysics” refers to the scientific methods that investigate the relationship between the *physical* properties of a stimulus (brightness, contrast, presentation time, etc.) and the *psychological* effects (perceptions) they generate. Psychophysics traces its origins to the very beginning of experimental psychology in the mid-19th century, and its basic methodologies have been used to investigate everything from low-level perceptual mechanisms in humans and animals (Blough, 1958; Jameson, Highnote, &

Wasserman, 2001) to higher-level phenomena such as scene perception and the principles of learning (Dobres & Watanabe, 2012; Henderson & Hollingworth, 1999). This family of methods is, in many ways, ideal for investigations of typefaces. Psychophysical studies do not require elaborate equipment and can usually be conducted using a standard desktop computer and screen. Usable data can be collected rapidly and at minimal cost.

There is already a considerable body of psychophysical research on the topic of legibility. Some of the earliest psychophysical investigations concerned the legibility of the English alphabet (Sanford, 1888), which has since been carried over into modern investigations of digital typography (Beier & Larson, 2010; Chaparro, Shaikh, Chaparro, & Merkle, 2010; Fox, Chaparro, & Merkle, 2007). Most of these legibility investigations concern relatively fundamental visual or cognitive features, such as size (Huang, Patrick Rau, & Liu, 2009; Legge & Bigelow, 2011), digit span (Chien-Hsiung Chen & Chien, 2005), lexical frequency (Grainger & Segui, 1990; Yan, Tian, Bai, & Rayner, 2006), spatial frequency (Paterson, McGowan, & Jordan, 2013), visual crowding (Pelli & Tillman, 2008; Pelli et al., 2007; L. Wang et al., 2008), and reading in peripheral vision (He, Legge, & Yu, 2013; Legge, Mansfield, & Chung, 2001).

There have been relatively few psychophysical investigations of the effects of typeface on legibility, at least at the level of the *design* of the typeface itself (as described above, investigations of spacing, color, contrast, and other “extrinsic” factors are much more common). However, there have been isolated studies that investigate the design properties of typefaces as they relate to reading and comprehension in visually-normal and visually-impaired children (Bessems, 2012), how intrinsic design factors mediate the perception of type, (A.-H. Wang & Chen, 2003), and the design of a typeface for a specific purpose (Vinot & Athenes, 2012). Wang and Chen (2002) found that extrinsic properties mediated legibility, but that typeface design itself did not, whereas Vinot and Athènes (2012) attempted to iteratively design a new typeface, as opposed to comparing existing and commonly used typefaces. Other research has examined differences between traditional paper-based reading and screen-based reading, generally finding that reading from a screen is inferior to reading from paper, and that the difference in performance is primarily attributable to factors that affect the intrinsic quality of on-screen type, such as screen resolution and typeface design, rather than extrinsic factors such as monitor brightness and contrast (Gould et al., 1987a; Gould, Alfaro, Finn, Haupt, & Minuto, 1987b).

In the study described here, we develop a methodology that allows us to rapidly assess the legibility of two typefaces—the same as those used in our previous simulator work. We also examine the effect of text polarity (the choice of foreground and background

colors for the text display), by comparing the two typefaces under positive polarity (black on white) and negative polarity (white on black) conditions. The methodology works by presenting words on screen for a very brief duration, enforcing glance-like behavior. In this way, the methodology parallels the occlusion testing standard commonly used in the driving research field, with a much simpler and easily reproducible setup. The difficulty of the task is determined by the length of time that words are presented on-screen, and difficulty is continuously calibrated until a stable level of task performance accuracy is reached. A more legible typeface should require a shorter on-screen presentation time (“stimulus duration”) to reach equivalent levels of difficulty.

Methods

Pilot Testing

Several pilot studies were conducted to optimize the parameters of the experimental method and gauge the reliability of experimental measures. Samples were drawn from a mixture of internal staff and participants recruited from the Boston area. The pilot studies utilized a split design, in which participants were first calibrated to a target level of difficulty, and then data on participants’ reaction time to stimuli presented in each typeface were collected at this individualized (fixed) difficulty level. The first pilot study collected data from a small sample of 9 department employees, and indicated that the experimental method and its primary dependent measure (reaction time) were able to expose differences between typefaces and color conditions. This testing was expanded to a small external sample of 10 participants. After minor adjustments to the experiment protocol, a further 18 participants were recruited from internal and external sources. At this point, we determined that although our chosen measures were working well, the experiment method itself would benefit from being shortened. A total of 49 participants were run using the shortened methodology, at which point it became clear that the reaction time measure was not as sensitive to differences between typefaces or colors as one would prefer in an optimal methodology that minimizes the number of experimental participants required to observe a statistically significant effect. We deemed the design to be somewhat overcomplicated, and reconfigured it to use the difficulty thresholds themselves, and not reaction times, as the primary dependent measure. A description of that finalized experiment protocol follows.

Participants

A total of 67 participants between the ages of 20 and 75 were recruited for the primary study. All participants gave their written, informed consent to participate, as outlined by the institutional review board of the Massachusetts Institute of Technology. To ensure that the sample was similar to those recruited for the AgeLab’s previous typographic

research (Reimer et al., in press), participants were required to have driven a car at least once per week and to be in reasonably good health for their age. Exclusion criteria included experience of a major medical illnesses or hospitalization in the last six months, conditions that impair vision (other than typical nearsightedness or farsightedness), or a history of epilepsy, Parkinson’s disease, Alzheimer’s disease, dementia, mild cognitive impairment, or other 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. Corrected near and far visual acuities did not differ significantly between genders ($p > 0.05$ for all statistical comparisons of visual acuity, Wilcoxon signed rank tests).

Of the 67 participants, 11 failed to reach a stable stimulus duration threshold (see “miscalibration”, below) and were excluded from analysis. Six were excluded due to technical problems with the equipment or software. Two participants were excluded from analysis because the target sample distribution had already been reached. This left a total of 48 participants, equally split between males and females. Age distribution did not differ significantly between genders ($t_{(45)} = 0.34$, $p = .737$). Summary statistics for men and women are given in Table 1.

Table 1: Sample sizes, mean, standard deviation, and range of ages for men and women.

Gender	n	Mean	Standard Deviation	Range
Female	24	46.3	12.8	25 - 64
Male	24	45.0	14.4	23 - 65

Task, Apparatus, & Stimuli

Task

This experimental methodology reduces the test of legibility to a 1-interval forced choice decision task. A schematic of the task is presented in Figure 1. Each trial begins with a 1000ms display of a fixation rectangle (200px by 100px), centered on the screen, indicating the general area where stimuli will appear (all stimuli and masks are displayed at the screen’s exact center). The fixation rectangle is followed by a 200ms mask composed of non-letter characters. Then a single word (or pseudoword) stimulus is displayed for a variable presentation time, as determined by an adaptive staircase procedure (see below). This is immediately followed by another 200ms mask. Finally, the participant is prompted to decide whether the stimulus was a word or pseudoword.

Participants are given a maximum of 5000ms to respond by pressing one of two keys on a numeric keypad (one button corresponds to “word”, the other to “pseudoword”). Subjects were not provided with feedback regarding the accuracy of their responses, other than during the practice section. Each mask was unique, constructed by randomly selecting eight characters from a small pool of non-letter characters. The sandwiching of the stimulus between the two masks minimizes the stimulus’s visible persistence in iconic memory, ensuring that it will only be perceptually accessible for the intended presentation time (Coltheart, 1980).

The experiment began with a series of ten practice trials, with stimulus duration set to 1000ms. After five consecutive correct answers, participants were permitted to move on to the main experiment. If the participant reached the end of the ten trials without making five consecutive correct responses, he/she was allowed to repeat the practice block. If the participant was still unable to complete the practice at this minimum performance criterion, he/she would be excluded from the experiment (no participants were excluded for this reason).

