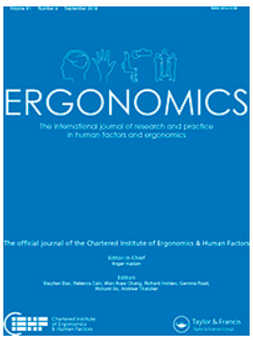

Ergonomics

ISSN: 0014-0139 (Print) 1366-5847 (Online) Journal homepage: https://www.tandfonline.com/loi/terg20

Glanceable, legible typography over complex backgrounds

Ben D. Sawyer, Benjamin Wolfe, Jonathan Dobres, Nadine Chahine, Bruce Mehler & Bryan Reimer

To cite this article: Ben D. Sawyer, Benjamin Wolfe, Jonathan Dobres, Nadine Chahine, Bruce Mehler & Bryan Reimer (2020): Glanceable, legible typography over complex backgrounds, Ergonomics, DOI: 10.1080/00140139.2020.1758348

To link to this article: https://doi.org/10.1080/00140139.2020.1758348

Published online: 19 May 2020.

ERGONOMICS
https://doi.org/10.1080/00140139.2020.1758348

Taylor & Francis
Taylor & Francis Group

Glanceable, legible typography over complex backgrounds

Ben D. Sawyer[a,b], Benjamin Wolfe[c,b], Jonathan Dobres[b], Nadine Chahine[d], Bruce Mehler[b] and Bryan Reimer[b]

[a]Industrial Engineering and Management Systems, University of Central Florida, Orlando, FL, USA; [b]AgeLab, Massachusetts Institute of Technology, Cambridge, MA, USA; [c]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA; [d]ArabicType, London, UK

ABSTRACT

Modern digital interfaces display typeface in ways new to the 500 year old art of typography, driving a shift in reading from primarily long-form to increasingly short-form. In safety-critical settings, such at-a-glance reading competes with the need to understand the environment. To keep both type and the environment legible, a variety of 'middle layer' approaches are employed. But what is the best approach to presenting type over complex backgrounds so as to preserve legibility? This work tests and ranks middle layers in three studies. In the first study, Gaussian blur and semi-transparent 'scrim' middle layer techniques best maximise legibility. In the second, an optimal combination of the two is identified. In the third, letter-localised middle layers are tested, with results favouring drop-shadows. These results, discussed in mixed reality (MR) including overlays, virtual reality (VR), and augmented reality (AR), considers a future in which glanceable reading amidst complex backgrounds is common.

Practitioner summary: Typography over complex backgrounds, meant to be read and understood at a glance, was once niche but today is a growing design challenge for graphical user interface HCI. We provide a technique, evidence-based strategies, and illuminating results for maximising legibility of glanceable typography over complex backgrounds.

Abbreviations: AR: augmented reality; VR: virtual reality; HUD: head-up display; OLED: organic light-emitting diode; UX: user experience; MS: millisecond; CM: centimeter

ARTICLE HISTORY
Received 6 June 2019
Accepted 22 February 2020

KEYWORDS
Perception; vision and lighting; environmental ergonomics; information displays; human-machine systems; mixed reality virtual environments human-computer interaction

Introduction

The applied question of legibility, how text may be presented so as to be more easily read (c.f. Graham 2012), is not a new topic of study (see, for example, Roethlein 1912). What is, however, radically new is a growing prevalence of types of reading and typographic presentation once niche. Two of these are the focus of the present work: (a) reading at-a-glance and (b) doing so with type situated over complex backgrounds. This seemingly inadvisable situation is becoming commonplace in various types of interfaces, and is particularly prominent in the proliferation of computationally powerful portable electronics with high resolution screens and see Hancock, Sawyer, and Stafford 2015. Music track information, for example, is layered over album art, distance information over a camera feed, or street names over satellite imagery. Text displayed as a layer over the natural world is indeed promising to become increasingly common, as advances in augmented reality (AR) in support of ubiquitous computing (Weiser 1993), the current proliferation of virtual reality (VR), and indeed the emerging technological feasibility of the entire mixed reality continuum (Milgram and Kishino 1994; Milgram et al. 1995). Amidst this continued rise of glanceable reading over complex backgrounds, data-grounded efforts to understand how to best design in such situations are vital.

At-a-glance reading is not new, and indeed signage has arguably been 'augmenting' reality since shortly after the advent of writing, but a growing proportion of reading is now directed towards glanceable content. That change, and the growth in content designed to be consumed at speed, has recently been driven largely by wide-scale adoption of digital displays intended to be read on-the-go. Smartphones, for

CONTACT Ben D. Sawyer bsawyer@mit.edu; sawyer@inhumanfactors.com Industrial Engineering and Management Systems, University of Central Florida, Orlando, FL 32816-2993, USA; AgeLab, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/terg.

example, certainly support long-form reading, but interaction with these devices is largely 'glanceable'. Such short-form glanceable reading can be distinguished by the viewer's rapid acquisition of textual information in a single or small set of fixation(s): the weather, a text message, or name of a caller are all examples of modern day glanceable reading. The degree to which an interface can facilitate reading-at-a-glance is crucially important, especially in safety-critical operational contexts. Consider surface transport: when a driver looks to their GPS to read the name of the street on which they must soon turn, they are moving their attention away from potential roadway hazards. Recent research assessing the impact of reading short messages while driving describes not only impact to onroad event response time (Caird et al. 2014), but differential impact when reading text in different typefaces (Reimer et al. 2014). Crucially, it is important in such situations to not simply recommend the user focus on the operation task (driving, etc.) but also to consider the optimisation of the messages, so as to design the interfere to be minimally demanding on attention (and see Skrypchuk, Langdon, Sawyer, and Clarkson 2019; Skrypchuk, Langdon, Sawyer, et al. 2019).

The background over which text is presented is a neglected aspect of the literature, and one in need of expansion as the convention of the white page is continually challenged. Digital reading devices allow readers to select all matter of backgrounds, and AR head-up displays (HUDs) have been put forth as a way to mitigate the divided attention by bringing reading inline with visually complex environments. For example, a windshield HUD might allow information to be presented to a driver without necessitating glances away from the roadway, a security camera might provide location information over feed, or a display might deliver notification information superimposed over a user-selected background. While benefits can be imagined in each case, these technologies incur new design complications related to the display of text over complex, real-world backgrounds. Performance gains are far from certain (see Sawyer et al. 2014, Rusch et al. 2013). How can a designer provide a multitasking user with crucial information, before returning their gaze to an equally crucial ongoing task? Not addressed in this work are systems where visual depth cues are presented, the ongoing evolution of associated technologies and resultant focal issues, including accommodation-vergence conflict (and see Hoffman et al. 2008). We here focus on the applied question of textual legibility in glanceable reading over various backgrounds. How can typographic design be leveraged to facilitate fast glanceable reading, superimposed over both simple and complex backgrounds? This paper addresses this relationship between typography and legibility rather differently from previous works on reading. We investigate reading not as a long-form task, but as a task performed at speed for single words, and not only on simple, monochromatic backgrounds, but also on complex backgrounds.

