Displays 41 (2016) 42–49

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

The effects of Chinese typeface design, stroke weight, and contrast polarity on glance based legibility

Jonathan Dobres [a,*], Nadine Chahine [b], Bryan Reimer [a], David Gould [b], Nan Zhao [c]

[a] AgeLab, Massachusetts Institute of Technology and New England University Transportation Center, 77 Massachusetts Avenue, E40-209, Cambridge, MA 02139, United States
[b] Monotype Imaging, Inc., 600 Unicorn Park Drive, Woburn, MA 01801, United States
[c] Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road, Chaoyang District, Beijing 100101, China

ARTICLE INFO

Article history:
Received 16 February 2015
Received in revised form 20 November 2015
Accepted 3 December 2015
Available online 8 December 2015

Keywords:
Legibility
Typeface
Visual perception
Font weight
Contrast polarity

ABSTRACT

Modern interfaces increasingly rely on screens filled with digital text to display information to users. 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., 2014). Here we outline a methodology for rapidly and flexibly investigating the legibility of typefaces on digital screens 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 accurately 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 ("MT YingHei") could be read 33.1% faster than the least legible typeface in this glance-induced context. A second study examined two different weights of the MT YingHei type family (medium and bold), as well as two contrast polarity (color) conditions to investigate how these variations impact legibility thresholds. Results indicate that bold weight text is easier to read in this enforced glance-like context, and that positive polarity text (black on white) is easier to read compared to white on black text under the lighting conditions considered. These results are discussed in terms of contextual factors that may mediate glance-reading behavior, as well as how type design interacts with the practical limitations of a moderate density pixel grid.

1. Introduction

The meteoric rise of the smartphone, and the incorporation of electronic displays into an increasing number of technologies (in-vehicle devices, wearables, subway signage, advertising, etc.) have made digital screens essential to daily life. Use of these information sources has resulted in a new kind of reading behavior, markedly different from traditional long-form or "embedded" reading. Instead, more and more text is now read in short bursts of attention or otherwise performed in brief glances. Where once text might have been static and predictable, as in the dependable columns of a newspaper, a sign papered on the wall, or the display of text in a vehicle instrument cluster, it is now dynamic, shifting and changing to suit the next article, function, or app. While high-resolution displays (300-400PPI) have become increasingly common in smartphone hardware specifications, most desktop displays and in-vehicle device screens continue to rely on lower pixel densities (80-110PPI) [1,2]. Combined with the diverse environments in which displays may be used, it becomes clear that the display's readability may vary considerably, dependent on a large number of interacting factors.

As the interfaces that display information to us become more complex and the characteristics of what we can display become more flexible, it becomes necessary to form an empirical understanding of what makes text easier to encode, understand, and retain. Clear information presentation, particularly in the context of at-a-glance reading behavior, will be essential to global technology development, deployment, and marketing.

Previous research has shown that even something as subtle as a display's typeface can significantly impact reading behavior and task completion time [3]. 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

* Corresponding author.
E-mail address: jdobres@mit.edu (J. Dobres).

http://dx.doi.org/10.1016/j.displa.2015.12.001
0141-9382/

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.

The results of this study led to the development of a simplified psychophysical technique—a lexical decision task used in combination with a thresholding procedure—for assessing typeface legibility, the results of which were consistent with the results obtained from the original driving simulator study [4]. This simplified method allows for multiple interacting typographic factors to be examined and compared simultaneously in a more controlled framework. The method greatly increases the speed and flexibility of investigation while providing results that extend beyond the automotive context. It was hypothesized that the lexical decision task used would generalize to studies of non-Latin character sets, such as Simplified Chinese. Given the rapid adoption of smartphones, in-vehicle technology, and other types of digital screens by Chinese consumers [5], a psychophysical investigation of Chinese legibility factors was developed to assess this hypothesis.

Chinese typographic styles can be classified into several broad categories, including Kai, Ming, and Hei [6]. Kai typefaces resemble handwritten calligraphy. Ming (also called Song) typefaces were designed for printing, and therefore use simpler stroke patterns, but still contain subtle detailing and fine stroke widths similar to Western serif typefaces such as Times New Roman. Hei typefaces feature thicker, more regular strokes and a minimalistic design aesthetic, placing them on par with Western sans-serif typefaces such as Helvetica.

Several studies have attempted to examine the effect of typographic style on the legibility of Chinese characters. Early work in this area showed that character identification accuracy suffered when a handwriting-like Kai typeface was used to render characters, as opposed to one made for digital screens [7]. Similar work has shown the Kai style to be inferior to the Ming style on digital screens [6]. On the other hand, subsequent research using a reading comprehension paradigm failed to find an effect of typeface style [8]. Other studies of Chinese typography that use paradigms in which text is read with little or no time pressure, such as tests of reading comprehension or character search tasks, have also failed to find an effect of typeface design [9–11]. All of the aforementioned studies typically compare two typeface styles, most commonly Ming and Kai. Shieh et al. [7] reported using a Kai typeface and a "computer" typeface (the exact typeface is not reported, but the paper's figures suggest a Hei typeface). More recent work that examined the typography of digital roadway signage in a simulated driving task found an advantage of Hei-style typefaces [12]. At present, it appears that the relatively subtle differences within Chinese type styles are under-studied, perhaps because the methodologies typically used to investigate legibility lack sufficient sensitivity to reveal differences within styles.

