An Exploratory Psychophysical Investigation of the Relative Glance Legibility of Chinese Typefaces

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

In-vehicle user interfaces increasingly rely on screens filled with digital text to display information to the driver. As these interfaces have the potential to increase the demands placed upon the driver, it is important to design them in a way that minimizes attention to the device and thus keeps the driver focused on the road. 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 (Reimer et al., in press). Here we outline a methodology for rapidly and flexibly investigating the legibility of typefaces in glance-like contexts, and apply this method to a comparison of 5 Simplified Chinese typefaces. We find that the legibility of the typefaces, measured as the minimum presentation time needed to read character strings and respond to a yes/no lexical decision task, is sensitive to differences in the typeface’s design characteristics. The most legible typeface under study could be read 33.1% faster than the least legible typeface in this glance-induced context. Benefits and limitations of the methodology are discussed.

Background

As technological advances in mobile computing have been brought inside the vehicle, the complexity of in-vehicle interfaces has increased dramatically. Embedded in-vehicle interfaces now offer the driver access to a large variety of applications for convenience and “infotainment” functions. Many of these applications present textual information on an in-vehicle display, from live weather reports to integrated navigation systems with lists of potential routes and turn-by-turn directions. As a result, 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. While there have been some efforts to minimize or outright ban these types of activities in the vehicle (Michigan Vehicle Code, 2014), the benefits of such
systems ease navigation and may enhance safety, and as such, efficient optimization of the interfaces should be prioritized.

The flexibility that digital displays provide for the optimization of content presentation is a necessary solution to the problems that arise when a large number of functions can be accessed through a single interface, but these types of displays also come with new design challenges. One must consider that as these displays become everricher sources of information, the driver’s attention may be pulled toward the device, 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 (Sodhi, Reimer, & Llamazares, 2002).

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 (Reimer et al., in press). That study compared two seemingly 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 and real impact on driver behavior.

Given the complexities inherent in typographic design, the number of factors that a designer or interface engineer may wish to test, and the associated impact on legibility, a method for investigating the legibility of typefaces that minimizes time, cost, and risk is desirable. Here we present a psychophysical methodology that investigates the legibility of typefaces in brief glances. The methodology is flexible, resource efficient, can be deployed on a desktop computer, and is sufficiently sensitive to subtle differences in typographic design.

**Methods**

**Participants**

A total of 22 participants between the ages of 30 and 75 were recruited for this study, equally split between men and women (mean age 44.7 years, SD 9.9 years, no significant age difference between genders [p = .726, t-test]). All participants gave their written, informed consent to participate, as outlined by the Committee on the Use of Humans as Experiment Subjects (COUHES) of the Massachusetts Institute of Technology. Owing to cultural/local factors that can affect the interpretation of Chinese
script, participants were required to be native readers of Simplified Chinese from mainland China. Participants also had to be in self-reported 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. 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.

**Apparatus**

The experiment was run on a 2.5GHz Mac Mini running Mac OS X 10.9.1. Stimuli were created and displayed using Matlab running the Psychtoolbox 3. Stimuli were displayed on an Asus 24” (60.96cm) LCD monitor. The monitor had a resolution of 1920 x 1080 pixels and a refresh rate of 109.9Hz. Participants responded to stimuli using a standard keyboard. The experiment was conducted in a quiet, dimly lit room. Participants were seated approximately 27.5” from the screen. Head restraints were not used, though participants were encouraged to be mindful of their posture and to avoid leaning toward the screen.

**Stimuli**

*Words and Pseudowords*

Stimuli in this experiment were Mandarin words and pseudowords written in Simplified Chinese characters. Each stimulus was composed of a pair of Simplified Chinese characters that, when read left to right, either formed a single, commonly understood word/concept, or did not do so. Two-character words were selected from a compiled list of words and characters ordered by their frequency of occurrence in Chinese movie subtitles (Q. Cai & Brysbaert, 2010). Low frequency words were chosen, and these were also balanced for the frequency rate of the first character and the number of strokes occurring in each character.