Primary data collection (400 trials total) began after the practice block. Every 50 trials (approximately every 4-5 minutes), participants were allowed to take a short rest of up to 30 seconds (the participant could terminate the rest periods early if so desired). There was a mandatory 5-minute break after the 200th trial, during which participants listened to an audio recording. The recording contained a short selection from a biography of Benjamin Franklin, and was provided only to fill time during the extended break. Data collection lasted approximately 30 minutes.

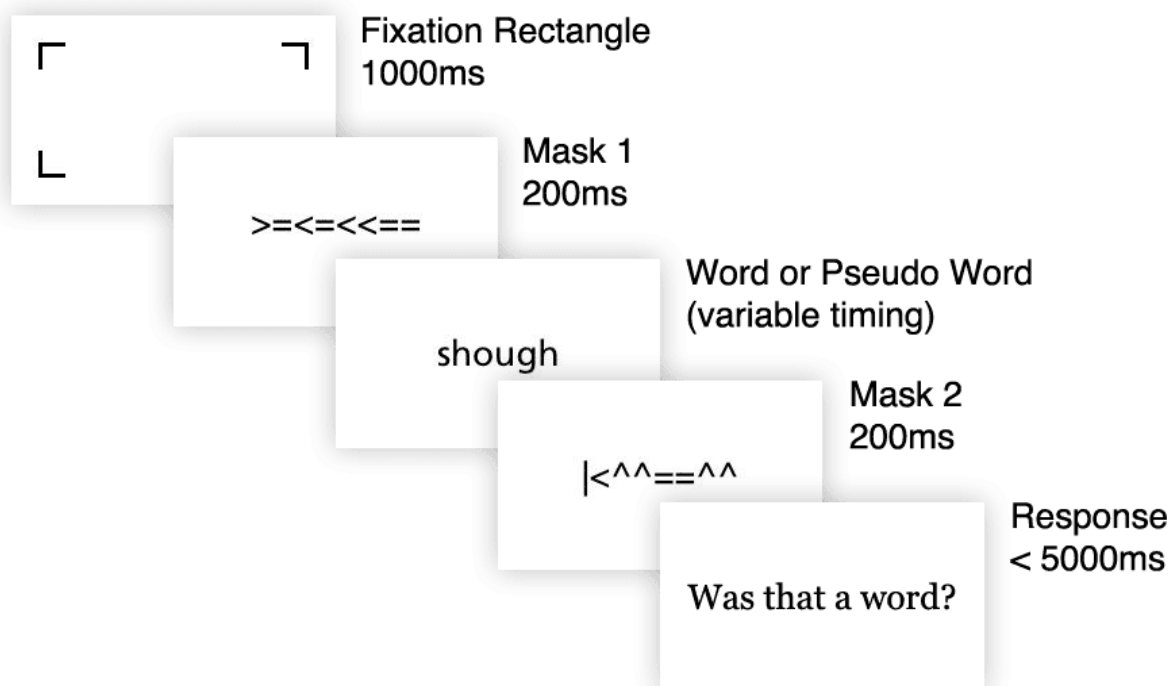


Figure 1: The structure of an individual trial of the experiment. See Methods for details.

Apparatus

The experiment was run on a 2.4GHz Mac Mini running Mac OS X 10.6.8. Stimuli were created and displayed using Matlab (Natick, MA) running the Psychtoolbox 3 (Brainard, 1997; Pelli, 1997). The Psychtoolbox software is specifically designed for this type of research, and contains a variety of tools geared toward the precise and reliable control of stimulus timing. Stimuli were displayed on a Dell 24" (60.96cm) LCD monitor with its brightness set to the lowest possible level (1 cd/m² when displaying pure black, 113 cd/m² when displaying pure white). The monitor had a resolution of 1920 x 1200 pixels and a refresh rate of 60Hz. The experiment was conducted in a quiet, dimly lit room.

Stimuli

The primary stimuli of this experiment were words selected from an online orthographic database (Medler & Binder, 2005). To generate a suitably large list of reasonably common words, word length was restricted to 6 letters, orthographic neighborhood size was restricted to between 1 and 5 (inclusive), word frequency was set to 2-5 per million (inclusive), and constrained bigram frequency was set to a minimum of 600 per million. All other search parameters were unconstrained. Pseudowords, also 6 letters long, were generated from the same database using constrained trigrams. This resulted in pseudowords made of pronounceable combinations of letters, and closely resembled

the list of real words. The resulting pools of words and pseudowords are provided in Appendix A.

This experiment included a total of 4 experimental conditions: 2 typefaces x 2 polarities (100 trials per condition). The typefaces were “Frutiger” (a humanist typeface) and “Eurostile” (a square grotesque). Standard versions of Frutiger and Eurostile were modified to equalize their optical heights based on the height of each typeface’s capital “H” (Reimer et al., in press). The positive polarity condition displayed black text (RGB: 0, 0, 0) on a white background (RGB: 255, 255, 255), while the negative polarity condition displayed the opposite (same color values). Samples of each typeface and polarity are shown in Figure 2. Each combination of polarity and typeface was presented in a separate block, and the order of blocks was counterbalanced across participants. Each typeface/polarity condition contained 50 word trials and 50 pseudoword trials, randomly interleaved. Word order was randomized for each participant. Polarity conditions were always presented consecutively. For example, a participant might see all positive polarity trials first, followed by all negative polarity trials, or vice versa. The transition between polarities always occurred after the extended break after the 200th trial.

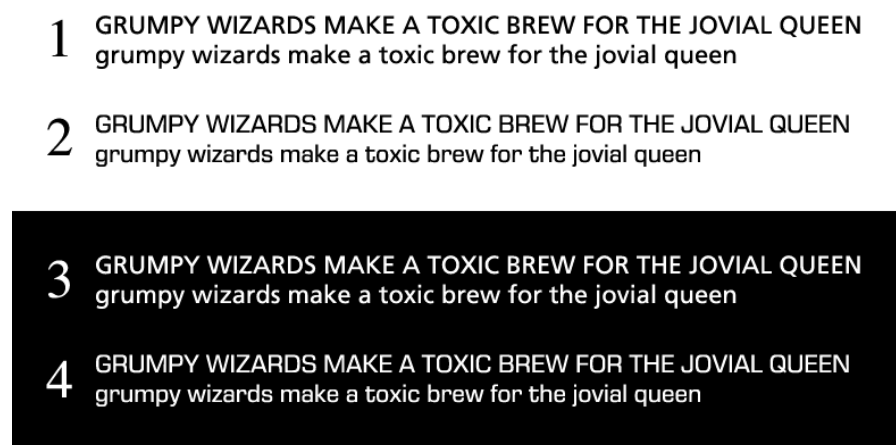


Figure 2: Pangram type samples for the four combinations of typeface and polarity (color) used in this experiment. Figure displays all uppercase and lowercase letters, though only lowercase text was used in the experiments. Samples 1 and 3 display Frutiger; samples 2 and 4 display Eurostile. Figure was rendered in Adobe Photoshop CS5.

Typefaces were scaled such that their capital letter height was 4mm on screen, though all stimuli were presented in lowercase type. Whereas the previous simulator-based study presented the typefaces of interest on pre-rendered graphics mimicking a menu system, in the present study, text was rendered “live” using Matlab’s default algorithms, which in turn rely on the operating system’s (Mac OS) font rendering. Notably, Matlab’s font rendering does not support sub-pixel anti-aliasing.

To mimic the fixed visual distance of an automotive interface, participants were seated such that their eyes were approximately 27" (68.58cm) from the screen, and were instructed to try to maintain that distance throughout the session (word stimuli were therefore displayed at an optical size of approximately 20.1 arcmin). Head restraints were not used, thus allowing for the kind of positional variability that is likely to be encountered in real-world reading scenarios. The 4mm screen character height and the distant positioning of the participants' eyes from the screen were consistent with our earlier driving simulation study (Reimer et al., in press) and ISO standard 15008 (International Standards Organization, 2009) for automotive displays, which recommends an effective character size ≥ 20 arcmin.

A serif typeface that looked substantially different from the two typefaces of interest, "Georgia", was used to display practice trial stimuli and all prompt text. Text set in Georgia was also displayed at approximately double the size of the word and pseudoword stimuli.