Typographic style interacts with a number of other display factors, including the colors used for the background and foreground of the display. Several studies of the effects of color and/or contrast polarity have shown mixed results, with several indicating legibility benefits for positive polarity (dark on light) displays [8,11], another showing advantages for negative polarity (light on dark) displays [10], and at least one study that failed to show an effect of display color [7]. Recent research suggests that positive polarity displays provide a legibility advantage over negative polarity displays, and that this is most likely due to pupillary dilation in the presence of darker backgrounds, which provide less illumination than a light background [13–15]. The balance of evidence seems to suggest a legibility advantage for positive polarity digital text, consistent with early research on the legibility of printed materials [16].

In addition to text polarity, the weight, or line thickness, of the typeface can affect legibility. Studies of Latin text show that legibility is usually optimal when using typical medium weight fonts, and that legibility can be hindered if the font weight is extremely light or bold [17,18]. However, it should be noted that Sheedy et al's [19] research suggests that bold weight fonts may produce a legibility advantage in difficult "near threshold" reading conditions. To our knowledge, no comparable studies of the effect of stroke weight have been carried out on Chinese typefaces, at least as can be determined from English language literature reviews.

In summary, there has been relatively little research to date on how certain design factors, such as typeface style, font weight, and contrast polarity, affect the legibility of Chinese characters on digital screens. While several studies have compared Ming and Kai style typefaces, there has been little work done with Hei typefaces, and critically, no work to date has examined possible design differences within a single typeface style (such as the many different Hei-style typefaces currently available). The methods used in these studies typically rely on reading comprehension, character search, or serial presentation tasks. Studies have shown that such self-paced methods are not sufficiently sensitive to differences between typeface styles, while techniques that place constraints on evaluation time have been able to reveal such differences [7]. These results suggest that typeface design may play a more prominent role in constrained glance-like reading contexts. At the same time, investigations of the effects of font weight and display contrast polarity on Chinese legibility are sparse.

Here we present two studies that employ a psychophysical technique for enforcing glance-like reading behavior to examine these issues in more detail. Study I examines the relative legibility of five Simplified Chinese typefaces, four of which are within the Hei-style family. Study II expands upon these findings by choosing the most legible typeface from Study I and presenting it in two different weights and in two different contrast polarities.

2. Study I

2.1. Materials and methods

2.1.1. Participants

A total of 34 participants who natively read Simplified Chinese were recruited for this study. Of these, 5 were excluded from analysis due to an apparent failure to understand the task, 6 were excluded because they failed to reach a stable threshold estimate in the allotted time (see Section 2.1.5, and [4] for a fuller explanation of this criterion), and 1 was excluded due to technical problems with the equipment. This left a total of 22 participants between the ages of 30 and 75, equally split between men and women (men: mean age = 43.9, SD = 10.3; women: mean age = 45.5, SD = 10.1). There was no significant difference in age between genders ($t(20.0) = 0.356$, $p = 0.726$). 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 and were compensated for their involvement in the study. All data were collected by trained MIT staff in university-owned facilities.

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 illness 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.

2.1.2. Apparatus

The experiment was conducted using a 2.5 GHz Mac Mini running Mac OS X 10.9.1. Stimuli were created and displayed using Matlab running the Psychtoolbox 3 [20,21]. Stimuli were displayed on an Asus 24" (60.96 cm) LCD monitor. The monitor had a resolution of 1920 × 1080 pixels and a refresh rate of 109.9 Hz. Participants responded to stimuli using a standard keyboard. The experiment was conducted in a quiet, dimly lit room. Participants were seated approximately 27.5" (0.7 m) from the screen, within a recommended range specified in international testing guidelines [22]. Head restraints were not used, as we wished to allow some freedom of motion, as would be the case when performing real-world tasks such as dual-task driving. Participants were encouraged to be mindful of their posture and to avoid leaning toward the screen.

2.1.3. Stimuli

2.1.3.1. 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 [23]. 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 [24]. As this complexity interferes with reading speed [25], the characters chosen were moderately to highly complex in terms of stroke count, with a range of 9–20 strokes per character.

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.

2.1.3.2. Typefaces. Each participant saw stimuli displayed in 5 different typefaces (see Fig. 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. A Ming style typeface was included as a way of verifying the sensitivity of the methodology, as it was expected that a Ming typeface would be less legible than the Hei typefaces, owing to the Ming style's use of fine detailing [6]. The presented order of the 5 typefaces was randomized for each participant. A sixth typeface drawn in the calligraphic Kai style, Monotype's "M Kai PRC Medium", was used for a short set of practice trials.

