Applied Ergonomics 70 (2018) 240–246

Contents lists available at ScienceDirect

Applied Ergonomics

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

The effects of visual crowding, text size, and positional uncertainty on text legibility at a glance

Jonathan Dobres[a,*], Benjamin Wolfe[a], Nadine Chahine[b], Bryan Reimer[a]

[a] MIT AgeLab and New England University Transportation Center, 77 Massachusetts Avenue, E40-275, Cambridge, MA 02139, USA
[b] Monotype Imaging, Inc., 600 Unicorn Park Drive, Woburn, MA 01801, USA

ARTICLE INFO

Keywords:
Typography
Legibility
Visual crowding

ABSTRACT

Reading at a glance, once a relatively infrequent mode of reading, is becoming common. Mobile interaction paradigms increasingly dominate the way in which users obtain information about the world, which often requires reading at a glance, whether from a smartphone, wearable device, or in-vehicle interface. Recent research in these areas has shown that a number of factors can affect text legibility when words are briefly presented in isolation. Here we expand upon this work by examining how legibility is affected by more crowded presentations. Word arrays were combined with a lexical decision task, in which the size of the text elements and the inter-line spacing (leading) between individual items were manipulated to gauge their relative impacts on text legibility. In addition, a single-word presentation condition that randomized the location of presentation was compared with previous work that held position constant. Results show that larger text was more legible than smaller text. Wider leading significantly enhanced legibility as well, but contrary to expectations, wider leading did not fully counteract decrements in legibility at smaller text sizes. Single-word stimuli presented with random positioning were more difficult to read than stationary counterparts from earlier studies. Finally, crowded displays required much greater processing time compared to single-word displays. These results have implications for modern interface design, which often present interactions in the form of scrollable and/or selectable lists. The present findings are of practical interest to the wide community of graphic designers and interface engineers responsible for developing our interfaces of daily use.

1. Introduction

Over the last decade, the advent of mobile computing has fundamentally impacted the ways in which users interact with their devices and the information accessible from them. This rapid evolution has led to changes in common use cases for human-computer interaction. Interfaces that were once relatively stable and simplistic, such as the infotainment dashboards of motor vehicles, have now become digital front-ends full of dynamically changing content. Where once computers were reliably anchored to desktops, now they can be pulled from pockets or read from the wrist. As the use cases have evolved, so too have the user behaviors associated with them. Perhaps most prominently, users are now accustomed to reading pieces of text in brief glances, a behavior previously limited to the intake of information from signage or in-vehicle gauges.

Although the factors affecting legibility have been of interest to researchers for over a century, most of this body of work has focused on long-form paragraph reading or threshold acuity assessments. Both of these are unpressured perception paradigms, in which the observer reads and responds at his or her own pace. In contrast, more modern mobile interaction paradigms, and especially in-vehicle interfaces, often force users to multitask, placing constraints on the amount of time the user has available to perceive the interface and process secondary information. Recent laboratory-based research examining legibility at a glance has shown that legibility can be affected by the typeface family used and the size at which it is set (Dobres et al., 2016a, b), the boldness or weight of the font (Dobres et al., 2016b), and the amount of ambient illumination available in the interface's environment (Dobres et al., 2017a). Furthermore, these laboratory assessments are corroborated by driving simulator research showing that the typeface used for a menu interface can affect the amount of time spent glancing off the road to the device screen, the total time needed to complete secondary tasks, and the number of errors made (Reimer et al., 2014).

While the aforementioned body of work represents a first step in the investigation of glance-based legibility, the work also highlights the large number of interacting factors that affect complex reading behavior. Those studies presented their target stimuli in isolation, without distracting elements, and thus neglected the impact of visually crowded

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

https://doi.org/10.1016/j.apergo.2018.03.007
Received 17 July 2017; Received in revised form 22 December 2017; Accepted 12 March 2018

displays on legibility. Crowded displays are common in modern HCI designs, as users are often required to scroll through lists of information or differentiate between columns of a design layout, as when reading a website. Note that by "crowding", we are referring generally to displays which feature distractor elements alongside some target of interest. This is not to be confused with the strict classical definition of visual crowding, which specifically refers to a loss of visual discriminability experienced when a target presented in the visual periphery is surrounded by flanking elements ((Bouma, 1970, 1973); for extensive reviews of this phenomenon, see (Pelli and Tillman, 2008; Pelli et al., 2007; Whitney and Levi, 2011)). While these types of phenomena grant insight into the workings of visual perception and processing, they are of less direct use to practicing engineers and designers. Instead, in the present paper we rely on a looser definition of visual crowding, which refers simply to the presence and density of task-irrelevant visual information in a given display.

Even so, there is reason to believe that classical crowding affects legibility even in central (foveal) vision (Chung et al., 2007; J.-Y. Zhang et al., 2009). Over a century ago, Roethlein found that fonts which featured more negative space (i.e., were less visually cramped) were associated with greater legibility (Roethlein, 1912), a finding supported at least in part by more recent research showing that lighter-weight fonts, which have more negative space by definition, may be more legible than thicker fonts (Dobres et al., 2016b). Along the same lines, increasing the spacing between individual characters generally aids legibility, even when words are read foveally under normal conditions (Montani et al., 2014; Perea and Gomez, 2012; Perea et al., 2011).