Adaptive Staircase Procedures

During the four main data collection blocks, task difficulty was controlled via an adaptive staircase procedure (Leek, 2001; Levitt, 1971). This technique changes the difficulty of the task based on the participant's pattern of correct and incorrect responses. Using a "3-down, 1-up" rule, the task is made more difficult (stimulus duration is decreased) after three consecutive correct responses, and made easier (stimulus duration is increased) after one incorrect response. Following this rule, stimulus duration will converge on a difficulty that produces 79.4% accuracy (Leek, 2001). In this "sweet spot", accuracy is high enough to ensure that the participant is not randomly guessing, but still low enough to ensure that the participant is experiencing some degree of uncertainty in his/her responses, and thus may be affected by factors such as subtle differences in typeface legibility. The movement of a typical staircase is shown in Figure 3 (left panel).

Staircase levels were reset at the start of each typeface/polarity block, allowing for the calculation of separate stimulus duration thresholds for each of the 4 conditions (stimulus duration was constrained to be at least 33.4ms and at most 1000ms). Thresholds were calculated as the median presentation time during the final 20 trials of each condition. These thresholds form our primary dependent measure. Each condition is calibrated to the same hypothetical accuracy level. Therefore, a less legible typeface should require a longer presentation time (and thus a higher threshold) to reach the same accuracy level as a more legible typeface.

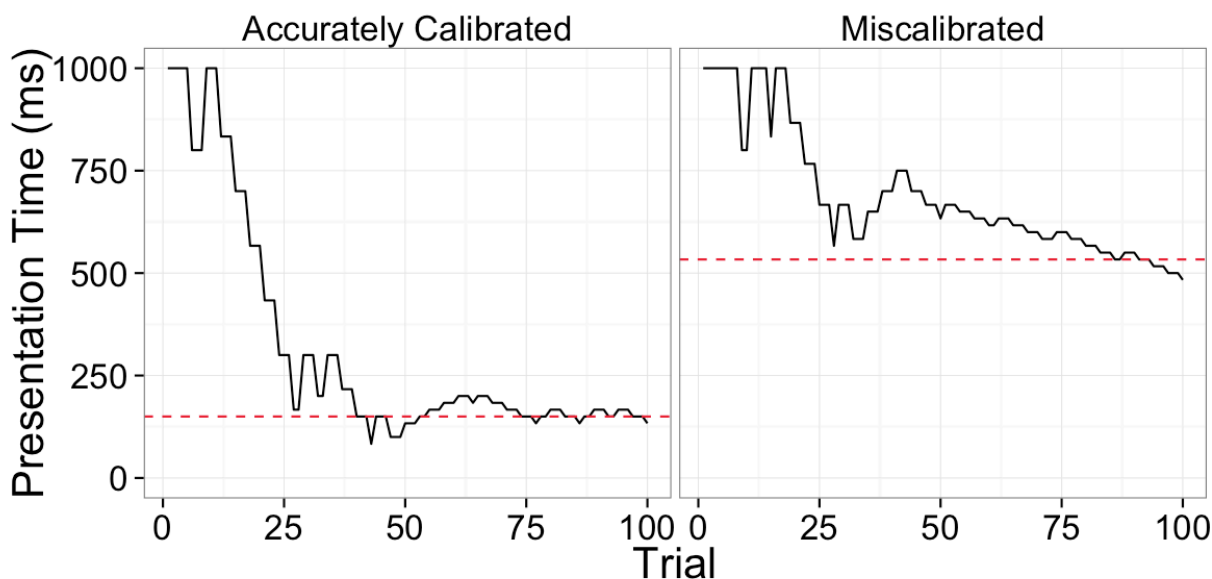


Figure 3: Selected data showing accurately calibrated (left) and miscalibrated (right) thresholds. The solid lines show the staircase values (time on screen) as trials proceed. The dashed red lines show the calculated threshold levels (median of the last 20 trials of the staircase).

Miscalibration

Each condition resets the staircase to a stimulus duration of 1000ms. This is an “easy” setting that gives the participant ample time to read the single stimulus word (or pseudoword). As the participant makes more correct responses, stimulus duration rapidly “steps” downward, until the participant makes an incorrect response and triggers a reversal (a change from a series of correct responses to incorrect responses, or vice versa). At that point, the step size is reduced, on the assumption that it has neared the participant’s accuracy threshold and will need finer-grained control of stimulus difficulty to converge on the threshold. The minimum possible step size is dependent on the refresh rate of the monitor being used (in this case, 16.7ms). Figure 3 (left panel) shows a nearly ideal staircase. Stimulus duration decreases rapidly until the participant’s threshold (red line) is approached around the 25th trial, at which point the staircase begins a long series of reversals in smaller and smaller steps that revolve around the participant’s true threshold (i.e. minimum presentation time to achieve 79.4% accuracy).

However, a series of early incorrect responses, whether due to unintended button presses or coincidental difficulty with the vocabulary, can produce erroneous reversals that prematurely slow the movement of the staircase, resulting in failure to reach the participant’s true threshold during the allotted trials. Figure 3 (right panel) shows a miscalibrated staircase. A series of early reversals reduces the step size too quickly, and as a result, the staircase has not settled around a threshold and is still moving

downward steadily at the end of the condition. This results in an inaccurate threshold estimate.

Participants were excluded from analysis if any of their calculated threshold values was greater than 300ms, or if a participant's staircase was still in the process of steadily descending when the condition ended. Each participant experienced 4 conditions, and therefore, 4 staircases. This resulted in a somewhat higher incidence of miscalibrations than we had anticipated. We believe that the staircase procedures can be further optimized to reduce the chances of miscalibration (see Discussion).

Data Analysis

As previously stated, thresholds were obtained for each of the 4 typeface/polarity conditions by calculating the median stimulus duration of each condition's final 20 trials. In addition to response accuracy, reaction times were also recorded for each trial. Primary data were analyzed in a 2 x 2 repeated-measures design (typeface x polarity). Although we included participants across a wide 20-75 year age range, an investigation of the differences in perception due to age was not a goal of the present study. Therefore, in most statistical tests we include gender as predictor and age as a covariate of the main effects of interest. All statistics were computed and visualized using R (R Core Team, 2014).

Results

Response Accuracy

Since task difficulty fluctuates in accordance with the staircase, mean response accuracy was calculated for the last 20 trials of each condition, when the staircase had stabilized for most participants. Summary statistics are shown in Table 2. As expected, response accuracy did not differ significantly between typeface ($F_{(1, 46)} = .06, p = .805$) or polarity ($F_{(1, 46)} = .20, p = .655$). These factors also did not interact significantly ($F_{(1, 46)} = .62, p = .436$). This indicates that participants were calibrated to a consistent level of accuracy across conditions, which is to be expected and is a goal when using adaptive threshold procedures. In other words, each calibrated threshold (one for each typeface and polarity) represents the amount of time that a stimulus had to be displayed on screen, such that the participant could read it with 79.4% accuracy. Across conditions, response accuracy was 78.8% on average, which is not statistically different from the staircase calibration point of 79.4% ($t_{(47)} = -0.94, p = 0.350$). This indicates that the staircase was able to converge on a stable threshold estimate in the allotted time.

Table 2: Means (and standard deviations) of response accuracy for each of the 4 conditions.

Typeface	Positive Polarity	Negative Polarity
Humanist	79.0% (7.4%)	78.4% (7.7%)
Square Grotesque	78.3% (7.0%)	79.6% (7.5%)

Reaction Time Changes

It has been suggested that reaction times reflect the amount of time needed to process stimuli relevant to a decision. A longer reaction time indicates that a greater amount of cognitive “computation time” is necessary before a decision point is reached. If reaction times therefore reflect a kind of cognitive uncertainty, we may find that incorrect responses have slower reaction times compared to correct ones (Ratcliff & McKoon, 2008; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008). This is indeed the case in the present data. Reaction times were significantly slower for incorrect responses compared to correct ones (576ms vs. 453ms, respectively, $F_{(1, 46)} = 58.3$, $p < .001$). Put another way, reaction times for incorrect responses were 27.2% slower compared to correct responses. Similarly, reaction times to pseudoword trials were significantly slower compared to word trials (492ms vs. 452ms, $F_{(1, 46)} = 20.9$, $p < .001$), suggesting that participants needed more time to process more novel or linguistically confusing stimuli.