Fig. 1. Examples of the 5 typefaces examined in Study I, as rendered in Adobe Photoshop CS5.

All typefaces were scaled such that character heights averaged 5 mm on screen (a mean of 16 pixels, subtending approximately 21.79 arcmin from an assumed distance of 0.7 m). The 5 mm text height was chosen because it was found to be relatively common among a small sample of Chinese-language in-vehicle interfaces. 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.

2.1.4. Task

Participants performed a 2-alternative forced-choice lexical decision task as illustrated in Fig. 2. Each trial of the experiment began with the presentation of a fixation rectangle lasting 1000 ms. This was followed by a mask stimulus presented for 200 ms, which was in turn followed by a word or pseudoword character pair presented for a variable duration. The word/pseudoword stimulus was followed by another 200 ms mask, and finally, a prompt screen that asked the participant to determine whether the character pair had been a word or pseudoword. Participants were instructed to interpret the meaning of the character pairs as read from left to right. These instructions were presented in Simplified Chinese, with two example character pairings for clarity. The participant made his response by pressing one of two keys on the keyboard. The next trial began after a 2-s 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.

2.1.5. Adaptive staircase procedure

During the 5 main data collection blocks, task difficulty was controlled via an adaptive staircase procedure [26,27]. This technique changes the difficulty of the task based on a participant's

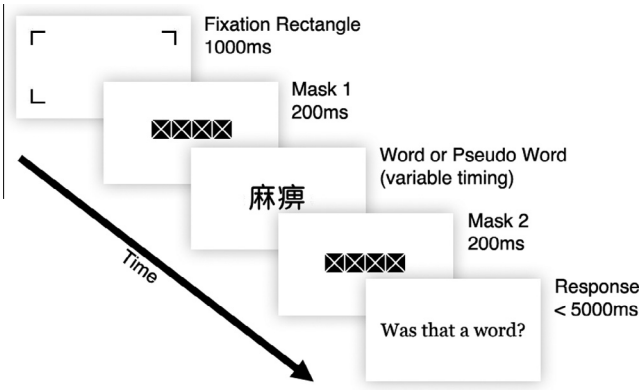

Fig. 2. A schematic of a single trial of the experiment task.

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 [27].

We modified the staircase algorithm to accommodate the experiment's workflow in the following ways. First, stimulus duration was initially decremented in a controlled manner to allow the participant to adapt to the expected task difficulty. At the start of each typeface/polarity block, stimulus duration was set at 800 ms. Three trials were performed at this setting, regardless of the participant's responses. Stimulus duration was then decremented to 600 ms for the next 3 trials, 400 ms for 3 trials after that, and finally, 200 ms for another 3 trials. Staircase control of stimulus duration was initiated on the 13th trial of the condition.

The staircase's step size (the increment by which stimulus duration was adjusted, not to be confused with stimulus duration itself) was gradually decreased throughout each condition, allowing the staircase to make finer adjustments as the condition progressed. Step size was initially set to 24 frames (200 ms), and was reduced by a factor of 20% after every 3 staircase reversals (when the staircase switched from increasing to decreasing difficulty or vice versa). Over the course of 100 trials per condition, step size reached a minimum of 1 frame. Third, stimulus duration was constrained to be at least 33.6 ms and at most 1000 ms. While the 119.9 Hz monitor used in this study was capable of a minimum presentation time of 8.3 ms, it was felt that this value made the stimulus practically invisible and constituted a nearly impossible task difficulty, particularly for older participants. A floor of 33.6 ms was implemented to reduce participant frustration and increase the number of trial responses informed by veridical perception.

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 5 conditions. 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.

In some cases, a sequence of early staircase reversals, caused either by participant confusion or erroneous responses, could cause step size to be reduced too early, and thus threshold estimates were fail to stabilize in the allotted time. As described in Section 2.1.1, these "miscalibrated" participants were excluded from analysis.

2.1.6. 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 [28].

3. Results

3.1. 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.

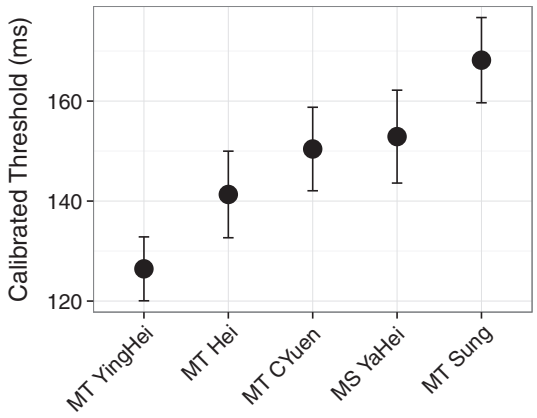

Fig. 3. Mean presentation time thresholds for each typeface in rank order. Error bars represent one within-subject standard error.