Questions of legibility have interested researchers nearly from the outset of modern studies of perception and psychology, a timespan that subtends significant changes in printing, display, rendering and other technologies. The earliest studies of eye movements—done to understand how a reader's gaze moves across a sentence and the page—laid the groundwork for the last century of eye movement research (Javal 1878, as described in Huey 1908). About the same time as the advent of hot lead type, typographical manipulations impacting text legibility were first being studied with the explicit goal of improving reading speed and accuracy (c.f. Sanford 1888). Perhaps the earliest direct precursor to our work is that of Roethlein, who presented individual words in diverse typefaces to subjects, and determined minimum viewing distance for words of a given size (1912). During this same period, printing technology has moved from hand-set cold type to hot lead type to modern offset printing, and since the 1990s has increasingly eschewed paper for digital text on screens. These digital interfaces have enabled greater design freedom, but present new challenges inherent to the medium and how it differs from printed text on paper. This, notably, includes the ease with which type elements can be superimposed over photos, illustrations, and other richly complex backgrounds.

Prior work out of the MIT AgeLab in this space began with the assessment of legibility differences in the glance-based context of a driving simulation experiment (Reimer et al. 2014). More recently, our efforts have couched legibility in typographical design as a vision science-grounded empirical effort (Dobres et al. 2018; Dobres et al. 2016; Sawyer et al. 2017, 2020; Wolfe et al. 2016). These investigations have used calibrated legibility thresholds, or the duration of a fixation during which an average reader can make a decision about a single word, as well as accuracy data, as a window into understanding legibility. Using this approach, the importance to legibility of (a) size (c.f. Roethlein 1912) has been showed to hold true on modern backlit digital displays (Dobres et al. 2016). Likewise, (b) choice of typeface has been shown to be

important for legibility (Dobres et al. 2016; Sawyer et al. 2017) and to effect higher reading speed in skilled adult readers (Wallace et al., 2020). Legibility thresholds have been used for benchmarking the relative legibility of typefaces in 'typographic bakeoffs' (as in Sawyer et al. 2017) which could potentially allow the vast libraries of typefaces presently available to be benchmarked in terms of legibility. Such efforts could span language and character set, while effects across polarity and typeface are consistent between English and Italian (Dobres, Chahine, and Reimer 2017), Chinese characters likewise exhibit typeface-based legibility differences (Dobres et al. 2016b) that, perhaps for reasons particular to the script, did not parallel those observed with Latin character legibility. Significant impacts on legibility, as measured by increased thresholds, were found in research for both (c) all-lowercase typeface and (d) condensed typeface (Sawyer et al. 2017). Though the lowercase finding is counter-intuitive, we believe that this result is driven by the smaller area size that lowercase letterforms occupy. Finally, legibility costs (e.g. longer duration thresholds) have been shown for (e) typeface weight in Latin (Dobres et al. 2016) and Chinese characters (Dobres et al. 2016), where heavier weights in Latin were less legible, while Chinese results showed improved legibility with heavier weights.

Legibility research has identified a number of important considerations beyond those mentioned above. For example, studies comparing polarities on digital screens favour positive polarity (Dobres et al. 2016; Dobres, Chahine, and Reimer 2017). The importance of text size, polarity and background illumination illustrate a more complex relationship with the favorability of positive polarity being amplified in low ambient conditions (Dobres, Chahine, and Reimer 2017). These findings have emerged alongside a user experience (UX) movement towards negative polarity design driven by the new technology of organic light-emitting diode (OLED) displays, which consume less power as they display more black. Previous works on design and typography do note the importance of maximising contrast between background colours and text colours (Graham 2012). It is only recently that technology has allowed users to choose, on-the-fly, between positive polarity, dark text on a light background, and negative polarity, light text on a dark background. Likewise, positional uncertainty, or displaying information in a different place over time, has been shown to hinder legibility (Dobres et al. 2018). Together, this body of work speaks to a range of questions for glanceable reading on screens, but moving beyond screens to augmented reality poses its own challenges.

It is an open question how much these prior findings will translate to AR, where text must be presented over the user's view of the larger environment. Legible typographical or symbolic overlays in such environments are the foundation of head-up displays (HUDs), head/helmet mounted displays (HMDs), automotive backup cameras, and other AR systems. The question of legibility in these environments is particularly timely: overlay technologies are becoming relatively financially and computationally inexpensive, and therefore AR implementations are increasingly common. Backup cameras, dashboard mounted HUDs, and other overlay technologies in vehicle UX formerly associated only with luxury vehicles are now being found on midrange models. Such interfaces may soon be standard in every new vehicle. The text overlay technology literature is presently dominated by evaluations of specific technologies, while generalised research into the psychophysical underpinnings of legibility in such systems is relatively sparse. Subtitled video has been studied in this regard (Neve and Jenniskens 1994), indicating that reading is facilitated with solid backgrounds and high contrast, even at the cost of occluding a portion of the video. A slightly larger body of literature has more directly addressed questions of text legibility in AR, both in HUDs and HMDs. Work here includes an investigation of optimal text colour and background for AR by Debernardis et al. (2014), who suggest white text on a fully occlusive blue background to maximise legibility. In addition, Gabbard and colleagues examined the effects of background texture, visual distance, and ambient illumination (Gabbard, Swan, and Hix 2006), all of which impact legibility. They also suggest fully occlusive opaque squares, or 'billboards', as the background to textual information to minimise reaction time. However, this is complicated by the optical requirements of HUDs and HMDs, since both ambient illumination and the capabilities of the AR device limit the ability to provide these maximally useful conditions in the real-world (c.f. Kress and Starner 2013). Modelling efforts have also provided optimal luminance at–the-eye curves for HMD iconography presented in military vehicles (Harding, Martin, and Rash 2005, 2007), further emphasising the need to consider the problem in its real-world context. While some of these studies get close, none of them speak to our question: how can designers best present text in augmented reality?

As such, the present work embarked on three studies to better understand which design strategies might

best facilitate typeface legibility over complex backgrounds. Holding other known variables of consequence constant, Experiment I investigates so-called 'middle layer' strategies, in which the background is digitally manipulated. Experiment II compares the best of the strategies emerging from this work in order to find an ideal level of background manipulation. Experiment III looks to typeface manipulation for legibility (i.e. manipulating the foreground elements rather than the background), specifically outline and shadow, both in combination with and in the absence of a middle layer. In all studies, the overarching theme is the search for typographic permutations which maximise legibility for at-a-glance reading in AR applications. To these questions we bring an age and gender balanced population of individuals who, as with all readers, use glanceable typography in their day-to-day lives.

Experiment I

Designers, and ultimately users, have a need for crucial interface elements to remain legible across a wide range of possible complex backgrounds. One common solution to this dilemma is the adoption of 'middle layer' techniques, in which a manipulation to the complex background layer enhances legibility of a foreground text layer. Existing psychophysical work investigating the legibility effects of such techniques is sparse, but a sizable body of work supports the need for middle layer approaches. Background pattern spatial frequency strongly affects readability, with backgrounds exerting a stronger masking effect when pattern width is comparable to letter weight, and letter size itself exerting no significant influence (Petkov and Westenberg 2003). Consider 'billboards' of colour interposed behind content and background, which facilitate reading text faster and more accurately (Gabbard, Swan, and Hix 2006). Billboards also often completely obscure the background, an aesthetic drawback in some systems, but a practical safety drawback when vital information is present, as in automotive backup cameras.