Figure 4 shows the average reaction time (binned by 10 trials) across the 30 minute session. The data are presented in chronological order (each color in the figure corresponds not to a typeface/polarity, but to the subjects’ 1st, 2nd, 3rd, and 4th conditions in the order presented). Reaction time at first decreases rapidly as the participant becomes more comfortable with the experiment, and then levels off. Reaction time generally declines within conditions (colors in Figure 4), even after the mid-condition rest periods (a rest period is denoted by a break in the lines in Figure 4). However, when a new condition is begun, reaction time temporarily increases, and then undergoes another fairly rapid decrease. This can be quantified statistically by computing a reaction time slope for each block (a group of 5 bins in Figure 4). Slopes for each condition’s first block average -25ms, whereas slopes for each condition’s second block average a marginal -0.2ms. The difference in slopes is highly significant ($F_{(1, 47)} = 16.3$, $p < .001$).

Participants could opt to terminate the rest periods early (other than the 5-minute mid-experiment rest), raising the possibility that this reaction time effect might be due to coincidental differences in rest lengths. However, excluding the mid-experiment rest, there were no significant differences in rest lengths ($F_{(1, 47)} = 0.1$, $p = .771$). We hypothesize that the difference in reaction times observed between conditions may be

due to a “task switching cost”, a cognitive effect in which switching to a new task requires an adjustment period (Monsell, 2003). The presence of task switching costs here is consistent with earlier work suggesting that the visual system tunes its letter recognition capabilities as more time is spent reading a typeface, and that processing costs are incurred when typefaces are mixed (Gauthier, Wong, Hayward, & Cheung, 2006; Sanocki, 1987; 1992; Sanocki & Dyson, 2011; Walker, 2008). Also of note, reaction times appear to increase substantially during the final block, which may be indicative of fatigue, distraction, or both.

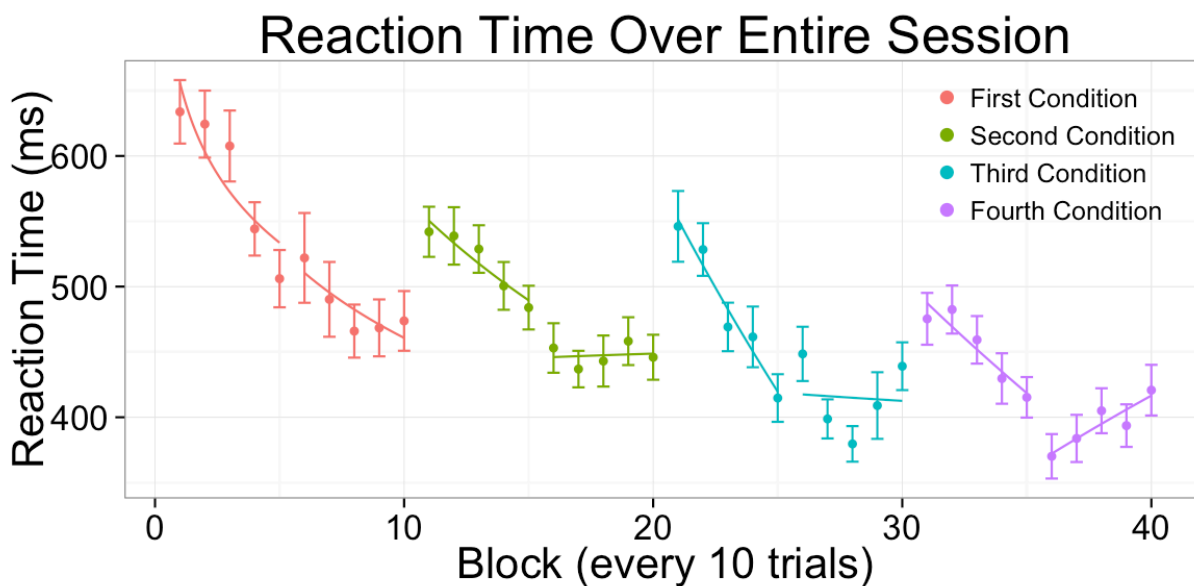


Figure 4: Reaction time over the course of the entire experiment session. Each color corresponds to one of the four experimental conditions, presented here simply in the order in which participants saw them. Breaks in the lines represent rest periods (every 50 trials, or every 5 points here). Reaction time decreases sharply until the first rest period, then stabilizes. Note that reaction times do not increase after rests that are part of the same condition (same color), but do increase whenever a new condition is started.

Stimulus Duration Threshold Differences

Stimulus duration thresholds (again, calculated as the median stimulus duration during the last 20 trials of each condition) are presented in Table 3 and Figure 5. Thresholds for the humanist typeface were significantly shorter than thresholds for square grotesque ($F_{(1, 46)} = 7.32, p < .01$), suggesting that humanist type is more legible. In essence, the humanist typeface required less viewing time to formulate an accurate response. Thresholds were also significantly lower for positive polarity (black on white) text than for negative polarity (white on black) ($F_{(1, 46)} = 55.3, p < .001$). Typeface and polarity did not interact significantly ($F_{(1, 46)} = 0.44, p = .510$), suggesting that the humanist typeface carries the same legibility benefit regardless of polarity condition.

There was no significant difference in thresholds between genders ($F_{(1, 46)} = 0.03$, $p = .863$).

Table 3: Means (and standard deviations) of threshold presentation times (in ms) for each of the 4 conditions.

Typeface	Positive Polarity	Negative Polarity
Humanist	82.3 (31.4)	112.7 (49.3)
Square Grotesque	88.2 (42.3)	124.0 (57.1)

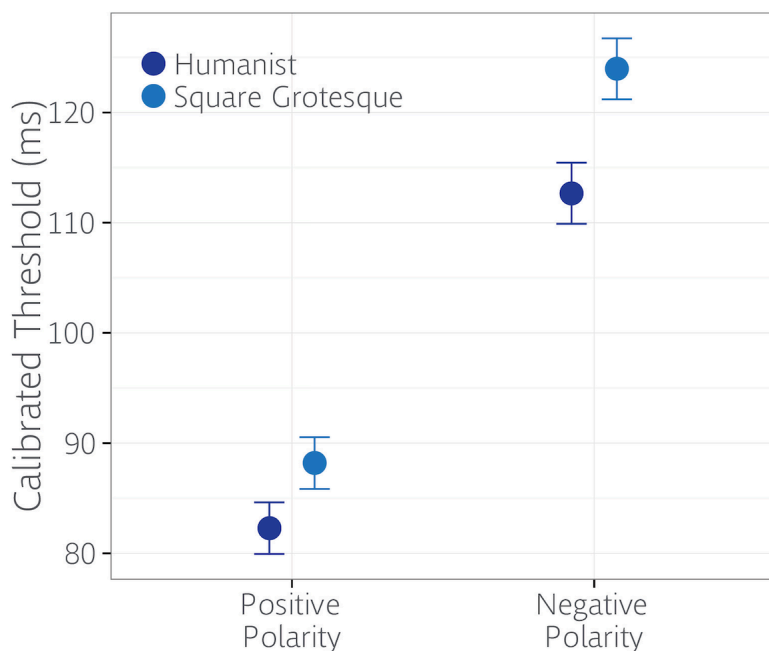


Figure 5: Calibrated presentation time thresholds for each condition. Note that thresholds are consistently lower for the humanist typeface compared to the square grotesque, and thresholds are consistently lower in the positive polarity condition (black on white) compared to the negative (white on black).