However, such intra-character crowding effects are relatively subtle compared to effects arising from the overall density of text on the page or display. One factor governing the density of text is the leading, or vertical space that separates lines of text. Research on this issue, though sparse, has consistently shown increased leading to be markedly beneficial for legibility compared to "set solid" or maximally dense typesetting (at least up to extreme leadings greater than twice the line height), and that readers subjectively prefer text set with some amount of leading (Bentley, 1921; Paterson and Tinker, 1944, 1947; Poulton, 1972; Tinker, 1963; Wilkins and Nimmo-Smith, 1987). This line of work also highlights the fact that the effect of increased leading may be amplified in variation with other factors, such as text size, horizontal line width, or the typeface used (Becker et al., 1970; Paterson and Tinker, 1944; Tinker, 1963). These same findings are also apparent in investigations of digital typography (Holleran and Bauersfeld, 1993; van Nes, 1986), albeit in paradigms that require unpressured responses from the reader.

In the present study, we expand this line of research by examining the effect of crowded displays on glance-based legibility, more consistent with modern, mobile-oriented reading behaviors. Legibility is assessed using a lexical decision task, which requires readers to classify a briefly presented stimulus as either a word or a nonsense pseudoword. The display time of these stimuli is adjusted via a staircase procedure in accordance with participant performance, to arrive at a reading time threshold per condition studied. This allows for the legibility of different typographic configurations to be compared. Legibility is assessed at two text sizes and with two different degrees of leading. In addition, a condition that presents the lexical decision task in isolation, without crowding distractors, is also compared. We hypothesize that 1) crowded displays will require more time for accurate reading compared to isolated displays, 2) text set at a larger size will be read more easily than text at a smaller size, 3) text with more generous leading (more vertical spacing) will be read more easily than more crowded text, 4) wider leading should ameliorate decrements in legibility arising from smaller text sizes, 5) the effect of crowding will be more pronounced at the smaller text size, and 6) older readers will experience greater increases in reading time for the more difficult conditions.

2. Methods

2.1. Participants

37 participants (between the ages of 35 and 75) were recruited from the Massachusetts Institute of Technology AgeLab's participant pool. Prior to participating in the study, all participants provided written informed consent in accordance with the MIT Institutional Review Board as required by the Declaration of Helsinki. Participants were required to be in "reasonably good health" as reported to experimenters. Participants were excluded from the study if they had experienced a major medical illness or had been hospitalized in the previous six months, or if they had medical conditions that impair vision (beyond those which can be treated with corrective lenses). Participants were excluded if they reported a history or diagnosis of epilepsy, Parkinson's disease, Alzheimer's disease, dementia or mild cognitive impairment, or other neurological problems. All participants were native English speakers, and had normal or corrected-to-normal vision (e.g., with glasses or contacts worn for the experiment). In addition, all participants were assessed 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. No participants were excluded due to excessively low acuity (summary statistics are provided in Table 1). As the table indicates, one participant had unusually low distance acuity but adequate near acuity. The opposite is true for one participant with unusually low near acuity. Excluding these participants from analysis did not appreciably change the reported results. In the interest of providing a more varied and robust sample, these participants are retained. Participants were permitted to choose whether they wore corrective lenses during the experiment, and were asked to abide by their choice throughout data collection.

Of the 37 total participants, 7 were excluded, leaving a final pool of 30: 5 failed to reach a stable stimulus duration threshold, resulting in unreliable measurements, and 2 were excluded due to a failure to comply with the experiment protocol. Failure to reach a stable threshold was defined as a calculated threshold value of greater than 600 ms, or if a participant's staircase showed no reversals in the last twenty trials, indicating that the participant had failed to achieve a stable level of performance.

This left a total of 30 participants (mean age = 53.0 years). Age distribution did not differ significantly between genders (t(28) = 0.32, p = 0.754, *t*-test). Descriptive statistics for the final participant pool are provided in Table 2.

2.2. Task, apparatus, & stimuli

2.2.1. Task

Participants performed a 1-interval forced choice lexical decision task (Meyer and Schvaneveldt, 1971), modified to accommodate an array of distractor words. The stimulus sequence and timings are shown Fig. 1. The word arrays contained a single target word or pseudoword embedded within an array (three columns and five rows) of distractor words. The target word/pseudoword was always presented in the array's center column, and never appeared in the top or bottom row, ensuring that the target was always crowded on all sides. Each lexical decision trial begins with a 1000 ms screen cueing the participant to

Table 1
Summary statistics for binocular acuity measures (all measures taken with optical correction worn).

Acuity	Mean	SD	Min	**25th percentile**	**50th percentile**	**75th percentile**	Max
Near	32.17	8.68	25	25	30	30	60
Distance	25.87	11.01	13	20	25	28.75	70

Table 2
Descriptive statistics for the participant sample.

Gender	n	Mean	SD	Minimum	Maximum
Female	14	53.79	13.16	35	72
Male	16	52.25	13.39	35	73

search for the target on a particular line (line 2, 3, or 4, randomized across trials). This is followed by a fixation rectangle, displayed for 400 ms (700px by 400px, or approximately 13.19° by 7.63° at viewing distance), centered on the screen, indicating the general area where stimuli will appear. The fixation rectangle is followed by a 200 ms masking array composed of randomized punctuation characters matching the size of the word array. Subsequently, a word (or pseudoword) stimulus is displayed embedded within a 5 row x 3 column array of distractor words. Presentation time of this key stimulus array was varied, as determined by an adaptive staircase procedure. This is immediately followed by another 200 ms masking array. Finally, the participant is prompted to decide whether the stimulus was a word or pseudoword. Participants are given a maximum of 5000 ms to respond by pressing either the '1' or '3' key of the numeric keypad (the keys corresponded to "word" and "pseudoword", respectively, and were marked with either green or red tape for clarity). Participants were not provided with feedback regarding the accuracy of their responses, other than during the practice section described below. Each mask was unique, constructed by randomly selecting eight characters from a small pool of punctuation characters. The sandwiching of the stimulus between the two masks minimizes the persistence of the stimulus 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 fixed at 1000 ms. 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. 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. Prompt text set in Georgia was also displayed at approximately double the size of the word and pseudoword stimuli.