Designers generally compromise, blurring or uniformly dimming/brightening the background layer. This improves contrast between background and text, a vital cue in human perception (Legge et al. 1990) with a well-understood range of acceptable values, at least in office environments (ISO 9241). Designers here face a balance: improving contrast removes information about a background which may fulfil an aesthetic (e.g. album art) and/or situation awareness (e.g. backup camera) function. Surprisingly, there has been almost no work that directly examines the relationship between manipulations of a middle layer and the legibility of the foreground. As dynamic interfaces enter the vehicle and other time critical contexts, it will be important to understand how dynamic background elements and the middle layers that mediate them affect legibility at a glance. Optimal levels of such middle layer 'strength' are generally judged by the individual designer.

To investigate an objective optimal formulation of middle layers, Experiment I investigates the glance legibility of text set against complex backgrounds. In addition to an unmodified control condition, several types of middle layers suggested by the literature and our own experience are studied at various intensities. These include the traditional semi-transparent layering and Gaussian blurring, as well as the removal of horizontal and vertical information in the spatial frequency domain. We hypothesised that each of these middle layer approaches would improve legibility.

Experiment I methods

A total of 42 participants between the ages of 35 and 75 were recruited, a range chosen from previous research (see Wolfe et al. 2016), and provided written informed consent prior to participating. Of these, five participants were excluded for having response accuracies of less than 50% (worse than chance performance) in several assessed conditions, two participants were excluded because their display time threshold estimates were 200 ms or greater (outliers in this sample), and 1 was excluded due to technical failure. This left a total of 34 participants in the analysis sample, 18 female (mean age = 58.8) and 16 male (mean age = 58.6).

Background images were sampled from the ImageNet database (Deng et al. 2009). To ensure a wide array of 'unpredictable' images, candidates were sampled from categories that spanned naturalistic, artificial, sparse, and crowded categories. Keywords included: metal, wood, orchestra, grating, window, blind, alluvial, mountain, crowd, pattern, plant, city, road, and tree. Because the images were to be used as backgrounds in the experiment, and white text stimuli would be displayed at their centre, the image pool was reduced according to the following parameters: (a) width between 500 and 800 pixels, (b) height of at least 350 pixels, (c) average image brightness between 102 and 107 (inclusive, out of a maximum value of 255), and (d) within a 10-pixel square at the

image centre, an average brightness of between 77 and 102 (inclusive, out of a maximum value of 255). This left a total of 160 eligible images in the pool. In the actual experiment, background images were scaled up by 25% in each dimension to fully utilise the high-resolution monitor (see section 'Apparatus and setting'). Screen area beyond the image boundary was filled with a uniform grey (RGB: 110, Display: Acer H257HU, Brightness 350 cd/m^2).

The present study examines 21 middle layer conditions (see below) using a lexical decision task (as in Dobres et al. 2016; Sawyer et al. 2017). A simple yes/no decision was made as to whether a string of letters was a true word (record) or a psuedoword (throps), a pronounceable non-word. All stimuli were six letters in length. True words were selected from an online orthographic database (see Medler & Binder 2005), while pseudowords were generated as detailed by Dobres et al. (2016), and included in their Supplemental Materials for that work. The display time of each word/pseudoword was adjusted depending on performance. Before experimental trials began, participants completed two independent legibility threshold calibration trials utilising an adaptive staircase procedure. As in work by Dobres and colleagues (2016), text stimuli during this section were set in Frutiger at a 4 mm capital letter height, in black (hex: #000000) against a plain white background (hex: #ffffff). The white background was a 700×400 pixel rectangle at the centre of the screen. The edges of the rectangle were Gaussian blurred (SD $=$ 50 pixels) to minimise afterimage effects. Stimulus duration was set at 800 ms then decreased every three trials, to 6,00,400 ms, and finally 200 ms. Thereafter, a '3-down 1-up' rule incremented stimulus duration by 200 ms for three trials, and 20% less for each three trials thereafter, with a floor of 33.4 ms. This established the minimum amount of time needed for each participant to read a word presented under highly legible conditions with approximately 79.4% accuracy. The median value of these two blocks (rounded to the nearest monitor refresh interval) was taken as representative of the participant's 79.4% accuracy thresholds. This threshold value was then used as the display time for all stimuli across the 21 conditions of interest. Therefore, the dependent variable in all middle layer conditions was subject performance, given a fixed stimulus duration.

Each participant completed 800 lexical decision trials, which was deemed to be the upper limit of what could be completed in a single session. As described above, each session began with two threshold calibration blocks (100 trials each). The remaining 600 trials were used to test the 21 middle layer conditions, as described below. All experimental trials were randomly interleaved, as were the exact background images used per condition. The same font and text size from the thresholding portion were used throughout, and text was set in white (hex: #ffffff). Each image in the pool of 160 was processed to produce an appearance consistent with each of the 21 tested conditions, thus backgrounds could be repeated across conditions.

A random selection of backgrounds was presented without modification across 40 trials of the experiment as a control condition.

A semi-transparent layer was presented across 140 trials of the experiment, or 28 per intensity level. The image's mean colour was calculated, and then the image was interpolated towards the mean. An interpolation of 0% would leave the image unmodified, while 100% would simplify the image to its uniform mean colour. Interpolation levels used were 10%, 20%, 30%, 40%, and 50% (see examples in Figure 1).

A Gaussian blur layer was presented across 140 trials, or 28 per intensity level. Backgrounds were blurred with a 2 D Gaussian-weighted function, an aesthetically pleasing effect that is increasingly common in interface design. The technique is easily accessible in image editors, as a function in many common image processing libraries. Although a Gaussian kernel expression would be used in most vision science research, blur intensity levels are here expressed as the standard deviation (SD) of the Gaussian function, in pixels. This is their usual expression in the applied domain of digital displays. Blur intensities (SDs) used were 1.00, 1.75, 2.50, 3.25, and 4.00 (see examples in Figure 2). Note that Gaussian blurs cannot occur within 'fraction pixels', and we make no claim to have fractional pixels. Rather, antialiasing and other technologies which smooth pixels perceptually do play a role here to create a difference in the display characteristics of, for example, SD 1.00 and 1.75. This does mean that our results in this work explicitly reference our hardware and software, which work together to influence such technologies (for another example, see Dobres et al. 2016).