Performance accuracy did not differ between typefaces ($F_{(4,\ 84)} = 1.13$, $p = .348$). In a model that also includes age as a factor, age had no significant effect on performance accuracy ($F_{(1,\ 20)} = 1.93$, $p = .180$). 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.

3.2. Reaction time

In this type of experiment, longer reaction times may indicate increased processing and higher uncertainty or difficulty [29]. Therefore, less legible typefaces might produce longer reaction times. However, reaction time did not differ significantly between typefaces ($F_{(4,\ 84)} = 0.80$, $p = .530$). Reaction times were also unaffected by participants' ages ($F_{(1,\ 20)} = 0.07$, $p = .793$).

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.

3.3. Presentation time threshold

Fig. 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. Finally, the addition of age to this statistical model showed no main effect of age on threshold measures ($F_{(1,\ 20)} = .289$, $p = .597$), nor any significant interaction with typeface ($F_{(4,\ 80)} = 0.25$, $p = .907$).

4. Study II

4.1. Materials and methods

This experiment utilizes the same apparatus, stimuli, task configurations, and data analysis techniques as those used in Study I, except where noted below.

Table 1
Posthoc test results for the effect of typeface on presentation time threshold. Significant ($p < .05$) or borderline significant ($p < .10$) differences are indicated with (*).

Typeface A	Typeface B	*t*	*df*	*p*
MT CYuen	MS YaHei	0.20	21	0.847
MT CYuen	MT Hei	0.65	21	0.520
MT CYuen	MT Sung	1.26	21	0.220
MT CYuen	MT YingHei	2.02	21	0.056*
MS YaHei	MT Hei	0.71	21	0.487
MS YaHei	MT Sung	1.06	21	0.299
MS YaHei	MT YingHei	2.23	21	0.037*
MT Hei	MT Sung	2.21	21	0.038*
MT Hei	MT YingHei	−1.41	21	0.173
MT Sung	MT YingHei	3.34	21	0.003*

4.2. Facilities

Given the difficulty of obtaining a sample of older native readers of Simplified Chinese for Study I, primary recruitment and data collection for Study II were contracted to a local focus group facility with expertise in procuring highly specific demographic groups. Facility staff were trained by the lead MIT investigator in appropriate consent procedures, as well as how to conduct all aspects of the experiment. The rooms used at the facility were adjacent to one-way viewing rooms, which allowed the staff to monitor participants for noncompliant behavior during data collection (no participants were withdrawn for this reason).

4.3. Participants

A total of 30 participants between the ages of 31 and 60 were recruited for this study. All participants were informed of their rights as research participants and gave verbal informed consent to participate, a procedure deemed sufficient in consultation with the Committee on the Use of Humans as Experiment Subjects (COUHES) of the Massachusetts Institute of Technology. Participants were screened for eligibility using the same criteria as in Study I.

The research staff withdrew 1 participant due to an overt inability to understand task instructions. An additional 5 participants were excluded from analysis, as their data strongly suggested that they either misunderstood how to perform the task or had difficulty maintaining a clear understanding of the word/pseudoword concept as it applies to Simplified Chinese.

This resulted in a sample of 24 participants (14 men, mean age = 44.2, SD = 6.7; 10 women, mean age = 37.0, SD = 6.3). Although women were significantly younger than men in this sample ($t(20.2) = 2.68$, $p = .014$), there were no apparent gender effects in the variables of interest. Additionally, this age distribution did not differ significantly from Study I's sample ($t(39) = 1.33$, $p = .190$).

4.4. Stimuli

Word and pseudoword character pairs were drawn from the same pool as Study I, and followed the same randomization rules described in Section 2.3. Each participant saw stimuli displayed in 4 different configurations: 2 weights (medium and bold) × 2 polarities (positive and negative), as shown in Fig. 4. The conditions were all variations of the MYingHei typeface, which Study I showed to require the least amount of time for accurate on-screen reading. As in Study I, each condition was tested over the course of 100 trials.

The bold weight of MYingHei was drawn with strokes that were approximately 25.8% thicker than the medium weight used in Study I, though it should be kept in mind that stroke thickness

Fig. 4. Examples of the 4 conditions used in Study II. Rendered in Adobe Photoshop CS5.

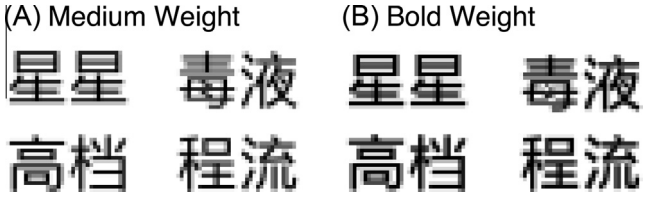

Fig. 5. Typeface samples taken directly from the Psychtoolbox frame buffer and enlarged to show rendering artifacts. (A) MYingHei Medium and (B) MYingHei Bold.

was not uniform across or even within characters. At a display size of 5 mm, characters were approximately 16 pixels in height. As shown in Fig. 5's comparison images, at this size, the practical difference in stroke weight resulted in bold strokes that appeared darker than their medium weight counterparts, but did not result in the expansion of strokes into new pixels. In other words, the characters in the medium and bold weights shared the same overall width and height, though with a different weight of the strokes. This increased their visual density, as the extra weight put on the strokes ate away from the negative space within each character.