2.2.2. *Apparatus*

The experiment was conducted in a quiet, dimly lit room. A comfortable level of ambient illumination was provided during the experiment by two low-power floor lamps directed toward the room's ceiling, resulting in an ambient illumination level of approximately 23 lux near the participant's eyes. The experiment was run on a 2.4 GHz Mac Mini running Mac OS X 10.6.8. The experiment was performed using the PsychoPy library for Python (Peirce, 2008) and stimuli were displayed on an Acer 27" (68 cm) LCD monitor. The monitor had a resolution of 2560 × 1200 pixels and a refresh rate of 60 Hz. All text was rendered using PsychoPy's font rendering capabilities (via the Pygame and Pyglet software libraries), which do not support subpixel anti-aliasing and instead use grayscale font smoothing to ensure accurate presentation of letterforms.

2.2.3. *Stimuli*

The primary stimuli of this experiment were English words selected from an online orthographic database (Medler and Binder, 2005). To generate a sufficiently large list of usefully common words, word length was restricted to 6 letters; orthographic neighborhood size (the number

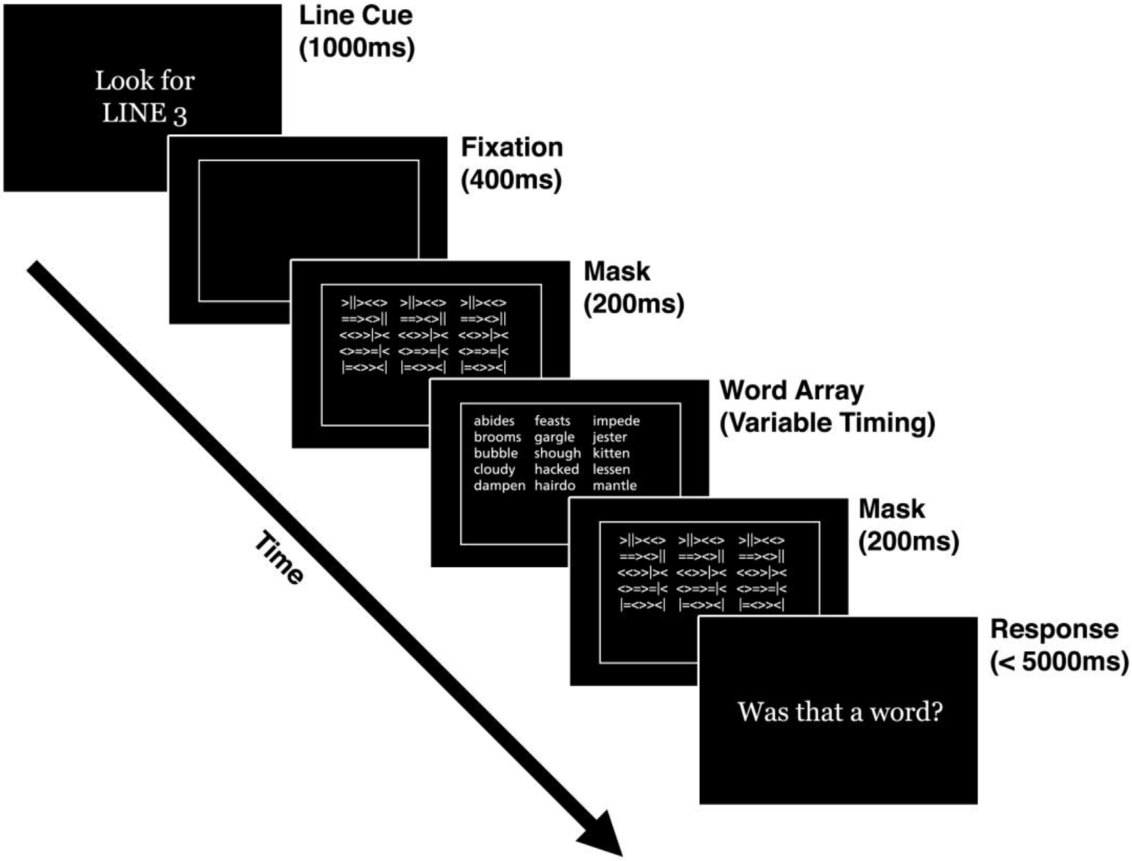

Fig. 1. Schematic of a single trial of the lexical decision task (not to scale). The position of the array within the fixation rectangle and the row on which the target appeared were randomized for every trial (targets always appeared in the middle column). In this example, the target is the pseudoword "shough", in the center of the panel.

of words of the same length that differ by exactly one letter) was restricted to between 1 and 5 (inclusive); word frequency was set to 2–5 per million (inclusive); and bigram frequency (the frequency of a specific two-letter set of characters in a specific word position) was constrained to a minimum of 600 per million. All other search parameters were unconstrained. This ensured a list of relatively common English words that were suitably varied in letter combination. 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 in English.

The pool of distractor words had no items in common with the target word pool. The lexical database was searched with the same parameters as above, except that word frequency was set to a maximum of 1 per million (no minimum), thus generating a list of words that appear more rarely in the English lexicon. This resulting word list was manually inspected by the experimenters to remove offensive or potentially upsetting terms.