More complex characters require finer visual acuity to identify (J.-Y. Zhang, Zhang, Xue, Liu, & Yu, 2007). As this complexity interferes with reading speed (J.-Y. Zhang, Zhang, Xue, Liu, & Yu, 2009), the characters chosen were moderately to highly complex in terms of stroke count, with a range of 9–20 strokes per character. Higher stroke counts were selected since we hypothesized that this is where the effect of typeface style would be seen most prominently.
Pseudowords were created by swapping the character order of the word stimuli. If the resulting combination made a word (as determined by comparing the flipped pair to the list of known words and in consultation with a Chinese linguist and a group of native Chinese readers), it was discarded. The remaining combinations made up the list of pseudowords, also balanced for the number of strokes per character. The presented order of words and pseudowords was randomized for each participant, and no stimuli were repeated during a session.

**Typefaces**

Each participant saw stimuli displayed in 5 different typefaces (see Figure 1): Monotype’s “MHeiGB18030C Medium”, hereafter referred to as MT Hei; Monotype’s “MYingHei 18030C Medium”, hereafter referred to as MT YingHei; Monotype’s “CYuen2PRC SemiBold”, hereafter referred to as MT CYuen; Microsoft’s “YaHei Regular”, hereafter referred to as “MS YaHei”; and Monotype’s “MSung PRC Medium”, hereafter referred to as “MT Sung”. All typefaces were of a medium or semi-bold weight, ensuring that character strokes were of similar thickness. All typefaces with the exception of MT Sung were of the modern Hei style (MT CYuen melds characteristics of Hei with a “Rounded Gothic” style), while MT Sung is drawn in the more traditional Ming style. Since Ming typefaces are widely considered to be less legible than the Hei style, a Ming style typeface was included as a way of verifying the sensitivity of the methodology (D. Cai, Chi, & You, 2001). A sixth typeface drawn in the calligraphic Kai style, Monotype’s “M Kai PRC Medium”, was used for a short set of practice trials.

All typefaces were scaled such that character heights averaged 5mm on screen. This height is commonly used for Chinese in-vehicle HMI designs. Text was displayed in pure black (RGB: 0, 0, 0) against a background of pure white (RGB: 255, 255, 255) at the center of the screen.
Figure 1. Examples of the 5 typefaces under study, as rendered in Adobe InDesign.

Task

Participants performed a 2-alternative forced-choice lexical decision task as illustrated in Figure 2. Each trial of the experiment began with the presentation of a fixation rectangle lasting 1000ms. This was followed by a mask stimulus presented for 200ms, which was in turn followed by a word or pseudoword character pair presented for a variable duration (see Section 2.5). The word/pseudoword stimulus was followed by another 200ms mask, and finally, a prompt screen that asked the participant to determine whether the character pair had been a word or pseudoword. The participant made his response by pressing one of two keys on the keyboard. The next trial began after a 2-second intertrial interval. All on-screen stimuli were centered on the screen. Each typeface was presented for 100 trials, or 500 trials altogether. Short rest periods were
inserted after every 50 trials. Prior to primary data collection, participants completed a series of practice trials with a novel typeface to ensure sufficient familiarity with the task.

Figure 2. A schematic of a single trial of the experiment task.

Adaptive Staircase Procedure

The 5 typefaces were each presented in separate blocks, and block order was counterbalanced between participants. Each block began by presenting stimuli for 800ms. After 3 trials, presentation time was reduced to 600ms, then 400ms, then 200ms. After these 12 trials, task difficulty was controlled via an adaptive staircase procedure (Leek, 2001; Levitt, 1971). Task difficulty increased (presentation time was reduced) whenever the participant made three consecutive correct responses, and difficulty was decreased (presentation time increased) after each incorrect response. Following this “3-up, 1-down” rule, task difficulty will converge on a presentation time threshold corresponding to 79.4% accuracy. A more legible typeface should converge on a lower threshold (shorter presentation time) compared to a less legible typeface.