Positive polarity text was displayed in pure black (RGB: 0, 0, 0) against a background of pure white (RGB: 255, 255, 255), while negative polarity configurations reversed these values. Each condition was exposed in a separate block. The order of the conditions was counterbalanced across participants, with the exception that polarity conditions remain contiguous (in other words, participants either saw all positive polarity trials or all negative polarity trials first, while the weights could be shown in any order).

5. Results

5.1. Performance accuracy

Consistent with Study I, performance accuracy did not differ between conditions ($F_{(3, 69)} = 0.62$, $p = .602$) and did not vary with age ($F_{(1, 22)} = 0.01$, $p = .944$). 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 79.6%.

5.2. Reaction time

As in Study I, reaction time did not differ significantly between conditions ($F_{(3, 69)} = 0.48$, $p = .698$) or vary significantly with age ($F_{(1, 22)} = 0.81$, $p = .377$). In contrast, reaction times did differ significantly between correct and incorrect responses ($F_{(1, 23)} = 30.16$,

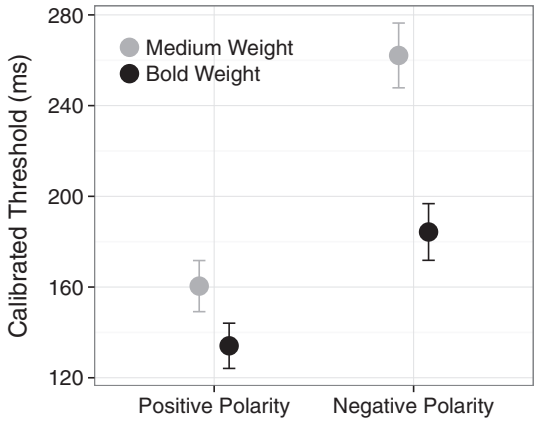

Fig. 6. Mean presentation time thresholds for each condition in Study II. Error bars represent one within-subject standard error.

$p < .001$), as well as between words and pseudowords ($F_{(1,\ 23)} = 35.85$, $p < .001$).

5.3. Presentation time threshold

Fig. 6 shows presentation time threshold values for each of the 4 conditions in Study II. Positive polarity text had significantly lower legibility thresholds compared to negative polarity text ($F_{(1,\ 22)} = 21.13$, $p < .001$). At the same time, bold weight text had lower thresholds compared to medium weight text. ($F_{(1,\ 22)} = 13.90$, $p = .001$). Weight and polarity interacted significantly ($F_{(1,\ 22)} = 4.81$, $p = .039$), indicating that the effect of weight is more pronounced in the negative polarity condition. Subsequent *t*-tests on the effect of weight within each polarity confirm this. Considering negative polarity alone, the effect of weight is significant ($t(23) = 3.66$, $p = .001$). Considering positive polarity alone, there is a weaker apparent effect of weight near the border of significance ($t(23) = 1.93$, $p = .066$). Similarly, post hoc comparisons of the two text polarities within each weight condition show a significant effect of polarity within medium weight conditions ($t(23) = 4.67$, $p = .001$) and bold weight conditions ($t(23) = 2.79$, $p = .010$). Age may have also weakly affected threshold measurements in this sample, as the effect of age is near statistical significance ($F_{(1,\ 22)} = 4.21$, $p = .052$).

The positive polarity, medium weight condition in Study II corresponds to the MYingHei condition in Study I, allowing for the two threshold estimates to be compared directly. Statistical testing shows that both studies arrive at similar estimates of thresholds for this condition (Study I = 126 ms, Study II = 160 ms; $t(44.0) = 1.35$, $p = .184$). This indicates that the calibration methodology is able to arrive at similar estimates reliably across different experiment samples.

6. General discussion

6.1. Main findings

The psychophysical technique described here adapts methods from classical vision science and applies them to real-world assessments of the legibility of Chinese typography on digital screens. In the context of this experiment's glance-like demands, the MT YingHei typeface proved the most legible, as it required the lowest mean presentation times for accurate reading. The Ming style MT Sung typeface had the highest thresholds, consistent with the hypothesis that its more complex strokes and large terminal endings would result in reduced legibility when set against the pixel grid. A deeper examination of two weights and contrast polarities of MT YingHei indicate that (1) positive polarity text requires less time for accurate reading, (2) bold weight text requires less time than medium weight, and (3) the bold weight advantage is more pronounced in the less legible negative polarity conditions.