2.2.4. Randomization of target position

As the primary goal of this experiment is to examine the effects of visual crowding on reading accuracy, it is essential that the crowding elements remain sufficiently distracting for the duration of the experiment. To prevent participants from simply attending to a narrow region of the screen, several steps were taken to increase randomness in stimulus presentation and prevent such a strategy from being useful to participants. As described above, arrays were presented within a marked region of the screen. The exact position of the array within the area was jittered on each trial, with horizontal and vertical origin coordinates randomly chosen from two independent uniform distributions. In addition, the target word/pseudoword could appear in one of three positions in the middle column of the array (only in the second, third, or fourth row). Immediately prior to the presentation of the array (as shown in Fig. 1, first panel), participants were cued as to which row would contain the target.

2.2.5. Conditions tested

A total of 5 experimental conditions were tested. Word arrays were set at an onscreen text size of either 4 mm or 3 mm, and rows were separated with a leading (inter-line spacing) of either 0% of the font size or 33% of the font size. With these values, a 5-line 3 mm array with 33% leading subtends the same visual angle as a 5-line 4 mm array with 0% leading. These two factors were fully crossed, producing four array conditions. Array columns were separated by margins 20 pixels wide (4.7 mm). In addition, a fifth condition was included in the experiment where the target word/pseudoword was presented alone in a random position within the fixation rectangle. This condition was included for comparison to previous studies from our lab that had presented single stimuli in a fixed location (Dobres et al., 2016a, b; 2017a, b), to gauge the effect of positional uncertainty.

All stimulus text was presented in Frutiger, a humanist style sans-serif typeface that has previously been shown to have desirable legibility characteristics, particularly in glance reading (Dobres et al., 2016a; Reimer et al., 2014). Text size was set by the height of the capital 'H' character (International Standards Organization, 2007), although all text was displayed in lowercase lettering. All stimuli were displayed as white text (RGB: 255, 255, 255), on a plain black background (RGB: 0, 0, 0).

Each condition was presented in a separate block to avoid confusion, and the order of blocks was randomized across participants to avoid order effects. Each condition contained 50 word trials and 50 pseudoword trials, randomly interleaved for a total of 100 trials per block. In addition, the order of target words/pseudowords was randomized for each participant. Primary data collection (500 trials in total) began after the practice block. Every 25 trials (approximately every 2–3 min), participants were allowed to take a short break of up to 30 s (the participant could terminate the rest periods early if they desired to do so). All participants completed all 5 conditions in a single sitting, which took approximately 1 h.

Participants performed the experiment at a viewing distance of approximately 27" (68.58 cm) from the screen (similar to the distance of a typical automotive interface), and were instructed to avoid changing the distance between themselves and the screen during the experiment (word stimuli were therefore displayed at a vertical size of approximately 20.1 arcmin). Head restraints were not used, which allowed for a degree of positional variability that is likely to be encountered in real-world reading scenarios. The 4 mm screen character height and the viewing distance were consistent with ISO standard 15008 (International Standards Organization, 2009) for automotive displays, which recommends character sizes > 20 arcmin.

2.3. Adaptive staircase procedures

During the five data collection blocks, task difficulty was manipulated using 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 (by decreasing stimulus duration) after three consecutive correct responses, and made easier (by increasing stimulus duration) after one incorrect response. Following this rule, stimulus duration will converge to a point where participants are correct in their judgment on 79.4% of trials (Leek, 2001).

To accommodate the experiment's workflow, we made the following modifications to the staircase algorithm. To begin, stimulus duration was initially decremented in a controlled manner to allow the participant to adapt to the difficulty of the task. At the start of each condition (e.g., at the beginning of each block), stimulus duration was set at 800 ms. The first three trials of each block were performed at this setting, regardless of the participant's responses. Stimulus duration was then reduced to 600 ms for the next 3 trials, 400 ms for 3 trials after that, and finally, to 200 ms for another 3 trials. Staircase control of stimulus duration took effect on the 13th trial of the condition.

Staircase step size (the increment of stimulus duration adjustment used, not to be confused with stimulus duration itself) was gradually reduced throughout each condition, allowing the staircase to make finer adjustments as the condition progressed. At the beginning of the block, step size was set to 12 frames (200 ms), and was reduced by a factor of 20% after every 3 reversals (a reversal being when the staircase switched from increasing to decreasing difficulty or vice versa). Over the course of 100 trials for each condition, step size reached a minimum of 1 screen refresh (16.67 ms). In addition, stimulus duration was constrained to be at least 33.4 ms (2 frames) and at most 1000 ms (60 frames). While the 60 Hz display used for this study was capable of a minimum presentation time of 16.7 ms, this duration was considered by the experimenters to be a nearly impossible level of difficulty, particularly for older participants. Therefore, a floor of 33.4 ms was implemented to reduce participant frustration.

The staircase procedure began anew at the start of each condition, 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 (of 79.4% accurate responses). As a result, a less legible condition should require a longer presentation time (reported as a higher threshold) to achieve the same accuracy level as a more legible typeface.

2.4. Data analysis

Thresholds were calculated for each condition by taking the median stimulus duration used on the last 20 trials of each condition. In addition, response accuracy and response times were recorded for each trial. Response time was calculated as the mean response time across trials per participant and condition, excluding each condition's first 20 trials to account for habituation (leaving 80 trials per condition). An initial omnibus model analyzed stimulus duration thresholds for the array-crowded conditions in a 2 × 2 repeated-measures design (size × leading), including between-participant factors of gender and age group (younger than 56, or 56 and over, bifurcating the sample). Age group and gender were found to have non-significant effects on reading time thresholds, and so a reduced model including only within-participant factors was used. The omnibus model was also applied to response times. Reading time thresholds for the single-word condition are compared against the crowded conditions, as well as data from two previous experiments from this lab (Dobres et al., 2016a). Measures of effect size for repeated-measures tests (eta-squared, or η^2) and two-group tests (Cohen's *d*) are provided for all significant effects (Olejnik and Algina, 2003; Bakeman, 2005). All statistics were computed and visualized using R (R Core Team., 2008).