Data Reduction and Analysis

Presentation time thresholds were calculated for each typeface by computing the median presentation time of the last 20 trials of each typeface condition. Participant responses were also saved for secondary analyses of reaction time and accuracy. Standard parametric statistical tests were used, including repeated-measures ANOVA and the Student’s t-test. All data were exported from Matlab and analyzed and visualized in R (R Core Team, 2014).
Results

Performance Accuracy

The use of an adaptive staircase procedure causes task difficulty to vary while stabilizing performance accuracy. Therefore accuracy, measured as percentage of correct responses, should not be different from the theoretical calibration point of 79.4%, and accuracy should not vary between typefaces.

Performance accuracy did not differ between typefaces ($F_{(4, 84)} = 1.13, p = .348$). Posthoc t-tests indicate that mean performance accuracy on each typeface was not significantly different from 79.4% (all $p > 0.05$). Across the entire sample, mean performance accuracy was 79.8%. This suggests that the adaptive staircase was able to converge on the expected stable threshold levels within the 100 trial limit for each condition.

Reaction Time

In this type of experiment, longer reaction times may indicate increased processing and higher uncertainty or difficulty (Ratcliff & McKoon, 2008). Therefore, less legible typefaces might produce longer reaction times. However, reaction time did not differ significantly between typefaces ($F_{(4, 84)} = 0.80, p = .530$).

In contrast, reaction times did differ significantly between correct and incorrect responses ($F_{(1, 21)} = 48.7, p < .001$), as well as between words and pseudowords ($F_{(1, 21)} = 99.8, p < .001$). This suggests that although reaction times are not sensitive to differences in legibility, they may reflect certain aspects of cognitive uncertainty.

Presentation Time Threshold

Figure 3 shows presentation time threshold values for each of the 5 typefaces under study. Thresholds differed significantly between typefaces ($F_{(4, 84)} = 2.75, p = .034$). Posthoc tests (Table 1), indicate that the effect was driven mostly by the MT YingHei typeface, which had a significantly shorter threshold compared to MT CYuen (borderline), MS YaHei, and MT Sung. MT Hei had a significantly lower threshold compared to MT Sung. Comparing the legibility of MT YingHei to the other typefaces, legibility thresholds for MT YingHei were 33.0% lower than MT Sung, 20.9% lower than MS YaHei, 19.0% lower than MT CYuen, and 11.8% lower than MT Hei.
Figure 3. Mean presentation time thresholds for each typeface. Error bars represent one within-subject standard error.

<table>
<thead>
<tr>
<th>Typeface A</th>
<th>Typeface B</th>
<th>t</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT CYuen</td>
<td>MS YaHei</td>
<td>0.20</td>
<td>21</td>
<td>0.847</td>
</tr>
<tr>
<td>MT CYuen</td>
<td>MT Hei</td>
<td>0.65</td>
<td>21</td>
<td>0.520</td>
</tr>
<tr>
<td>MT CYuen</td>
<td>MT Sung</td>
<td>1.26</td>
<td>21</td>
<td>0.220</td>
</tr>
<tr>
<td>MT CYuen</td>
<td>MT YingHei</td>
<td>2.02</td>
<td>21</td>
<td>0.056</td>
</tr>
<tr>
<td>MS YaHei</td>
<td>MT Hei</td>
<td>0.71</td>
<td>21</td>
<td>0.487</td>
</tr>
<tr>
<td>MS YaHei</td>
<td>MT Sung</td>
<td>1.06</td>
<td>21</td>
<td>0.299</td>
</tr>
<tr>
<td>MS YaHei</td>
<td>MT YingHei</td>
<td>2.23</td>
<td>21</td>
<td>0.037</td>
</tr>
<tr>
<td>MT Hei</td>
<td>MT Sung</td>
<td>2.21</td>
<td>21</td>
<td>0.038</td>
</tr>
<tr>
<td>MT Hei</td>
<td>MT YingHei</td>
<td>-1.41</td>
<td>21</td>
<td>0.173</td>
</tr>
<tr>
<td>MT Sung</td>
<td>MT YingHei</td>
<td>3.34</td>
<td>21</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Table 1. Posthoc test results for the effect of typeface on presentation time threshold. Significant or borderline significant results are shown in bold.
Discussion