Results indicated that the MT YingHei typeface, a Hei style typeface that employs a modern, minimalist approach to character formation, required the least amount of time to read accurately. This was followed in legibility by MT Hei. This finding was not surprising, as the two typefaces share many key features in common (see Fig. 1). The least legible typeface was the Ming style MT Sung, which features a more delicate, traditional design. Though popular in printed text, the fine detailing in MT Sung and other Ming typefaces may not be well-suited to the limitations of a digital pixel grid, especially when used in combination with a suboptimal anti-aliasing algorithm.

The hardware displays used in the present study had an effective resolution of 91.8PPI (36.1 PPCM). In combination with the Psychtoolbox's use of a grayscale font smoothing algorithm (without subpixel anti-aliasing), this makes the displays representative of mid-range displays commonly available today [2]. As shown in Fig. 5, at a practical display height of 16 pixels, the font's weight can have a marked impact on its overall appearance. While the bold weight does not make the font physically larger, it may have the effect of increasing stroke visibility and thus enhancing legibility. These results are consistent with Sheedy et al's results, which showed that bold weight fonts had a legibility advantage in near-threshold reading conditions [19]. Similarly, the present research enforces a threshold-like reading condition, albeit in the form of presentation time, rather than visual size. We argue that threshold reading conditions are of increasing relevance to practical use cases, as smartphones and other transient digital screens are likely to be read in brief or speeded glances at or near encoding thresholds, rather than at the more leisurely suprathreshold pace of traditional embedded reading.

Our results also suggest that positive polarity displays have a notable legibility advantage over negative polarity displays, at least in the laboratory lighting conditions studied here. Recent research in this area suggests that this advantage may stem from pupillary dilation effects [14,15,30]. In darker environments (as would be the case with a dark stimulus background), the pupil dilates over the imperfect surface of the eye, introducing sensory aberrations that hinder visual processing. Whether these results would be replicated in a brighter environment where the impact of screen illumination is lessened remains an open question, and presents a promising avenue for future research.

Lastly, presentation time thresholds were statistically similar between the two comparable conditions in each study, and overall performance accuracy was held almost precisely at the 79.4% calibration point. This suggests that the methodology is both replicable and reliable.

6.2. Limitations

These studies on the legibility of Chinese text were conducted by an English-speaking staff, and as a result, several difficulties in communication were encountered during data collection. Great care was taken to ensure that the pool of words and pseudowords were syntactically and culturally valid. For this reason, participants were required to have grown up in Mainland China, as the meanings of certain character pairings can change in certain localities. Despite these precautions, we found that some participants had difficulty understanding the concepts of "word" and "pseudoword", since even if a pair of characters does not form a commonly understood word, their individual characters may still be interpreted by some as concepts in the Mandarin language. In

addition, a pilot participant informed the researchers that it is often acceptable to read character pairings backwards if they do not read well forwards. While it is unclear whether this statement indicated confusion on the part of the participant, the experiment instructions were amended to specify left-to-right readings only. Even with these safeguards in place, behavioral data suggest that some participants had difficulty understanding either the task demands or the nature of the word/pseudoword distinction. As stated in the Methods section, data from such participants were excluded when either they had clear language issues (i.e. verbal communication difficulties with research assistant, understanding instructions, etc.) or difficulty reaching a stable threshold. In reality, these may not be mutually exclusive and individuals who were excluded for not reaching a stable threshold may also have been having difficulties with instructions. Care was taken to ensure that exclusions were minimized. Although all participants were required to possess English competency, the excluded cases suggest that a language barrier may have been present, though whether this affected the final results is difficult to determine. Future research may address these limitations by developing a more comprehensive methodology for assessing valid (word)/invalid (non-word) character sets, and using a Chinese-speaking experimental staff.

Lastly, this study examined the legibility of text presented at a pre-selected size of 5 mm, and did not attempt to examine a range of text sizes. While it may be expected that legibility thresholds should rise inversely with character size (smaller text resulting in higher thresholds), the effect of orthographic features unique to Chinese is unclear. For example, at small text sizes, characters can be further simplified to remove inessential strokes and improve their legibility under these conditions. The relationship between stroke reduction and text size, and their effects on glance legibility is not clear, and could represent a promising avenue of future work.

6.3. Conclusions

This study employed psychophysical techniques to quantify the legibility of digital Chinese typography as the minimum amount of time necessary to read a pair of characters with approximately 80% accuracy. Presentation time thresholds suggest that Hei style typefaces have the best on-screen legibility among the typefaces studied, perhaps due to their relatively simple, structured designs. A follow-up investigation demonstrated a strong legibility advantage for positive polarity text under laboratory lighting conditions, as well as an advantage for bold weight text. These results have important implications for the rendering of script-based characters on digital displays in a large emerging market. The psychophysical technique utilized (lexical decision) appears to be a repeatable methodology for evaluating relative legibility differences. Future work may extend the investigations presented here on typeface design, stroke weight, and contrast polarity to consider the influence of physical display characteristics, environmental influences, graphic design, language, and so forth.