3. Results

3.1. Response accuracy

Adaptive staircase procedures dynamically adjust the difficulty of the experimental task to achieve a target performance level. As such, we expect each participant and condition to produce a unique threshold value (in this case, corresponding to the stimulus display duration needed to achieve the target accuracy level), while overall performance accuracy should not vary between conditions. As expected, a repeated-measures ANOVA shows that performance accuracy did not differ significantly between the conditions under study (F(4, 116) = 0.13, p = 0.970), nor was accuracy in any individual condition significantly different from the theoretical calibration point of 79.4% accuracy (all $p > 0.684$, one-sample t-tests). On average across all participants and conditions, performance accuracy was 79.4%, identical to the targeted performance accuracy level. This indicates that the adaptive staircase procedures worked as intended and that stimulus display duration—the operational metric of legibility in this study—can be interpreted as the dependent measure.

3.2. Response times

An omnibus repeated-measures ANOVA specifying text size and leading as within-participant factors and gender and age group as between-participant factors indicated significant effects for gender (F(1, 29) = 7.18, p = 0.007, $\eta^2 = 0.210$) and age group (F(1, 29) = 17.29, $p < 0.001$, $\eta^2 = 0.507$). Women responded significantly more slowly than men (119 ms more slowly, on average), and older participants responded significantly more slowly than younger ones (179 ms more slowly, on average). Response times were not significantly affected by the typographic factors of leading (F(1, 29) = 0.23, p = 0.632) or text size (F(1, 29) = 0.287, p = 0.597), consistent with previous research in this area (Dobres et al., 2016a). Age and gender did not interact significantly with any of the typographic factors.

3.3. Display time thresholds

Fig. 2A illustrates threshold display times among the crowded arrays. An omnibus repeated-measures ANOVA specifying text size and leading as within-participant factors and gender and age group as between-participant factors indicated no significant effects for age group (F(1, 29) = 3.73, p = 0.054) or gender (F(1, 29) = 0.30, p = 0.582), nor did these demographic variables interact significantly with the within-participant variables. A reduced model that drops gender and age group as predictors indicates a significant main effect of type size (F (1, 29) = 24.54, $p < 0.001$, $\eta^2 = 0.550$), with larger text requiring less time for accurate reading compared to small text. A significant main effect of leading is also evident (F(1, 29) = 4.62, p = 0.040, $\eta^2 = 0.093$), with wider inter-line leading associated with faster reading times. These two factors did not interact significantly (F(1, 29) = 0.32, p = 0.574). Notably, 4 mm text with 0% leading had significantly lower display time thresholds (better performance) than 3 mm text with wider leading (t(29) = 3.22, p = 0.003, Cohen's *d* = 0.216).

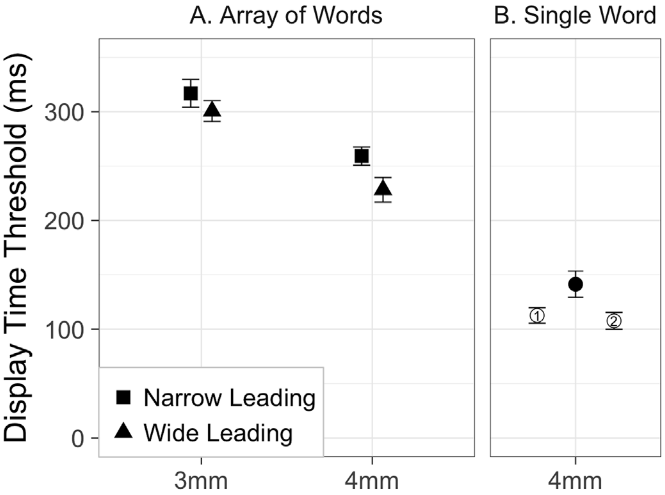

Fig. 2. Mean display time thresholds for all conditions under study. Error bars represent ± 1 within-subject standard error. A) Thresholds for crowded word arrays. B) Thresholds for single-word presentation conditions. Data from the conditions in the present study are shown in black (variable stimulus location), and data from two earlier studies from this lab (Dobres et al., 2016a) are shown in white (stable stimulus location). The points labeled “1” and “2” correspond to Dobres et al., 2016a Studies I and II, respectively.

A fifth condition (black data point in Fig. 2B) was included that presented the word/pseudoword stimuli in isolation, instead of in an array. Compared to the mean display time required for crowded displays, single-word displays required significantly less time for accurate reading (t(29) = 10.05, $p < 0.001$, Cohen's *d* = 1.06). The present experiment varied the location of the single-word stimulus across trials. Two typographic conditions from an earlier study conducted in the same lab (Dobres et al., 2016a) used identical text size, color, and font, and differed only in that stimulus location was held constant at the center of the screen (white data points in Fig. 2B). This allows for a comparison of the effect of positional variability on reading time. We note that Dobres et al., 2016a, Study I (Fig. 2B, white point labeled “1”) utilized recruitment criteria that were slightly different from subsequent studies, resulting in a sample that was significantly younger (by approximately 8 years on average) than in the later studies (all $t > 2.38$, $p < 0.021$). To account for this effect, we employed a linear model that specified study and participant age as independent variables. The analysis shows that age significantly affected reading time thresholds across studies (F(1, 104) = 11.33, p = 0.001, $\eta^2 = 0.092$) and that thresholds differed across studies (F(2, 104) = 3.56, p = 0.032, $\eta^2 = 0.065$). Age did not interact with study (F(2, 104) = 0.05, p = 0.950). Gender balancing in all three studies was similar, as confirmed by a chi-square test of independence (X2 = 0.097, p = 0.953). Posthoc testing confirms that the thresholds measured in the current study are significantly higher than those of the two previous studies (all $t > 2.04$, $p < 0.047$, Cohen's $d > 0.51$), while thresholds in the two earlier studies were not significantly different from each other (t(71) = 0.474, p = 0.637).