Summary of Findings

The methodology described here adapts techniques widely used in vision science to a more applied setting. The study’s primary dependent measure, presentation time thresholds automatically calibrated to each participant and typeface, has sufficient sensitivity to reveal differences in legibility between seemingly similar typefaces. In the context of this experiment’s glance-like demands, the MT YingHei typeface proved the most legible, as it had the lowest mean presentation time thresholds. The Ming style MT Sung typeface had the highest thresholds, consistent with the hypothesis that its more complex strokes and large terminal endings would be less legible on screen.

Measures of performance accuracy (percentage of correct responses) were roughly equal between typefaces and, on average, almost exactly in line with the theoretical calibration point of 79.4%. This suggests that the adaptive staircase procedures can successfully converge on an accurate estimate of participant thresholds even in the relatively brief allotted time. Finally, although participant reaction times were not sensitive to differences between typefaces, they were sensitive to differences between distributions of correct and incorrect responses, as well as between words and pseudowords.

Implications for Testing

Typographic design is complex, particularly when typefaces are embedded in an automotive user interface. Designers may wish to test an enormous number of interacting factors, such as contrast, color, size, weight, spacing, and more. The methodology outlined here is flexible and cost/time effective, and can be used to investigate a wide variety of issues without the prohibitive expense and risk of a fully simulated driving environment.

The brief on-screen presentation times enforce glance-like reading behavior. In this way, the methodology is broadly similar to occlusion testing. However, it is worth bearing in mind that although the methodology allows for a greater degree of experimental control, it also makes somewhat different demands of the participant. In this paradigm, reading is the only task and each trial is presented as a discrete event, whereas occlusion-testing methods often employ dual task paradigms that force the participant to balance multiple streams of continuous input.

This type of psychophysical experiment outlined here can be used to rapidly investigate how various design considerations impact legibility, to build a foundation of empirical
observations that may inform interface design, or to investigate the properties of existing interface implementations.

Figure 4: Illustrations of the five typefaces and relative magnitude of statistically significant threshold differences as compared to MYingHei.

Limitations

This task abstracts glance-like reading behavior to its most fundamental components. While this allows for a more “pure” examination of legibility issues, it also decontextualizes legibility from a specific environment. Whether this is a strength or a weakness will depend on one’s investigate goals, so the constraints of this type of design should be considered.

The typefaces used in this study were chosen because they are commonly used in a variety of interfaces. Efforts were made to choose fonts with similar stroke weights, but it is not possible to customize the extremely large character sets to guarantee perfectly equal weights. Therefore these findings should be interpreted as an investigation of 5 specific typefaces, and not necessarily an investigation of typographic style in the general sense.

The present study was designed and run by a staff of native English speakers, while the participants were native Mandarin speakers. As a result, we encountered several
unanticipated linguistic subtleties unique to the interpretation of Chinese script. For example, a few participants initially found the “word/non-word” distinction confusing, as individual Chinese characters are always “words”. Additionally, it was difficult to be sure that some participants fully understood the requirements of the experiment task, several participants were excluded from analysis for an apparent failure to perform the task correctly and 3 potential participants were excused was their English language skills were not strong enough to provide informed consent.

Conclusion

The methodology outlined here is a promising and efficient tool for future investigations of legibility and visual design. The methodology has simple technological requirements, is easy to implement, and is not intrinsically bound to one particular investigative question. We have shown that it can reveal differences between relatively similar-looking Chinese typefaces. The more abstract nature of the task, as well as the linguistic complexities unique to the study of Asian scripts, should be kept in mind when undertaking this type of investigation.

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


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
http://jdoabr.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
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
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](http://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](http://utc.mit.edu). For more information about the US Department of Transportation’s University Transportation Centers Program, please visit [www.rita.dot.gov/utc/](http://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](http://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](http://agelab.mit.edu).