A 'Fourier domain horizontal filtering', or Fourier layer, was presented across 140 trials, or 28 per level. The Gaussian blur technique, described above, can be thought of as a '2 D weighted average', but it may also be thought of as 'removal of high frequency image features at all orientations'. Images are commonly thought of as mosaics of pixel intensities,

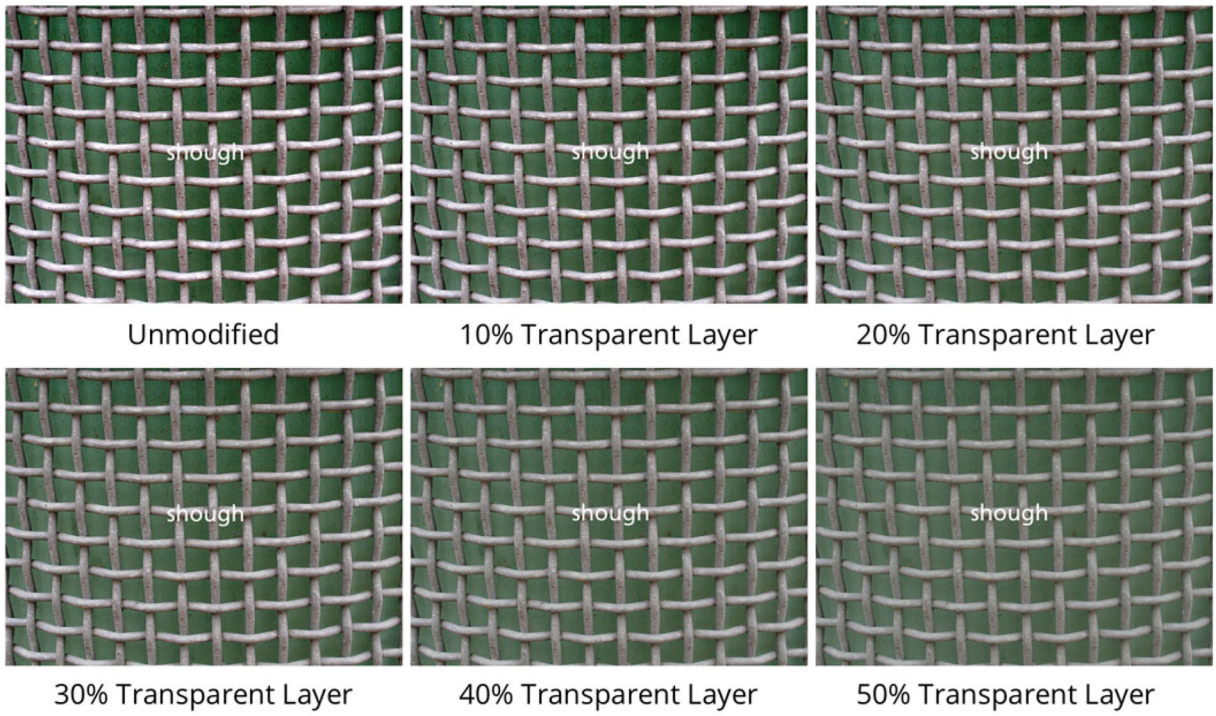

Figure 1. Examples of the middle layer resulting from interpolation with the image's mean colour, producing a semi-transparent layer effect. A pseudoword, like those displayed to participants, is shown in the foreground.

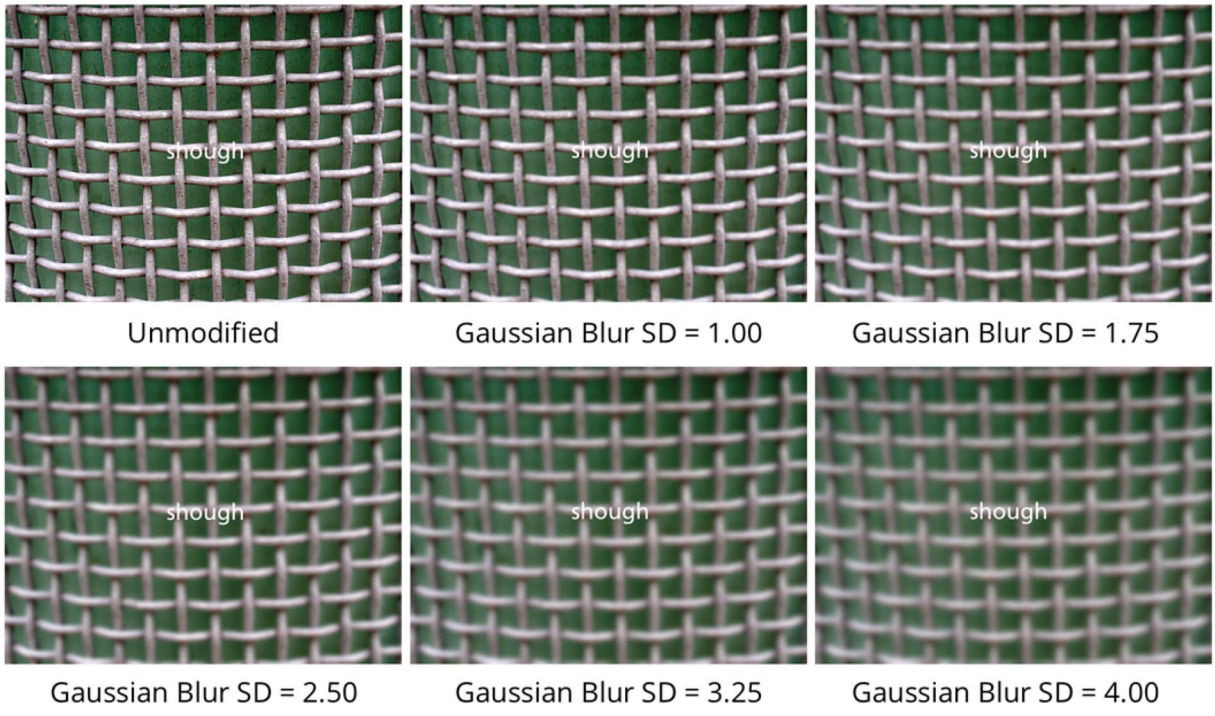

Figure 2. Examples of the middle layer resulting from Gaussian blur, a weighted average of pixels within a radius approximately three times larger than the specified SD. An SD of 1.0 would include weighted pixel values from a surrounding area of approximately 6 pixels in diameter. A pseudoword, like those displayed to participants, is shown in the foreground.

however, any image may also be thought of as a summation of waves of many orientations, frequencies, and powers (see Field 1987). The discrete Fourier transformation can be used to convert an image from Cartesian (or 'normal') space to Fourier (or 'frequency') space (see bottom row Figure 3). In a Fourier image, each pixel represents a wave, with the amplitude encoded as the pixel's brightness, its frequency as the