Although age effects were not a primary concern of the present study, the data do clearly demonstrate that stimulus duration thresholds across conditions increase significantly with age, as illustrated in Figure 6 ($t_{(47)} = 3.24$, $p < .01$). Stimulus duration thresholds among 20 year-olds averaged 70ms, versus 126ms for 65 year-olds, an increase of 81%. These findings are consistent with various well known age-related declines in perceptual processing (Faubert, 2002; Govenlock, Taylor, Sekuler, & Bennett, 2009; Habak & Faubert, 2000; Snowden & Kavanagh, 2006).

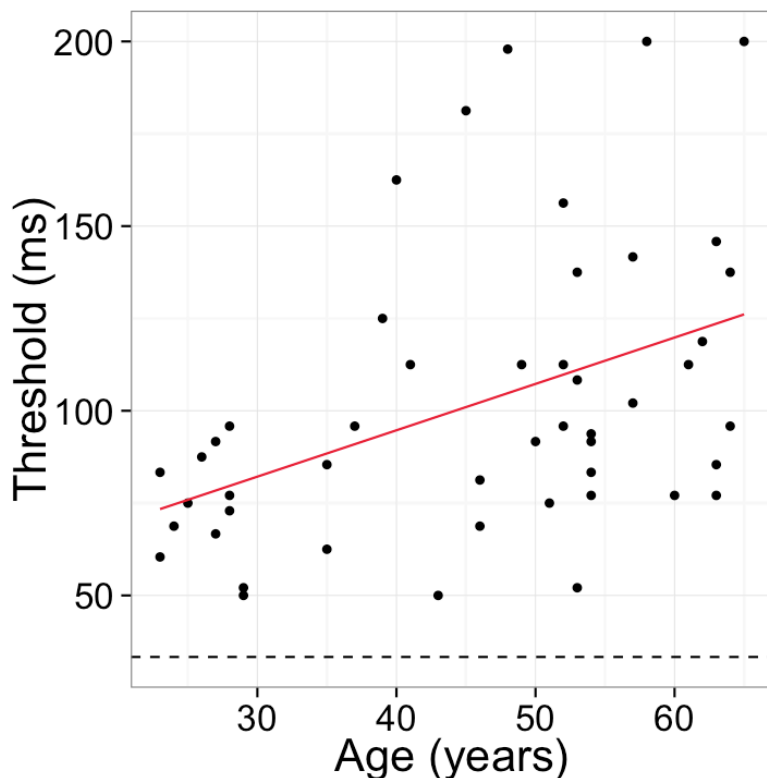


Figure 6: Each participant's average threshold across the 4 typeface/polarity conditions, visualized against the participant's age. The red line represents a simple linear regression through the data. The horizontal dashed line represents the shortest possible presentation duration allowed by the staircase procedure (33.4ms).

Discussion

Summary of Present Work

The present study adapted classical psychophysical techniques to an investigation of the relative legibility of two different typefaces across two different polarity (color) conditions. Participants performed a simple yes/no lexical decision task, with task difficulty controlled by an adaptive staircase in each of the four typeface/polarity conditions. The stimulus duration thresholds of these staircases (the time on screen necessary to reach 79.4% accuracy in each typeface/polarity condition) formed the primary dependent measure. We found that stimulus duration threshold levels were sensitive to differences between both typefaces and colors, with a humanist typeface showing a legibility advantage compared to a square grotesque. Stimulus duration thresholds were 8.8% shorter for humanist typefaces compared to square grotesque (considering participants 35 and older, this difference rises to 12.4%). Positive polarity text (black on white) showed a strong legibility advantage, with average stimulus durations 38.6% shorter than negative polarity text (40.6% considering participants 35

and older). Breaking down the typeface differences by polarity, humanist thresholds were 6.7% shorter in the positive polarity condition and 9.1% shorter in the negative polarity condition. Additionally, although reaction time measures were not sensitive to differences in typeface or polarity, they did reveal cognitive processing differences between correct and incorrect responses, as well as differences in processing words and pseudowords. There was also some evidence that switching to a new typeface and/or polarity may induce a quantifiable “task switching cost”.

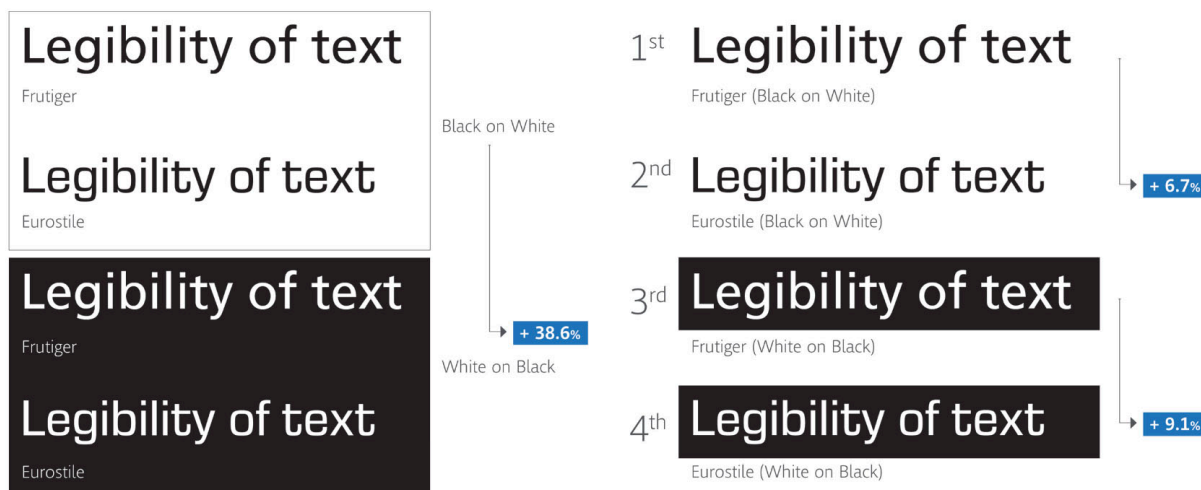


Figure 7: Illustration of relative legibility differences between typefaces and polarity conditions.

A number of participants experienced “miscalibration”, in which a series of early incorrect responses slowed the movement of the staircase and resulted in a failure to reach the participant’s true threshold before the end of the condition. This suggests that the staircase procedures could be optimized further so that more trials are presented near threshold values. For example, the first few trials of the block could gradually descend from an easy difficulty to a more challenging one, regardless of the participant’s responses. Once this initial slope is completed, staircase control can be activated with stimulus values that are closer to the observer’s likely threshold, and with a smaller step size that is less likely to cause the staircase to depart from the true threshold. Alternately, staircase values could be controlled with a QUEST procedure (Watson & Pelli, 1983), which selects stimulus values based on an assumed underlying statistical distribution, rather than a simpler adaptive rule set. Some pilot testing would be required to assess if these modifications can reduce the overall number of trials per calibration.

Comparisons to Previous Studies

In our previous typeface research (Reimer et al., in press) we employed simulated in-vehicle point-of-interest menu selection tasks in a simulated driving environment. Like the present study, those experiments found an advantage for a humanist typeface compared to square grotesque: participants spent significantly less time glancing at a menu set in humanist type. This effect was most apparent in male participants, who spent 10.6% more time looking at menus in the square grotesque typeface. Interestingly, female participants were found to spend less time glancing to the device than males. There was no statistically significant difference in glance duration by typeface for women in the original driving simulation experiments, which used the system's default "normal" brightness. A follow-up study that used the system's "bright" setting demonstrated a significant effect of typeface for both men and women, though the magnitude of the effect was more pronounced for men (9.1% vs. 3.3%). In contrast, relatively equivalent typeface effects were observed for both genders in the present study. The latter finding is more in line with general expectations based on legibility considerations alone.

While more work is clearly needed to assess generalizability, these results argue that stimulus presentation time is a valid surrogate for glance time. The effects of typeface observed in this study appear irrespective of polarity, suggesting that the earlier efforts to assess polarity in the driving simulator may have been impacted by factors extraneous to the assessment of legibility, such as illumination of the driving scene by the simulator's projector, moderate ambient room lighting used as an aid to reduce simulation sickness, etc. Environmental characteristics such as relative brightness and contrast considerations of daytime vs. nighttime lighting conditions are known to impact text presentation in different polarities on in-vehicle displays.