Fig. 3 shows display time thresholds for individual participants in all the conditions under study. A repeated-measures ANOVA that considers age and its interaction with condition as predictors indicates that the effect of age on display time threshold approaches, but does not reach, significance (F(1, 28) = 3.80, p = 0.061), with reading times trending toward an increase among older participants. While condition is highly significant in this model (consistent with the tests reported

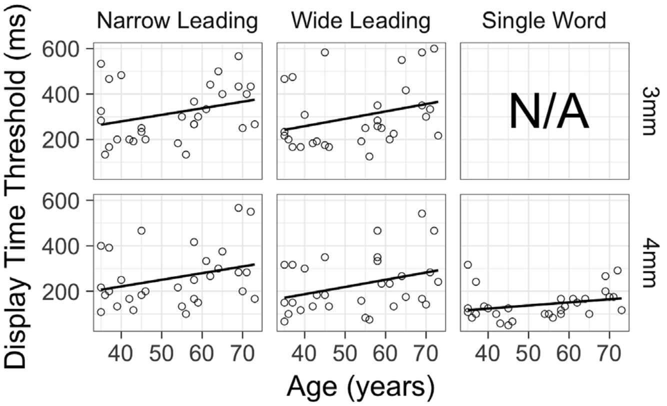

Fig. 3. Display time thresholds for each individual participant in each condition. Solid lines represent simple linear regressions through each set of points (display time predicted from age per condition).

above), age did not interact significantly with condition, suggesting that increases in reading times due to age are not significantly different across the conditions studied (F(4, 112) = 0.654, p = 0.625).

4. Discussion

In the present study, we examined two primary factors that can affect the at-a-glance legibility of modern HMI designs: the size of displayed text and the amount of inter-line spacing (leading). Consistent with our hypotheses, participants required 26.7% more time, on average, to read the smaller text displays accurately. In addition, displays that used narrower leading required 8.9% more time for accurate leading compared to displays with wider leading. Notably, these two factors did not interact significantly, and in contrast to expectations, wider leading did not fully compensate for decrements in legibility at smaller text sizes. The data also clearly show that crowded displays require substantially more time for visual intake than a stimulus presented in isolation, by an average of 95.4%. In contrast to our hypotheses, we observed that participant age was weakly correlated with legibility thresholds and did not reach statistical significance, nor did age interact significantly with leading or text size. Lastly, comparing the single-word condition of this experiment, which varied stimulus position, to earlier single-word data that used a stationary position, shows that positional variability itself contributes a significant 27.8% increase in reading times.

The present results, which show more widely leaded text to be more legible compared with closer-set text, is in accord with much of the historical literature on this topic. The present work expands that line of findings by showing that the benefits of wider leading transfer to digital displays read at a glance. At the same time, the results showing that smaller 3 mm text is considerably more difficult to read at a glance than 4 mm text, though perhaps unsurprising, agree with previous research on this topic (Dobres et al., 2016a; 2017a). It was expected, given previous work in this area (Becker et al., 1970; Paterson and Tinker, 1944; Tinker, 1963), that text size and leading would interact with each other, essentially producing a “multiplicative” effect on legibility thresholds. This was not the case in the present study. Moreover, wider leading produces relatively modest changes in legibility compared to changes in text size, suggesting that one cannot simply be assumed to offset the other.

One may further ask whether the effect of text size in the context of a crowded display is different from the effect of text size in an isolated display. The present study uses methods and a typographic configuration comparable to previous work (Dobres et al., 2016a). That study found that words presented in isolation set in 3 mm Frutiger required 26.4% more time for accurate reading compared to the 4 mm size. This is strikingly similar to the 26.7% difference between 3 mm and 4 mm sizes seen in the present data for crowded displays. Data from the earlier studies also allows for an examination of the effect of randomizing the stimulus position in the present study, which shows a separate 27.8% increase in reading times. Although real-world interfaces are unlikely to employ randomized positioning in their designs, it is plausible to assume that other external factors may “randomize” the location of interface elements in practice. For example, as users rapidly switch between two or more different applications with different layouts or interaction paradigms, momentary confusion could arise as their associated mental models come into conflict. In other contexts, such as when using an in-vehicle infotainment system while driving, the user is likely to be in a somewhat “randomized” position relative to the device, due both to the task switching cost of moving attention from roadway to device, and because of the physical movement caused by the vehicle.

The present data failed to demonstrate strong age-related effects on text legibility, other than a linear effect of age on legibility thresholds that approaches significance. This is in contrast to previous studies that have shown quite clear age effects (Dobres et al., 2016a; 2017b). It is possible that the use of variable positioning in this study led to increased variance in threshold measurements, thus weakening effects due to age. Alternatively, we may speculate that crowded displays are more difficult to read regardless of age, and may account for the shallower age slopes seen here.