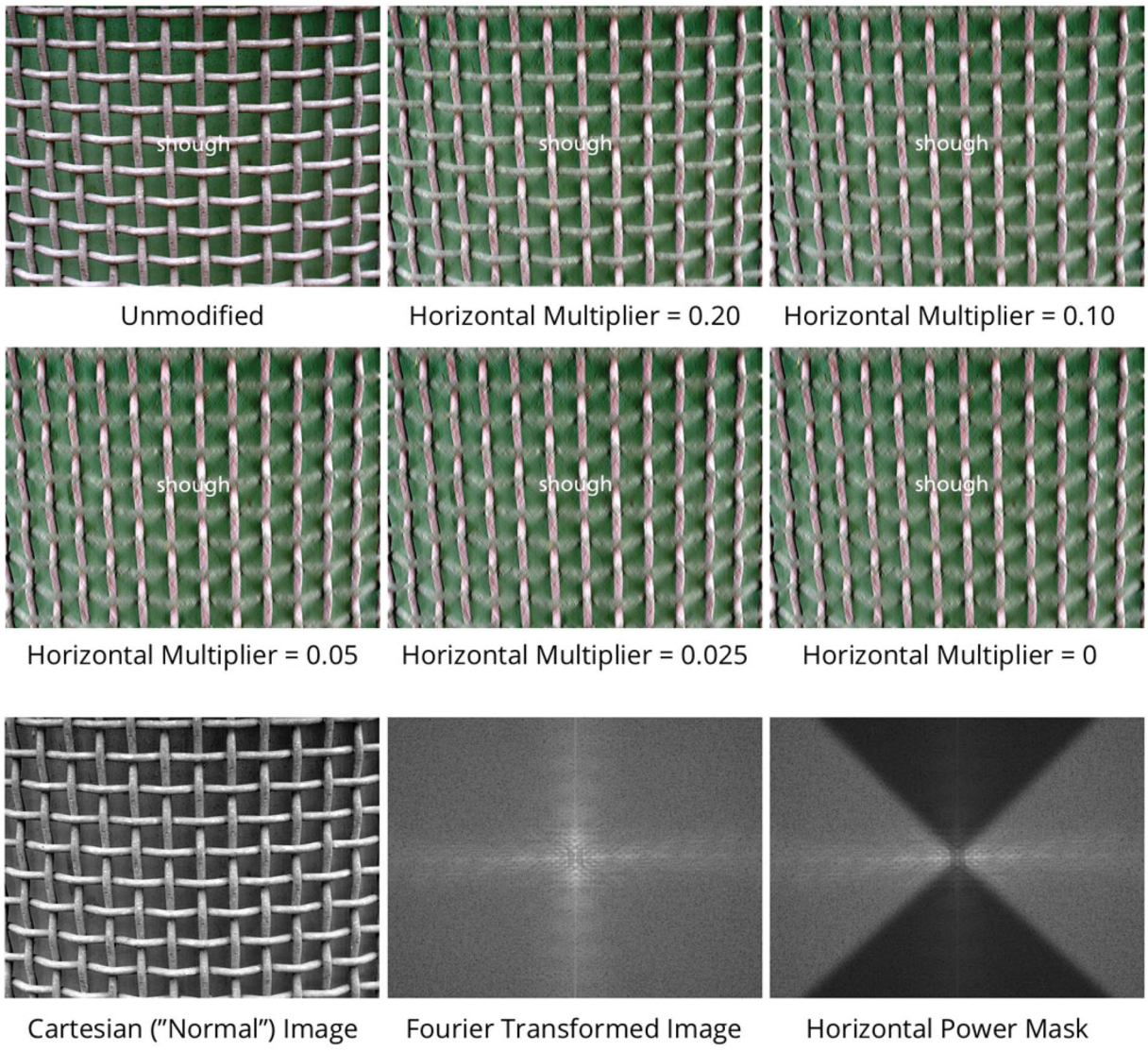

Figure 3. Examples of the middle layer resulting from filtering out horizontal information in the Fourier domain (top two rows). Illustration of Fourier domain transformation and filtering (bottom row). Note that the Fourier images have been log transformed and normalised to better expose the range of intensities present (untransformed, the image's centre pixel would have several orders of magnitude more power than any other pixel, resulting in an image that appears almost entirely black). A pseudoword, like those displayed to participants, is shown in the foreground.

pixel's distance from the image centre, and the wave's orientation as the pixel's angle from image centre. As can be seen in the example above, the image of the grate has prominent vertical and horizontal features. These appear in the Fourier image as bright horizontal and vertical bands radiating from the image centre. The final image at the bottom of Figure 3 shows a simple orientation mask applied in Fourier space, which reduces the power of all waves that are ±45 degrees from horizontal. When the image is transformed back to Cartesian space, the result is an image in which horizontal contours are blurred. The blurring is specific to horizontal contours, leaving vertical and off-vertical image data intact. If a Gaussian blur is 'reduction of high frequency image data at all orientations', then this technique is 'reduction of high frequency image data at selected orientations'. Images were generated that reduced horizontal power to varying extents. The scale for this effect is nonlinear and somewhat difficult to quantify at present, and is best thought of as a 'multiplier'. A multiplier of multiplier of 1.0 leaves the initial image intact, while a multiplier of 0.0 masks horizontal contours entirely.

Reducing image data in this way has the effect of darkening the image. To counteract this, the Fourier image's central pixel value was preserved and left unaffected by the mask. This ensures that, when the Fourier image is transformed back to Cartesian space,

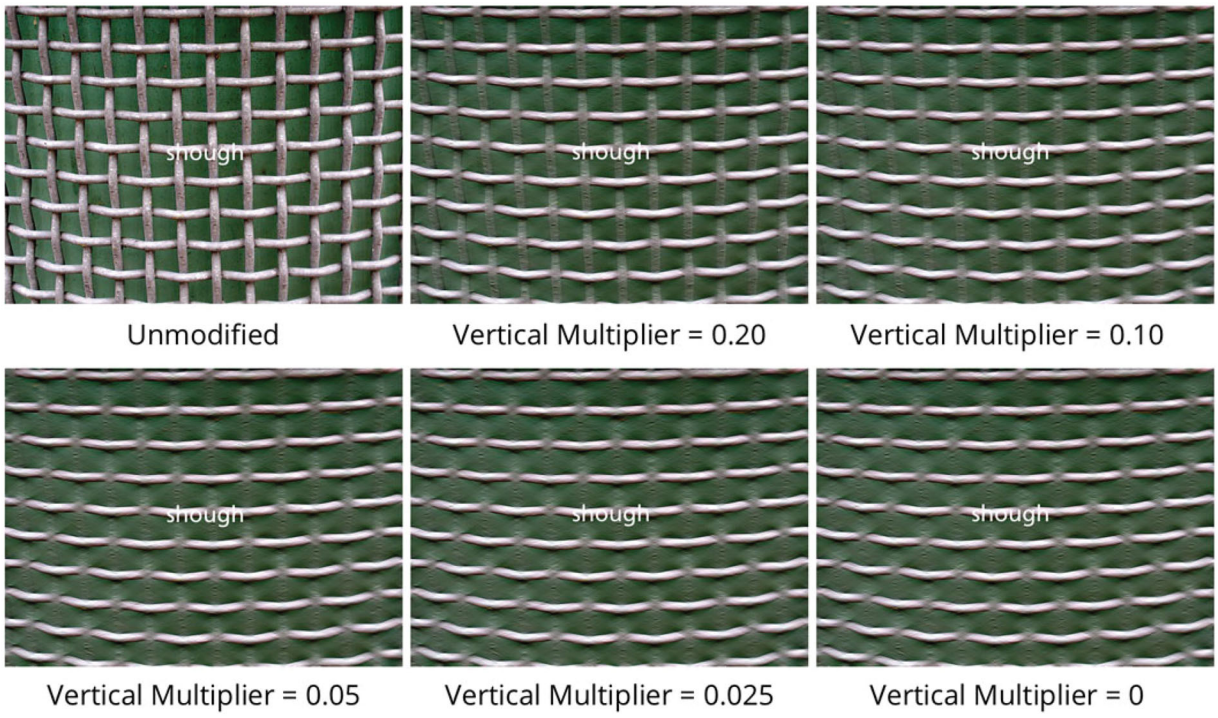

Figure 4. Examples of the middle layer resulting from filtering out vertical information in the Fourier domain. A pseudoword, like those displayed to participants, is shown in the foreground.

its mean brightness is left intact. While this preserves brightness, it also has the effect of compressing the image's colours towards the image's mean colour, similar to the semi-transparent layer technique described earlier. Therefore, as a final post-processing step, image colour values were rescaled to counteract the compression in colour space. While an imperfect solution at present, this generally had the effect of restoring image contrast and reducing the 'semi-transparent' appearance (Figure 4).