When comparing these studies, it is important to remember that our previous work studied legibility as part of a simulated real-world task (menu list option selection). As such, visual behaviors were influenced not only by the legibility of the typeface, but by a wide variety of psychomotor and situational factors that are beyond strict experimental or analytical control. In other words, the *context* of the previous study (the driving environment) may have limited the generalizability of its findings. The present study, in contrast, reduces the experimental task to the simplest possible yes/no decision, one mediated almost entirely by pure perceptual processes. The present findings are somewhat more abstract, but also more generalizable as a measure of pure legibility, as they remove any contextual aspect of an activity. The brief and variable stimulus durations enforce glance-like reading behavior, essentially forcing participants to read in brief glances, even in the absence of a larger, more complex task to manage (such as

driving). The current results are free of a specific context, and could therefore be a valid proxy for many types of glance-based reading, whether the glance in question is to the screen of an in-vehicle device, medical apparatus, the high-resolution display of an updating smartphone, or a 15-second commercial composed of millisecond-length shots.

The choice of task (abstracted decision-making versus fully simulated driving environment) also may explain the difference in gender effects found between the two studies. Women have been shown to more accurately evaluate the risks of certain driving situations, as well as their own driving abilities, as compared to men (Byrnes, Miller, & Schafer, 1999; DeJoy, 1992; Evans & Wasielewski, 1983). Women may simply adopt a different strategy for balancing attention on the roadway with attention to the device, which may cause the putative benefits of a typeface to “wash out” in the final metrics. Conversely, the present study reduces the test of legibility to its most fundamental components: the ability to accurately read a briefly presented word. With most extraneous behavioral factors removed, the benefits of the humanist typeface are now evident for both genders and in both polarity conditions. This is consistent with other work showing that positive polarity displays are more legible than negative polarity displays, as well as less likely to induce subtle visual aberrations due to the expansion of the pupil over the eyeball, and more likely to be preferred by those using them (Buchner & Baumgartner, 2007; Chan & Lee, 2005; Mayr & Buchner, 2010; Piepenbrock, Mayr, & Buchner, 2013; Taptagaporn & Saito, 1990).

Implications and Future Work

In summary, the methodology outlined in this paper can be used to investigate subtle aesthetic properties of typographic and graphic design by employing a relatively pure measurement of legibility. The methodology eliminates a number of confounding variables that are present when studying legibility using more typical glance time measures or in a specific interaction format such as menu selection. It is worth emphasizing that the threshold presentation time used as the primary dependent measure in this methodology bears a direct relationship to glance time requirements; the fact that the presentation time required to obtain a stable level of performance on the SOA task is shorter for the humanist than for the square grotesque typeface means that participants have demonstrated that they can carry out the same effective level of cognitive processing of information in less time when it is presented in the humanist typeface. An important advantage of the SOA methodology is that the reduced complexity, administration time, and data reduction and analysis costs make it practical to study many more subtle variations in how typographic information is displayed than would be practical to test under fully simulated or actual driving conditions.

Future work will need to assess the degree to which other aspects of the graphical user interface relate to the legibility of text rendered in different typefaces and across different polarities. Overall, the optimization of intrinsic and extrinsic features of type and the graphic design in which the text is presented may help reduce the demand of interface activities, making it easier for automotive manufacturers and suppliers to meet new governmental guidelines (National Highway Traffic Safety Administration, 2013).

Investment in further use of these psychophysical methods for the assessment of other attributes of typeface can be expected to provide a robust way to evaluate the relative tradeoffs between various intrinsic and extrinsic factors and help designers and engineers better balance the tradeoffs between “art” and “regulatory compliance”. Typography has always been, and likely will always remain, a nuanced and complex art. What makes a “best” typeface will always depend on the unique cultural, situational and visual considerations that a design is attempting to address. Bearing this mind, scientific methods like the one outlined here can be used as a valuable tool to guide designers as they explore or validate the demands of typography in other languages, environmental conditions, and even more complex visual scenarios.

Acknowledgements

This collaborative project was underwritten in part by Monotype Imaging Inc. through funding provided to MIT and in contribution of staff time. The authors would also like to acknowledge the US Department of Transportation’s Region I New England University Transportation Center at MIT for additional support. We would also like to acknowledge the contributions of Brahmi Pugh, Stephen Arredondo, Andrew McDonough, Justin Luccio, Peter Hamscher, and Jason Toby, who assisted in conducting study sessions.

References

- Beier, S., & Larson, K. (2010). Design Improvements for Frequently Misrecognized Letters 1. *Information Design Journal*, *18*(2), 118–137. doi:10.1075/idj.18.2.03bei
- Bessemans, A. (2012, October 10). *Letterontwerp voor kinderen met een visuele functiebeperking*. Leiden University.
- Blough, D. S. (1958). A Method for Obtaining Psychophysical Thresholds from the Pigeon. *Journal of the Experimental Analysis of Behavior*, *1*(1), 31–43. doi:10.1901/jeab.1958.1-31
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. doi:10.1163/156856897X00357
- Buchner, A., & Baumgartner, N. (2007). Text - background polarity affects performance irrespective of ambient illumination and colour contrast. *Ergonomics*, *50*(7), 1036–1063. doi:10.1080/00140130701306413

- Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*(3), 367–383.
- Chan, A., & Lee, P. (2005). Effect of display factors on Chinese reading times, comprehension scores and preferences. *Behaviour & Information Technology*, *24*(2), 81–91. doi:10.1080/0144929042000267073
- Chaparro, B. S., Shaikh, A. D., Chaparro, 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
- Chen, Chien-Hsiung, & Chien, Y.-H. (2005). Reading Chinese text on a small screen with RSVP. *Displays*, *26*(3), 103–108.
- Coltheart, M. (1980). Iconic memory and visible persistence. *Perception & Psychophysics*, *27*(3), 183–228.
- DeJoy, D. M. (1992). An examination of gender differences in traffic accident risk perception. *Accident Analysis & Prevention*, *24*(3), 237–246.
- Dobres, J., & Watanabe, T. (2012). Response feedback triggers long-term consolidation of perceptual learning independently of performance gains. *Journal of Vision*, *12*(8), 1–9. doi:10.1167/12.8.9
- Evans, L., & Wasielewski, P. (1983). Risky driving related to driver and vehicle characteristics. *Accident Analysis & Prevention*, *15*(2), 121–136.
- Faubert, J. (2002). Visual perception and aging. *Canadian Journal of Experimental Psychology*, *56*(3), 164–176.
- Fox, D., Chaparro, B. S., & Merkle, E. (2007, June 14). Examining Legibility of the Letter “e” and Number “0” Using Classification Tree Analysis. *Software Usability Research Laboratory, Wichita State University*. Retrieved March 17, 2014, from <http://usabilitynews.org/examining-legibility-of-the-letter-e-and-number-0-using-classification-tree-analysis/>
- Gauthier, I., Wong, A. C.-N., Hayward, W. G., & Cheung, O. S. (2006). Font tuning associated with expertise in letter perception. *Perception*, *35*(4), 541–559. doi:10.1068/p5313
- Gould, J. D., Alfaro, L., Barnes, V., Finn, R., Grischkowsky, N., & Minuto, A. (1987a). Reading is slower from CRT displays than from paper: attempts to isolate a single-variable explanation. *Human Factors: the Journal of the Human Factors and Ergonomics Society*, *29*(3), 269–299.
- Gould, J. D., Alfaro, L., Finn, R., Haupt, B., & Minuto, A. (1987b). Reading from CRT displays can be as fast as reading from paper. *Human Factors: the Journal of the Human Factors and Ergonomics Society*, *29*(5), 497–517.
- Govenlock, S. W., Taylor, C. P., Sekuler, A. B., & Bennett, P. J. (2009). The effect of aging on the orientational selectivity of the human visual system. *Vision Research*, *49*(1), 164–172. doi:10.1016/j.visres.2008.10.004
- Grainger, J., & Segui, J. (1990). Neighborhood frequency effects in visual word recognition: a comparison of lexical decision and masked identification latencies. *Perception & Psychophysics*, *47*(2), 191–198.