There is also a notable disconnect between measures of response time and reading time threshold. Reading time thresholds revealed the effects of typographic manipulations, but were not sensitive to demographic variables such as gender and age. Response time thresholds demonstrated the opposite pattern, showing sensitivity to demographic factors, but not typographic ones. These results are sensible in light of the experiment's design. The present experiment was structured such that a stimulus would be displayed for some fixed amount of time, only after which the “response window” would become available. Thus, a participant's response time was separated from the time allotted for the visual processing of the stimulus. We therefore speculate that display time thresholds represent a proxy for legibility (closely tied to visual processing), while response times in this experiment are more reflective of general cognitive-motor trends not specific to the typographic factors under study here. A large body of previous research suggests that slower response times are to be expected as participants age, and that in these types of psychophysical experiments, women tend to respond more slowly than men (Blough and Slavin, 1987; Fozard et al., 1994; Miller, 2001; Tun and Lachman, 2008).

4.1. Limitations

The present study drew on a sample of participants from a large metro area, and across a wide age range. While the repeated-measures design of the experiment and analyses reduces the impact of individual differences, considerable between-participant variability remains (as seen in Fig. 3). The influence of between-participant variability is particularly notable in the response time measures, which were sensitive to differences in age and gender and demonstrate effect sizes comparable to the significant experimental effects seen in the reading time threshold measures. As a result, there is evidence that the present study may have been underpowered for the detection of some of the statistical effects of interest. *A priori* power/sample size analyses are difficult or sometimes impossible to conduct for these types of mixed-model designs (Guo et al., 2013). Instead, we relied on prior experience suggesting that samples sizes similar to the one used here would be sufficient to successfully detect even relatively small effects under the present paradigm. In some cases, such as the borderline non-significant effect of age on legibility thresholds, it appears likely that an increased sample size would grant improved robustness and statistical power.

In contrast to many visual perception experiments, participants’ head positions in this experiment were not constrained by chin rests or other devices. This allowed for natural postural and positional variability that might be encountered in real-world device usage scenarios, particularly in-vehicle devices. The possibility exists that some participants may have leaned toward the screen, thus making experimental stimuli appear optically larger and therefore easier to read. However, we believe that such behavior is rare and unlikely to significantly affect experimental results in the aggregate. Previous experiments conducted in the same lab that required the continuous presence of a research assistant during data collection and used the same general paradigm suggested that participants maintained acceptable posture during the experiment session. We believe that an experiment measuring unconscious postural changes in response to difficult reading conditions, perhaps in correlation with a more thorough vision assessment, would make an intriguing follow-on study.

5. Conclusions

This study provides empirical evidence on how glance-based legibility is affected by crowded displays, text size, leading, and variable positioning, and how these factors do or do not interact with one another. As our interfaces of daily use increase in complexity and come to rely, frequently, on lists of neatly arranged textual information, it will be of paramount importance to recognize the trade-offs inherent in seemingly minor design decisions. For example, if a designer wishes to fit more information on the screen at once, he or she may reduce the text size, tighten the leading, choose a more visually compact typeface, or some combination of these. Studies such as these may help to provide actionable guidance on how each of these decisions can affect the legibility of the final design. The present data suggest, for instance, that changes to leading produce relatively small changes in legibility, whereas changes in size or overall information density are more dramatic, and therefore increased leading should not be expected to fully overcome the effect of reduced text size. Future research may extend our understanding of this design space along other dimensions, such as text contrast or compression. Consideration of the interaction of crowding with more dynamic real-life applications, such as smartphone use while walking or the use of in-vehicle displays while driving, may also warrant investigation to assess if the multi-tasking effect amplifies detriments due to increased information density.

Acknowledgements

This collaborative project was underwritten in part by Monotype Imaging Inc. through funding provided to MIT and in contribution of staff time and typographic expertise. We also wish to acknowledge the assistance of Hale McAnulty, Andrew Sipperly, and Anthony Pettinato in collecting the experimental data for this work.

References

Bakeman, R., 2005. Recommended effect size statistics for repeated measures designs. Behav. Res. Meth. 37 (3), 379–384.

Becker, D., Heinrich, J., Sichowsky, Von, R., Wendt, D., 1970. Reader preferences for typeface and leading. J. Typograp. Res., (Winter 1970) 61–66.

Bentley, M., 1921. Leading and legibility. Psychol. Monogr. 30 (3), 48–61. http://doi.org/10.1037/h0093140.

Blough, P.M., Slavin, L.K., 1987. Reaction time assessments of gender differences in visual-spatial performance. Percept. Psychophys. 41 (3), 276–281.

Bouma, H., 1970. Interaction effects in parafoveal letter recognition. Nature 226 (5241), 177–178.

Bouma, H., 1973. Visual interference in the parafoveal recognition of initial and final letters of words. Vis. Res. 13 (4), 767–782.

Chung, S.T.L., Li, R.W., Levi, D.M., 2007. Crowding between first- and second-order letter stimuli in normal foveal and peripheral vision. J. Vis. 7 (2) 10–10. http://doi.org/10.1167/7.2.10.

Coltheart, M., 1980. Iconic memory and visible persistence. Percept. Psychophys. 27 (3), 183–228.

Dobres, J., Chahine, N., Reimer, B., 2017a. Effects of ambient illumination, contrast polarity, and letter size on text legibility under glance-like reading. Appl. Ergon. 60 (C), 68–73. http://doi.org/10.1016/j.apergo.2016.11.001.

Dobres, J., Chahine, N., Reimer, B., Gould, D., Mehler, B., Coughlin, J.F., 2016a. Utilising psychophysical techniques to investigate the effects of age, typeface design, size and display polarity on glance legibility. Ergonomics 59 (10), 1377–1391. http://doi.org/10.1080/00140139.2015.1137637.

Dobres, J., Chrysler, S.T., Wolfe, B., Chahine, N., Reimer, B., 2017b. Empirical assessment of the legibility of the highway Gothic and clearview signage fonts. Transport. Res. Rec.: J. Transport. Res. Board 2624, 1–8. http://doi.org/10.3141/2624-01.