A 'Fourier domain vertical filtering' middle layer was applied over 140 trials, or 28 per intensity level. Images were filtered in the Fourier domain and post-processed using the same methods as described for horizontal domain filtering, with the exception that the filter was changed to affect waves that were ±45 degrees off vertical.

Apparatus and setting

The experiment was run on a Mac Mini running Mac OS 10.10.5 (2.5 Ghz Intel Core i5 CPU, 4GB of RAM) running PsychoPy (Peirce 2008). Stimuli were displayed on a high-resolution Acer monitor (21.77" × 12.24", 2560 × 1440 pixels, 60 Hz refresh rate). Participants viewed the screen from a distance of approximately 70 cm. Data were collected in two separate experiment rooms that used identical hardware and software configurations. Both rooms were quiet and dimly lit (as measured with a photometer to ensure equal ambient illumination <10 lux). Average reading time thresholds did not differ between rooms ($t(21) = 0.42$, $p = 0.682$). As a note, we retained this monitor and environment for experiments II and III. At the conclusion of the experimental portion of the study, participants were debriefed and paid for their time (Figure 5).

Experiment I results

Thresholds

Each participant completed two consecutive blocks of threshold assessment, and threshold estimates decreased significantly between the first and second assessment, consistent with a practice/familiarity effect $t(33) = 2.69$, $p = 0.011$. Participants required an average of 100 ms of display time to read highly legible text with 79.4% accuracy. This is slightly but significantly more time (82.3 vs 100 ms: $t(33) = 2.49$, $p = 0.018$) than was assessed in an early study of the same basic typographic configuration (Dobres et al. 2016). Across background image trials (i.e. excluding the threshold calibration sections at the start of each session), a Wilcoxon signed rank test revealed that response times decreased significantly over the time course of the experiment ($V = 535$, $p < 0.001$). Mean response time during the first 100 trials was 495 ms ($SE = 4.84$ ms), compared to 393 ms during the final

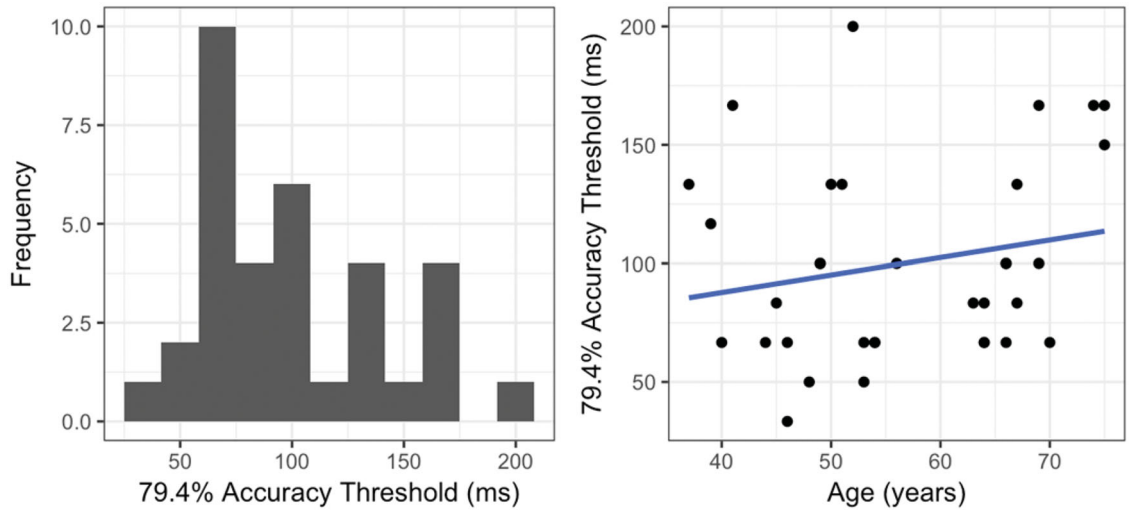

Figure 5. (Left Panel) Histogram of display time thresholds corresponding to approximately 79.4% response accuracy. (Right Panel) Legibility thresholds visualised against participant age. The blue line is a linear regression through the data.

100 trials ($SE = 4.00$ ms). Friedman's rank sum test revealed no significant differences among different background image conditions ($\chi^2 = 20.51$, $p = 0.427$). The lack of a significant difference in response times between conditions, combined with the fact that conditions were randomly interleaved among each other, supports the premise that the observed response time effect was consistent with well-known habituation and learning effects common to most psychophysical studies. Thresholds rose slightly with age, but this increase was not statistically significant $t(33) = 1.16$, $p = 0.255$.

The unmodified backgrounds were used as a control, a base reference point with no manipulation. When reading text set against a random selection of unmodified backgrounds, accuracy averaged 74.9%. As was expected, this is significantly lower than results obtained during threshold calibration against a plain white background $t(33) = 2.88$, $p = 0.007$. Accuracy in the control condition did not depend upon age $t(33) = 0.79$, $p = 0.434$, and ranged between 52.5% and 90%. Although the high end of this accuracy range was unexpected, only 2 participants had accuracies at this level in some conditions. Since their tentative removal from the sample had little effect on the overall pattern of results and a less restricted sample was considered beneficial, and they were retained in the analysis.

Semi-transparent layer

As shown in Figure 6, response accuracies were significantly different across the intensity levels of the semi-transparent middle layer $F(1, 33) = 11.90$, $p = 0.002$. Changes in accuracy follow a step-like pattern, with the higher middle layer intensities resulting in higher

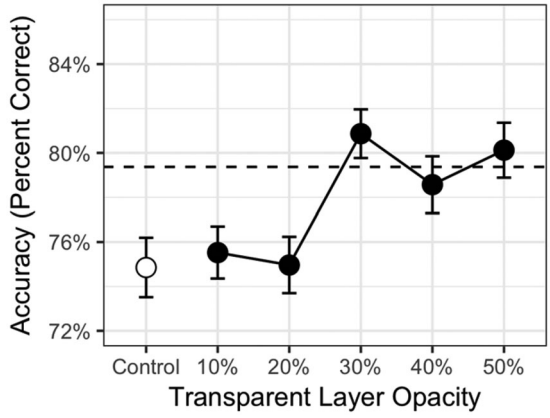

Figure 6. Response accuracies across different intensity levels of the semi-transparent middle layer condition (higher numbers on the *x*-axis indicate a stronger middle layer). Response accuracy in the unmodified control condition is shown in white. Error bars represent ±1 mean-adjusted SEM. The dashed horizontal line represents the theoretical accuracy level obtained during display time calibration under ideal conditions.

response accuracy. Table 1 summarises the key statistical relationships in these data. Performance at the 10% and 20% intensity levels were not significantly different from performance in the control condition, but were significantly lower than the high legibility accuracy point, suggesting that these intensities did little to mediate the background layer. Conversely, accuracies for intensity levels 30% and higher were significantly greater than the control condition, but not significantly different from the high legibility condition. This indicates that these intensity levels created legibility conditions comparable to the high legibility condition.