Acknowledgements

This collaborative project was underwritten in part by Monotype Imaging Inc. through funding provided to MIT and in contribution of technical support and typographical expertise. 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. Earlier presentation of Study I have appeared as a lab whitepaper and a conference proceeding [31,32].

Appendix. A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.displa.2015.12.001.

References

[1] Apple Inc., Compare iPhone Models, Store.Apple.com., n.d. <http://store.apple.com/us/iphone/family/iphone/compare> (accessed 15.01.15).
[2] w3schools, Browser Display Statistics, W3schools.com., n.d. <http://www.w3schools.com/browsers/browsers_Display.asp> (accessed 15.01.15).
[3] B. Reimer, B. Mehler, J. Dobres, J.F. Coughlin, S. Matteson, D. Gould, et al., Assessing the impact of typeface design in a text-rich automotive user interface, Ergonomics 57 (2014) 1643–1658, http://dx.doi.org/10.1080/00140139.2014.940000.
[4] J. Dobres, N. Chahine, B. Reimer, B. Mehler, J.F. Coughlin, Revealing Differences in Legibility Between Typefaces Using Psychophysical Techniques: Implications for Glance Time and Cognitive Processing, Massachusetts Institute of Technology, AgeLab, 2014.
[5] N.C.A.P.-L.P.R. Dan Pan, The acceptance and adoption of smartphone use among Chinese college students, in: P.-L.P. Rau (Ed.), 5th International Conference, CCD 2013. Held as Part of HCI International 2013, Springer, Las Vegas, NV, 2013, pp. 450–458.
[6] D. Cai, C.-F. Chi, M. You, The legibility threshold of Chinese characters in three-type styles, Int. J. Ind. Ergon. 27 (2001) 9–17.
[7] K.-K. Shieh, M.-T. Chen, J.-H. Chuang, Effects of color combination and typography on identification of characters briefly presented on VDTs, Int. J. Human-Comput. Interac. 9 (1997) 169–181, http://dx.doi.org/10.1207/s15327590ijhc0902_5.
[8] A.-H. Wang, C.-H. Chen, Effects of screen type, Chinese typography, text/background color combination, speed, and jump length for VDT leading display on users' reading performance, Int. J. Ind. Ergon. 31 (2003) 249–261.
[9] C.-H. Chen, Y.-H. Chien, Reading Chinese text on a small screen with RSVP, Displays 26 (2005) 103–108.
[10] Y.-J. Yau, C.-J. Chao, S.-L. Hwang, Optimization of Chinese interface design in motion environments, Displays 29 (2008) 308–315.
[11] I. Shen, K.-K. Shieh, C.-Y. Chao, D.-S. Lee, Lighting, font style, and polarity on visual performance and visual fatigue with electronic paper displays, Displays 30 (2009) 53–58.
[12] C.-J. Lai, An ergonomic study of Chinese font and color display on variable message signs, J. Chin. Instit. Ind. Eng. 25 (2008) 306–313.
[13] A. Buchner, S. Mayr, M. Brandt, The advantage of positive text-background polarity is due to high display luminance, Ergonomics 52 (2009) 882–886, http://dx.doi.org/10.1080/00140130802641635.
[14] C. Piepenbrock, S. Mayr, A. Buchner, Positive display polarity is particularly advantageous for small character sizes: implications for display design, Human Factors: J. Human Factors Ergonom. Soc. (2013), http://dx.doi.org/10.1177/0018720813515509.
[15] C. Piepenbrock, S. Mayr, A. Buchner, Smaller pupil size and better proofreading performance with positive than with negative polarity displays, Ergonomics 57 (2014) 1670–1677, http://dx.doi.org/10.1080/00140139.2014.948496.
[16] M.A. Tinker, D.G. Paterson, Studies of typographical factors influencing speed of reading. VII. Variations in color of print and background, J. Appl. Psychol. (1931).
[17] M. Luckiesh, F.K. Moss, Boldness as a factor in type-design and typography, J. Appl. Psychol. (1940).
[18] J.-B. Bernard, G. Kumar, J. Junge, S.T.L. Chung, The effect of letter-stroke boldness on reading speed in central and peripheral vision, Vision. Res. 84 (2013) 33–42, http://dx.doi.org/10.1016/j.visres.2013.03.005.
[19] J.E. Sheedy, M.V. Subbaram, A.B. Zimmerman, J.R. Hayes, Text legibility and the letter superiority effect, Human Factors: J. Human Factors Ergonom. Soc. 47 (2005) 797–815.
[20] D.H. Brainard, The psychophysics toolbox, Spat. Vis. 10 (1997) 433–436, http://dx.doi.org/10.1163/156856897X00357.
[21] D.G. Pelli, The VideoToolbox software for visual psychophysics: transforming numbers into movies, Spat. Vis. 10 (1997) 437–442.
[22] International Standards Organization, Ergonomic Aspects of Transport Information and Control Systems, Geneva, Switzerland, 2009.
[23] Q. Cai, M. Brysbaert, SUBTLEX-CH: Chinese word and character frequencies based on film subtitles, PLoS ONE 5 (2010) e10729, http://dx.doi.org/10.1371/journal.pone.0010729.
[24] J.-Y. Zhang, T. Zhang, F. Xue, L. Liu, C. Yu, Legibility variations of Chinese characters and implications for visual acuity measurement in Chinese reading population, Invest. Ophthalmol. Vis. Sci. 48 (2007) 2383–2390, http://dx.doi.org/10.1167/iovs.06-1195.
[25] J.-Y. Zhang, T. Zhang, F. Xue, L. Liu, C. Yu, Legibility of Chinese characters in peripheral vision and the top-down influences on crowding, Vision. Res. 49 (2009) 44–53, http://dx.doi.org/10.1016/j.visres.2008.09.021.
[26] H. Levitt, Transformed up-down methods in psychoacoustics, J. Acoust. Soc. Am. 49 (1971) 467–477, http://dx.doi.org/10.1121/1.1912375.
[27] M.R. Leek, Adaptive procedures in psychophysical research, Percept. Psychophys. 63 (2001) 1279–1292.