Dobres, J., Reimer, B., Chahine, N., 2016b. The effect of font weight and rendering system on glance-based text legibility. In: Presented at the Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Ann Arbor, MI, pp. 1–6. http://doi.org/10.1145/3003715.3005454.

Fozard, J.L., Vercryssen, M., Reynolds, S.L., Hancock, P.A., Quilter, R.E., 1994. Age differences and changes in reaction time: the baltimore longitudinal study of aging. J. Gerontol. 49 (4), P179–P189.

Guo, Y., Logan, H.L., Glueck, D.H., Muller, K.E., 2013. Selecting a sample size for studies with repeated measures. BMC Med. Res. Meth. 13, 100. http://doi.org/10.1186/1471-2288-13-100.

Holleran, P.A., Bauersfeld, K.G., 1993. Vertical Spacing of Computer-Presented Text. pp. 179–180. http://doi.org/10.1145/259964.260193.

International Standards Organization, 2007. Road Vehicles – Ergonomic Aspects of Transport Information and Control Systems – Occlusion Method to Assess Visual Demand Due to the Use of In-vehicle Systems (No. ISO 16673). International Standards Organization, Geneva.

International Standards Organization, 2009. Ergonomic Aspects of Transport Information and Control Systems. (No. 15008). Geneva, Switzerland.

Leek, M.R., 2001. Adaptive procedures in psychophysical research. Percept. Psychophys. 63 (8), 1279–1292.

Levitt, H., 1971. Transformed Up-Down methods in psychoacoustics. J. Acoust. Soc. Am. 49 (2B), 467–477. http://doi.org/http://dx.doi.org/10.1121/1.1912375.

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

Meyer, D.E., Schvaneveldt, R.W., 1971. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. J. Exp. Psychol. 90 (2), 227–234.

Miller, R.J., 2001. Gender differences in illusion response: the influence of spatial strategy and sex ratio. Sex. Roles 44 (3), 209–225.

Montani, V., Facoetti, A., Zorzi, M., 2014. The effect of decreased interletter spacing on orthographic processing. Psychonomic Bull. Rev. 22 (3), 824–832. http://doi.org/10.3758/s13423-014-0728-9.

Olejnik, S., Algina, J., 2003. Generalized eta and omega squared statistics: measures of effect size for some common research designs. Psychol. Meth. 8 (4), 434–447. http://doi.org/10.1037/1082-989X.8.4.434.

Paterson, D.G., Tinker, M.A., 1944. Eye movements in reading optimal and non-optimal typography. J. Exp. Psychol. 34 (1), 80–83. http://doi.org/10.1037/h0056763.

Paterson, D.G., Tinker, M.A., 1947. Influence of leading upon readability of newspaper type. J. Appl. Psychol. 31 (2), 160–163.

Peirce, J.W., 2008. Generating stimuli for neuroscience using PsychoPy. Front. Neuroinf. 2, 1–8. http://doi.org/10.3389/neuro.11.010.2008.

Pelli, D.G., Tillman, K.A., 2008. The uncrowded window of object recognition. Nat. Neurosci. 11 (10), 1129–1135. http://doi.org/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. J. Vis. 7 (2), 1–36. http://doi.org/10.1167/7.2.20.

Perea, M., Gomez, P., 2012. Increasing interletter spacing facilitates encoding of words. Psychonomic Bull. Rev. 19 (2), 332–338. http://doi.org/10.3758/s13423-011-0214-6.

Perea, M., Moret-Tatay, C., Gomez, P., 2011. The effects of interletter spacing in visual-word recognition. Acta Psychol. 137 (3), 345–351. http://doi.org/10.1016/j.actpsy.2011.04.003.

Poulton, E.C., 1972. Size, style, and vertical spacing in legibility of small typefaces. J. Appl. Psychol. 56 (2), 156–161.

R Core Team, 2018. R: a Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from. http://www.R-project.org/.

Reimer, B., Mehler, B., Dobres, J., Coughlin, J.F., Matteson, S., Gould, D., et al., 2014. Assessing the impact of typeface design in a text-rich automotive user interface. Ergonomics 57 (11), 1643–1658. http://doi.org/10.1080/00140139.2014.940000.

Roethlein, B.E., 1912. The relative legibility of different faces of printing types. Am. J. Psychol. 23 (1), 1. http://doi.org/10.2307/1413112.

Tinker, M.A., 1963. Influence of simultaneous variation in size of type, width of line, and leading for newspaper type. J. Appl. Psychol. 47 (6), 380–382. http://doi.org/10.1037/h0043573.

Tun, P.A., Lachman, M.E., 2008. Age differences in reaction time and attention in a national telephone sample of adults: education, sex, and task complexity matter. Dev. Psychol. 44 (5), 1421–1429. http://doi.org/10.1037/a0012845.

van Nes, F.L., 1986. Space, colour and typography on visual display terminals. Behav. Inf. Technol. 5 (2), 99–118. http://doi.org/10.1080/01449298608914504.

Whitney, D., Levi, D.M., 2011. Visual crowding: a fundamental limit on conscious perception and object recognition. Trends Cognit. Sci. 15 (4), 160–168. http://doi.org/10.1016/j.tics.2011.02.005.

Wilkins, A.J., Nimmo-Smith, M.I., 1987. The clarity and comfort of printed text. Ergonomics 30 (12), 1705–1720. http://doi.org/10.1080/00140138708966059.

Zhang, J.-Y., Zhang, T., Xue, F., Liu, L., Yu, C., 2009. Legibility of Chinese characters in peripheral vision and the top-down influences on crowding. Vis. Res. 49 (1), 44–53. http://doi.org/10.1016/j.visres.2008.09.021.