As shown in Figure 7, response accuracies were significantly different across the intensity levels of

Table 1. Results of statistical tests comparing the response accuracy obtained at each intensity level of the middle layer to the accuracy obtained in the control condition, as well as a comparison to the theoretical accuracy point of 79.4% obtained from a high legibility condition.

Middle layer opacity	Different from control?	Different from high legibility?
10%	No [$t = 0.35$, $p = 0.728$]	Yes [$t = -2.39$, $p = 0.023$]
20%	No [$t = 0.06$, $p = 0.950$]	Yes [$t = -2.91$, $p = 0.006$]
30%	Yes [$t = 3.60$, $p = 0.001$]	No [$t = 1.01$, $p = 0.318$]
40%	Yes* [$t = 2.03$, $p = 0.050$]	No [$t = -0.46$, $p = 0.651$]
50%	Yes [$t = 2.45$, $p = 0.020$]	No [$t = 0.49$, $p = 0.624$]

Asterisk denotes a borderline significant difference.

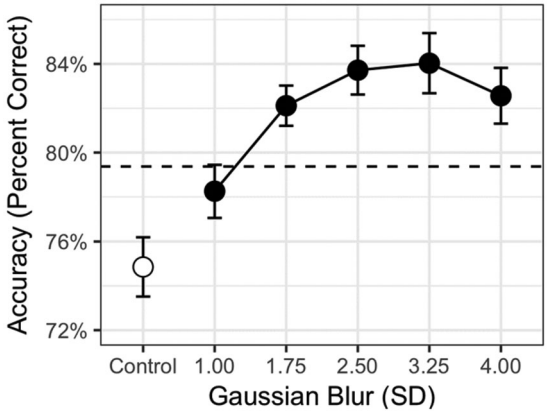

Figure 7. Response accuracies across different levels of the Gaussian blur middle layer condition (higher numbers on the *x*-axis indicate more blurring). Labelling as in Figure 6.

Table 2. Results of statistical tests comparing the response accuracies of the Gaussian blur middle layers to the control and high legibility conditions.

Gaussian blur SD	Different from control?	Different from high legibility?
1.00	Yes* [$t = 1.75$, $p = 0.089$]	No [$t = -0.77$, $p = 0.445$]
1.75	Yes [$t = 4.83$, $p = 0.000$]	Yes [$t = 2.15$, $p = 0.039$]
2.50	Yes [$t = 5.46$, $p = 0.000$]	Yes [$t = 3.11$, $p = 0.004$]
3.25	Yes [$t = 4.55$, $p = 0.000$]	Yes [$t = 3.26$, $p = 0.003$]
4.00	Yes [$t = 3.86$, $p = 0.001$]	Yes [$t = 2.25$, $p = 0.031$]

Asterisk denotes a borderline significant difference.

Gaussian blur middle layer $F(1, 33) = 9.97$, $p = 0.003$. Changes in accuracy follow a non-linear increasing pattern, with small amounts of blur resulting in larger gains in performance accuracy, while further blur plateaus performance. Table 2 summarises the key statistical relationships in these data. A Gaussian blur with an SD of 1.00 pixels results in a substantial increase in response accuracy. Though the increase in accuracy from the control condition is not quite statistically significant, the relationship trends towards significance and may reach this mark with a larger sample. A Gaussian blur of 1.75 pixels produces another large increase in accuracy. Not only is accuracy in this condition significantly greater than the control condition, it is also significantly greater than the accuracy point used in the high legibility condition. The same is true for blurs of 2.50, 3.25, and 4.00 pixels, which are not

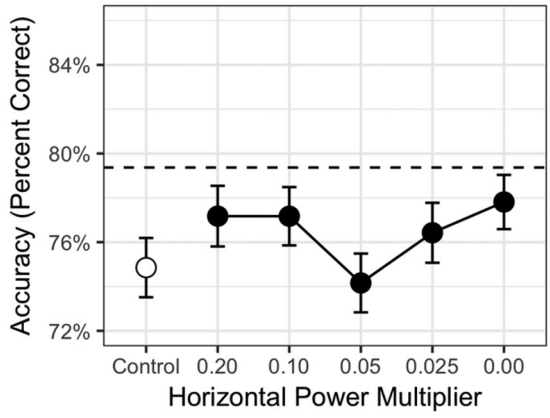

Figure 8. Response accuracies across different levels of the horizontal information filter middle layer condition (lower numbers on the *x*-axis indicate greater reductions in horizontal information; as in previous plots, the middle layer increases in intensity when read from left to right). Other labelling as in Figure 6.

significantly different from performance accuracy at the 1.75 blur level (all $p > 0.25$, paired t-tests) (Figure 8).

Accuracy did not differ between levels of the horizontal filter $F(1, 33) = 0.07$, $p = 0.791$. Averaged across intensity levels, response accuracy was not significantly different from accuracy in the control condition $t(33) = 1.13$, $p = 0.266$, but was significantly lower than the high legibility accuracy point $t(33) = 2.74$, $p = 0.010$. Table 3 summarises differences between each intensity level and the control condition or high legibility accuracy point. This table should be interpreted in light of the lack of statistical significance observed in the omnibus test, and is included here for completeness.

Fourier domain vertical filtering

Accuracy did not differ between levels of the vertical filter $F(1, 33) = 0.05$, $p = 0.818$ (see Figure 1). Averaged across intensity levels, response accuracy was significantly greater than accuracy in the control condition $t(33) = 3.57$, $p = 0.001$, but was not significantly different than the high legibility accuracy point $t(33) = 0.67$, $p = 0.509$. This table should be interpreted in light of the lack of statistical significance

Table 3. Results of statistical tests comparing the response accuracies of the horizontally filtered middle layers to the control and high legibility conditions.