- Habak, C., & Faubert, J. (2000). Larger effect of aging on the perception of higher-order stimuli. *Vision Research*, *40*(8), 943–950.
- He, Y., Legge, G. E., & Yu, D. (2013). Sensory and cognitive influences on the training-related improvement of reading speed in peripheral vision. *Journal of Vision*, *13*(7). doi:10.1167/13.7.14
- Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, *50*, 243–271.
- Huang, D.-L., Patrick Rau, P.-L., & Liu, Y. (2009). Effects of font size, display resolution and task type on reading Chinese fonts from mobile devices. *International Journal of Industrial Ergonomics*, *39*(1), 81–89.
- International Standards Organization. (2009). *Ergonomic aspects of transport information and control systems* (No. 15008). Geneva, Switzerland.
- Jameson, K. A., Highnote, S. M., & Wasserman, L. M. (2001). Richer color experience in observers with multiple photopigment opsin genes. *Psychonomic Bulletin & Review*, *8*(2), 244–261.
- Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, *63*(8), 1279–1292.
- 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., Mansfield, J. S., & Chung, S. T. (2001). Psychophysics of reading. XX. Linking letter recognition to reading speed in central and peripheral vision. *Vision Research*, *41*(6), 725–743.
- Levitt, H. (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477. doi:10.1121/1.1912375
- Mayr, S., & Buchner, A. (2010). After-effects of TFT-LCD display polarity and display colour on the detection of low-contrast objects. *Ergonomics*, *53*(7), 914–925. doi:10.1080/00140139.2010.484508
- Medler, D. A., & Binder, J. R. (Eds.). (2005). *MCWord*. Retrieved December 13, 2013, from <http://www.neuro.mcw.edu/mcword/>
- Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. doi:10.1016/S1364-6613(03)00028-7
- National Highway Traffic Safety Administration. (2013). *Visual-Manual NHTSA Driver Distraction Guidelines For In-Vehicle Electronic Devices* (No. NHTSA-2010-0053) (pp. 1–281).
- Paterson, K. B., McGowan, V. A., & Jordan, T. R. (2013). Filtered text reveals adult age differences in reading: evidence from eye movements. *Psychology and Aging*, *28*(2), 352–364. doi:10.1037/a0030350
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.
- Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition.

- Nature Neuroscience*, 11(10), 1129–1135. doi:10.1038/nn.2187
- Pelli, D. G., Tillman, K. A., Freeman, J., Su, M., Berger, T. D., & Majaj, N. J. (2007). Crowding and eccentricity determine reading rate. *Journal of Vision*, 7(2). doi:10.1167/7.2.20
- Piepenbrock, C., Mayr, S., & Buchner, A. (2013). Positive Display Polarity Is Particularly Advantageous for Small Character Sizes: Implications for Display Design. *Human Factors: the Journal of the Human Factors and Ergonomics Society*. doi:10.1177/0018720813515509
- 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. (in press). Assessing the Impact of Typeface Design in a Text Rich Automotive User Interface. *Ergonomics*.
- Sanford, E. C. (1888). The Relative Legibility of the Small Letters. *The American Journal of Psychology*, 1(3), 402–435.
- Sanocki, T. (1987). Visual knowledge underlying letter perception: font-specific, schematic tuning. *Journal of Experimental Psychology. Human Perception and Performance*, 13(2), 267–278.
- Sanocki, T. (1992). Effects of font-and letter-specific experience on the perceptual processing of letters. *American Journal of Psychology*, 105(3), 435–458.
- Sanocki, T., & Dyson, M. C. (2011). Letter processing and font information during reading: Beyond distinctiveness, where vision meets design. *Attention, Perception & Psychophysics*, 74(1), 132–145. doi:10.3758/s13414-011-0220-9
- Snowden, R., & Kavanagh, E. (2006). Motion perception in the ageing visual system: Minimum motion, motion coherence, and speed discrimination thresholds. *Perception*, 35(1), 9–24.
- Taptagaporn, S., & Saito, S. (1990). How Display Polarity and Lighting Conditions Affect the Pupil Size of VDT Operators. *Ergonomics*, 33(2), 201–208. doi:10.1080/00140139008927110
- Vinot, J. L., & Athenes, S. (2012). Legible, are you sure?: an experimentation-based typographical design in safety-critical context (pp. 2287–2296). Presented at the Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems, ACM.
- Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A Diffusion Model Account of Criterion Shifts in the Lexical Decision Task. *Journal of Memory and Language*, 58(1), 140–159. doi:10.1016/j.jml.2007.04.006
- Walker, P. (2008). Font tuning: A review and new experimental evidence. *Visual Cognition*, 16(8), 1022–1058.

- Wang, A.-H., & Chen, C.-H. (2003). Effects of screen type, Chinese typography, text/background color combination, speed, and jump length for VDT leading display on users' reading performance. *International Journal of Industrial Ergonomics*, 31(4), 249–261.
- Wang, L., Sato, H., Rau, P.-L. P., Fujimura, K., Gao, Q., & Asano, Y. (2008). Chinese text spacing on mobile phones for senior citizens. *Educational Gerontology*, 35(1), 77–90.
- Watson, A. B., & Pelli, D. G. (1983). QUEST: a Bayesian adaptive psychometric method. *Perception & Psychophysics*, 33(2), 113–120.
- Yan, G., Tian, H., Bai, X., & Rayner, K. (2006). The effect of word and character frequency on the eye movements of Chinese readers. *British Journal of Psychology (London, England : 1953)*, 97(Pt 2), 259–268. doi:10.1348/000712605X70066

Jonathan Dobres, Ph.D.

Jonathan Dobres is a Postdoctoral Associate at the Massachusetts Institute of Technology AgeLab. Dr. Dobres's research interests include human-computer interaction, user experience design, visual attention, and visual learning. He received a BA, MA, and PhD in Psychology from Boston University. His research examined how visual perception changes over time with training. Dr. Dobres's current research primarily concerns the visual and cognitive demands of performing tasks while driving, as well as how the visual properties of in-vehicle interfaces affect usability and driver performance. He has conducted several studies on the cognitive demands faced by drivers when using voice-based in-vehicle interfaces, as well as a series of psychophysical investigations of the legibility of typefaces displayed on-screen.

jdobres@mit.edu
[\(617\) 253 – 7728](tel:(617)253-7728)
<http://jdobr.es>

Nadine Chahine, Ph.D.

Nadine Chahine is an award winning Lebanese type designer working as the Arabic and Legibility Specialist at Monotype. She studied Graphic Design at the American University of Beirut, has an MA in Typeface Design from the University of Reading, UK, and a PhD from Leiden University, The Netherlands. In 2005 she joined Linotype, Germany, as the Arabic specialist and has been living in Germany since then.

Nadine's research focus is on legibility studies, and in particular, the Arabic script. She has won an Award for Excellence in Type Design from the Type Directors Club in New York in 2008 and 2011. Her typefaces include: the best-selling Frutiger Arabic, Neue Helvetica Arabic, Univers Next Arabic, Palatino and Palatino Sans Arabic, Koufiya, and BigVesta Arabic.

Nadine's work has been featured in the 5th edition of Megg's History of Graphic Design and in 2012 she was selected by Fast Company as one of its 100 Most Creative People in Business.

nadine.chahine@monotype.com
www.arabictype.com

Bryan Reimer, Ph.D.