[28] R Core Team, R: A Language and Environment for Statistical Computing, Vienna, Austria, 2015. <http://www.R-project.org/>.
[29] R. Ratcliff, G. McKoon, The diffusion decision model: theory and data for two-choice decision tasks, Neural Comput. 20 (2008) 873–922, http://dx.doi.org/10.1162/neco.2008.12-06-420.
[30] A. Buchner, N. Baumgartner, Text – background polarity affects performance irrespective of ambient illumination and colour contrast, Ergonomics 50 (2007) 1036–1063, http://dx.doi.org/10.1080/00140130701306413.
[31] J. Dobres, B. Reimer, B. Mehler, N. Chahine, D. Gould, A Pilot Study Measuring the Relative Legibility of Five Simplified Chinese Typefaces Using Psychophysical Methods, ACM, Seattle, WA, 2014, http://dx.doi.org/10.1145/2667317.2667339.
[32] J. Dobres, N. Chahine, B. Reimer, B. Mehler, An Exploratory Psychophysical Investigation of the Relative Glance Legibility of Chinese Typefaces, Massachusetts Institute of Technology, AgeLab, 2014.

Jonathan Dobres Massachusetts Institute of Technology, AgeLab Ph.D. in Psychology, 2013, Boston University.

Jonathan Dobres is a Research Scientist at the Massachusetts Institute of Technology AgeLab. His research interests include human-computer interaction, user experience design, visual attention, and visual learning. He received a BA, MA, and PhD in Psychology (Brain, Behavior, and Cognition) from Boston University. His research examined how visual perception changes over time with training. He has also worked for the Traumatic Brain Injury Model System at Spaulding Rehabilitation Hospital, part of a long-term national study on the effects of traumatic brain injuries. 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 across the lifespan.

Nadine Chahine Monotype Ph.D. in Arts, 2012, Leiden University, The Netherlands.

Nadine Chahine is an award winning Lebanese type designer working as the Arabic and Legibility Specialist at Monotype. She has an MA in Typeface Design from the University of Reading, UK, and a PhD in legibility studies from Leiden University, The Netherlands. Nadine's research focus is on eye movement and legibility studies for the Arabic, Latin, and Chinese scripts. She has won multiple design awards including 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. 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.

Bryan Reimer Massachusetts Institute of Technology, AgeLab Ph.D. in Industrial & Manufacturing Engineering, 2003, University of Rhode Island.

Bryan Reimer is a research engineer at the MIT AgeLab and associate director of the New England University Transportation Center. He received his PhD in industrial and manufacturing engineering from the University of Rhode Island in 2003. He directs work focused on how drivers across the life span are affected by new in-vehicle technologies and different types and levels of in-vehicle demand.

David Gould Monotype Director of Product Marketing M.S. in Systems Software Engineering, 1993, Boston University.

David Gould is a Market Research Director at Monotype with over 25 years' experience in the software industry. He received a Master's degree in Systems Software Engineering from Boston University and a Bachelor of Science in Biomedical Computing from the Rochester Institute of Technology. He is helping to define and drive research practices and initiatives that inform both exploratory and strategic planning for the company.

Nan Zhao Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, China Ph.D. in Cognitive Psychology, 2014, University of Chinese Academy of Sciences, China.

Nan Zhao is an Assistant Professor in the Institute of Psychology, Chinese Academy of Sciences. He has published journal and conference papers on several issues of human factors in driving, including DBQ, cell phone use in driving and drivers' change detection. Now Dr. Zhao's research also focuses on the internet behavior and human-computer interaction, especially how to understand and predict users' psychological and behavioral characteristics through Social Networking Services. Dr. Zhao is a joint trained graduate of the University of Chinese Academy of Sciences and Massachusetts Institute of Technology, with Ph.D. in Cognitive Psychology.