Bryan Reimer is a Research Engineer in the Massachusetts Institute of Technology AgeLab and the Associate Director of the New England University Transportation

Center. His research seeks to develop new models and methodologies to measure and understand human behavior in dynamic environments utilizing physiological signals, visual behavior monitoring, and overall performance measures. Dr. Reimer leads a multidisciplinary team of researchers and students focused on understanding how drivers respond to the increasing complexity of the operating environment and on finding solutions to the next generation of human factors challenges associated with distracted driving, automation and other in-vehicle technologies. He directs work focused on how drivers across the lifespan are affected by in-vehicle interfaces, safety systems, portable technologies, different types and levels of cognitive load. Dr. Reimer is a graduate of the University of Rhode Island with a Ph.D. in Industrial and Manufacturing Engineering.

reimer@mit.edu

[\(617\) 452-2177](tel:6174522177)

<http://web.mit.edu/reimer/www/>

Bruce Mehler, M.A.

Bruce Mehler is a Research Scientist in the Massachusetts Institute of Technology AgeLab and the New England University Transportation Center, and is the former Director of Applications & Development at NeuroDyne Medical Corporation. He has an extensive background in the development and application of non-invasive physiological monitoring technologies and research interests in workload assessment, individual differences in response to cognitive demand and stress in applied environments, and in how individuals adapt to new technologies. Mr. Mehler is an author of numerous peer reviewed journal and conference papers in the biobehavioral and transportation literature. He continues to maintain an interest in health status and behavior from his early work in behavioral medicine. He received an MA in Psychology from Boston University and a BS degree from the University of Washington.

bmehler@mit.edu

[\(617\) 253-3534](tel:6172533534)

<http://agelab.mit.edu/bruce-mehler>

About Monotype

Monotype is a leading global provider of typefaces, technology and expertise that enable the best user experience and ensure brand integrity. Based in Woburn, Mass., Monotype provides customers worldwide with typeface solutions for a broad range of creative applications and consumer devices. The company's libraries and e-commerce sites are home to many of the most widely used typefaces – including the Helvetica®,

Frutiger and Univers® families – as well as the next generation of type designs. Further information is available at www.monotype.com.

About the New England University Transportation Center & MIT Center for Transportation & Logistics

The New England University Transportation Center is a research, education and technology transfer program sponsored by the US Department of Transportation. Together the faculty, researchers and students sponsored by the New England Center conduct work in partnership with industry, state & local governments, foundations and other stakeholders to address the future transportation challenges of aging, new technologies and environmental change on the nation's transportation system. For more information about the New England University Transportation Center, visit utc.mit.edu. For more information about the US Department of Transportation's University Transportation Centers Program, please visit www.rita.dot.gov/utc/. The New England Center is based within MIT's Center for Transportation & Logistics, a world leader in supply chain management education and research. CTL has made significant contributions to transportation and supply chain logistics and helped numerous companies gain competitive advantage from its cutting edge research. For more information on CTL, visit ctl.mit.edu.

About the AgeLab

The Massachusetts Institute of Technology AgeLab conducts research in human behavior and technology to develop new ideas to improve the quality of life of older people. Based within MIT's Engineering Systems Division and Center for Transportation & Logistics, the AgeLab has assembled a multidisciplinary team of researchers, as well as government and industry partners, to develop innovations that will invent how we will live, work and play tomorrow. For more information about AgeLab, visit agelab.mit.edu.

Appendix A: Word & Pseudoword Lists

List of Word Stimuli

accent	ceased	easier	hidden	moment	polite	season	tended
acting	chairs	either	higher	months	poured	sector	thanks
adding	chance	enable	horses	mostly	powder	seemed	things
advice	change	ensure	hunger	mother	prayer	senior	thinks
agreed	charge	entire	images	motion	priest	series	thirty
allied	choose	extend	inches	moving	prince	served	though
allies	cities	extent	inside	narrow	prison	settle	threat
always	clever	farmer	island	nation	proper	severe	throat
argued	closed	father	issued	native	proved	sheets	thrown
around	closer	feared	issues	nature	purely	sighed	thrust
babies	clouds	fellow	jacket	nearby	raised	signed	ticket
barely	copies	fitted	joined	nearly	rarely	silver	toilet
beaten	corner	flower	kindly	neatly	reader	simply	trying
beauty	county	flying	kissed	needed	really	single	turned
became	course	folded	ladies	nodded	reason	sister	unable
become	courts	follow	larger	normal	record	smiled	united
behave	crying	forced	latest	notice	reduce	sought	valley
beings	damned	forces	lawyer	notion	reform	spoken	values
belief	decade	forest	learnt	number	refuse	square	varied
beside	decent	forget	lesson	opened	regard	stable	voices
bigger	decide	formal	levels	orange	region	stairs	wealth
border	deeper	former	lifted	others	relief	starts	weekly
bother	defeat	fought	likely	palace	remain	status	wholly
bottle	defend	future	linked	parent	remote	strain	widely
bought	demand	gained	listen	partly	remove	stream	window
branch	denied	garden	little	paused	repeat	stress	winter
bright	depend	gather	loaded	period	report	string	wished
brings	detail	gentle	lonely	person	resist	stroke	wishes
broken	device	gently	lovely	pieces	resort	strong	wonder
buying	dinner	golden	mainly	placed	reveal	subtle	wooden
cancer	direct	ground	manage	places	rising	sudden	worked
cannot	dishes	groups	manner	planet	rolled	suffer	worker
career	double	handed	marble	plants	rubber	summer	writer
carpet	drinks	handle	market	player	sacred	supper	
castle	driven	harder	mental	points	safely	supply	
caught	driver	health	misery	police	scenes	surely	
caused	during	heaven	models	policy	screen	survey	
causes	earned	helped	modest	polish	search	tackle	

List of Pseudoword Stimuli

acrope	carked	eurage	homedy	modert	quotor	sprels	witter
advind	caughs	euroed	humosh	momels	reares	stangs	wittly
adving	caurse	evener	humost	mothed	rectic	staten	wormas
affice	causen	expers	hundand	mothin	reinly	stater	yeldly
afries	causts	fabour	hunder	mothor	replow	stewer	
afriet	cenger	facket	hunned	motirs	resely	streed	
afring	cerend	faitan	incong	movels	rethes	stroat	
agends	chanty	failed	insels	mystem	reture	sublic	
agened	chapon	famner	insily	neemes	rhyths	sumost	
allone	chowth	fattes	inswer	norned	risted	surily	
alward	chrity	fellot	israge	nother	rister	swipos	
angind	cither	felpel	issuld	pacial	ritted	tabour	
appech	clidge	figher	itsely	pallow	rittle	talues	
appedy	coader	filist	itsend	parced	rivern	tencil	
argely	colled	finnot	jactor	parmal	rolike	thigma	
arount	collow	finths	karger	parmth	rubmit	thinge	
aroved	creful	fircle	kiless	paside	sacale	thirts	
arries	curost	flimal	kisted	pattle	savice	thison	
artiod	danded	fooker	kniman	pelief	sayind	thorus	
aventy	dardly	forcue	leampt	peoply	schoon	thould	
babief	darger	forrow	leason	peress	scorts	threet	
bantre	decket	fosays	lethin	perket	secome	threly	
beally	deemes	frient	likelf	perour	seeble	thrion	
becams	defust	fuhree	likels	perves	seeper	thurse	
becent	degare	futter	lisher	plever	seeple	tinalf	
befort	degarm	gargin	lithes	plints	seeply	tonger	
befugh	demaidd	genger	litted	pocial	shoond	trants	
begime	demain	gented	livies	poetto	shorge	tryins	
behale	dignal	geople	londer	poiled	shough	turiet	
behing	dinday	givelf	lonked	poings	shound	unched	
beight	docing	glowth	lonral	poisky	shroat	unione	
belped	drampy	goinge	lookes	pourse	siming	unlest	
belper	durder	growds	looket	pourth	sisted	uplike	
beturs	duries	havice	makins	posed	skilen	upware	
bourne	easped	havied	marded	prease	smiler	valack	
brisin	easted	havies	mather	preded	smiton	valest	
buddly	effore	haviet	mattle	prough	soctis	vioner	
bureer	enswer	heally	medger	prould	soctor	visser	
calker	enting	herear	miltor	pullow	speelf	wairty	
canate	escame	higger	misman	pummed	sporgy	wippen	
canner	essets	hights	modelf	puscle	sprect	witmal	