Horizontal multiplier	Different from control?	Different from high legibility?
0.2	No [$t = 1.15$, $p = 0.258$]	No [$t = -1.32$, $p = 0.198$]
0.1	No [$t = 1.12$, $p = 0.269$]	No [$t = -1.29$, $p = 0.206$]
0.05	No [$t = -0.39$, $p = 0.702$]	Yes [$t = -3.81$, $p = 0.001$]
0.025	No [$t = 0.76$, $p = 0.453$]	No [$t = -1.81$, $p = 0.080$]
0	No [$t = 1.80$, $p = 0.081$]	No [$t = -1.01$, $p = 0.318$]

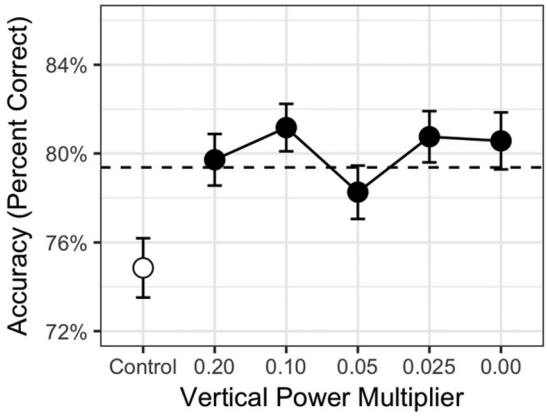

Figure 9. Response accuracies across different levels of the vertical information filter middle layer condition (lower numbers on the *x*-axis indicate greater reductions in vertical information; as in previous plots, the middle layer increases in intensity when read from left to right). Other labelling as in Figure 6.

Table 4. Results of statistical tests comparing the response accuracies of the vertically filtered middle layers to the control and high legibility conditions.

Vertical multiplier	Different from control?	Different from high legibility?
0.2	Yes [$t = 2.40$, $p = 0.022$]	No [$t = 0.22$, $p = 0.830$]
0.1	Yes [$t = 3.70$, $p = 0.001$]	No [$t = 1.17$, $p = 0.251$]
0.05	Yes [$t = 2.09$, $p = 0.045$]	No [$t = -0.82$, $p = 0.419$]
0.025	Yes [$t = 3.15$, $p = 0.003$]	No [$t = 0.88$, $p = 0.384$]
0	Yes [$t = 3.21$, $p = 0.003$]	No [$t = 0.80$, $p = 0.429$]

observed in the omnibus test, and is included here for completeness (Figure 9).

Table 4 summarises differences between each intensity level and the control condition or high legibility accuracy point. This table should be interpreted in light of the lack of statistical significance observed in the omnibus test, and is included here for completeness.

Experiment I discussion

Overall, most of the middle layer manipulations demonstrated a beneficial impact on glance legibility. This extends the findings of previous efforts, notably the work of Gabbard, Swan, and Hix (2006) in using opaque 'billboards'. Our approach uses middle layers which preserve, and therefore reveal, background information. Further, intensity levels were shown, broadly, to have an impact on legibility. Therefore, design decisions that have previously been primarily guided by aesthetic concerns should now be understood to have applied significance for operators; performance in glance-based legibility tasks hinges on both typeface and the intensity of the middle layer. However, our hypothesis held only for the conventional middle layer approaches of Gaussian blur and semi-transparent layering. We will discuss each of the middle layer types and performance at the tested levels below.

Results showed that the semi-transparent layer was ineffective at low intensity (10–20% opacity), but became highly effective at promoting legibility at 30% opacity or greater. Note that the method employed here linearly interpolates the background image with its own mean colour. This is somewhat different from what some interfaces do, which is to interpolate towards black to darken the image or white to lighten it. Figure 10 illustrates the difference between these techniques. At 30% opacity, the white and black layers have a noticeable presence against the background image, whereas the mean colour layer is perceived more as a subtle reduction in contrast. While it may be true that the white and black versions require less opacity to achieve improvements in legibility compared to the mean colour, it is true that mean colour interpolation remains a subtler effect. It is also worth considering that the data suggest that semi-transparent layering operates in a step function, with low levels of opacity having no effect. Once opacity has rapidly increased to a critical point, however, legibility gains are consistent.

Gaussian blurring proved to be the most effective middle layer technique, based on overall response accuracy. Small amounts of blur produced improvements in performance accuracy similar to the high legibility accuracy point. Larger amounts of blur produced performance accuracy significantly greater than this. Why this 'over improvement' occurred is unclear. Blur levels were relatively modest and certainly would not have created viewing conditions as amenable to legibility as the black-on-white condition, which was

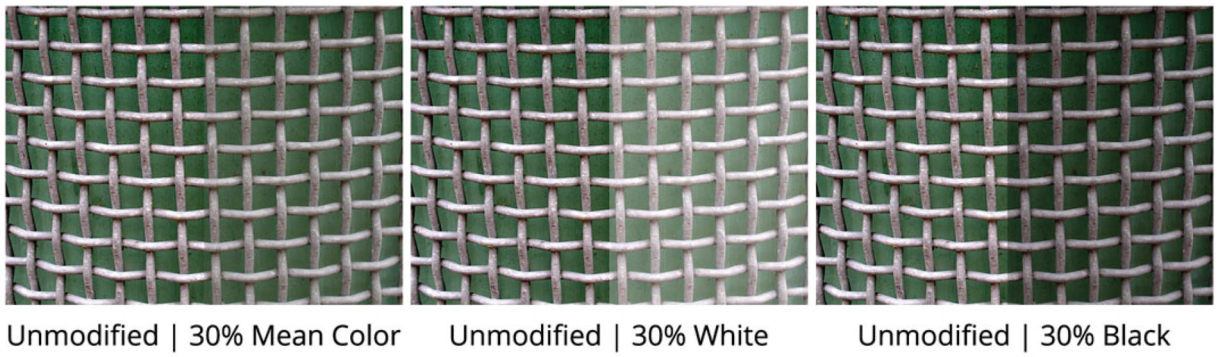

Figure 10. Comparisons of an unmodified background image with semi-transparent layers (30% opacity) of the mean colour, white, and black.

used to determine a stimulus display time threshold corresponding to 79.4% accuracy. Potentially, this is related to better edge detection against blurred middle layers, but could also be related to participants improving at performing the task as the session went on simply as a result of practice with the task.

One limitation that was identified during analysis was that, as threshold assessments were conducted only at the start of the session, it was possible that these thresholds could underestimate later performance gains. To proactively mitigate these effects, which may or may not have existed, in Experiment II and III we presented a second thresholding block, and ignored the values from the first block.

Removing horizontal information from the background images via Fourier domain filtering did not seem to improve foreground text legibility, in opposition to our hypothesis. Conversely, removing vertical information resulted in performance accuracy in line with the high legibility condition, regardless of the intensity of the filter. Figure 11 shows a filtered background that also contains a signature in the lower-right corner ('judith'). Fully filtering the vertical information renders the signature unrecognisable, while fully filtering the horizontal leaves the signature fairly legible. In other words, vertical filtering removes crucial text-like features that would compete with foreground text, but horizontal filtering does not. This is likely why the vertical filter was more effective at improving legibility, while the horizontal filter had little to no effect at all. Notably, the vertical filter does not improve with greater levels of information removal, also in opposition to our hypothesis.

From the present data, Experiment I provides clear design guidance: Gaussian blur and semi-transparent 'scrim' middle layer techniques prove best at maximising legibility. Note that these two techniques are not mutually exclusive, and that the computational requirements to produce each, or a combination, are modest. As such, it may be possible to use a combination of scrim and blur to produce even better results.

Experiment II

Building on the promising results of Experiment I, Experiment II was designed to provide a deeper understanding of whether the interaction between the two most common filtering techniques—semitransparent 'scrim' layers and Gaussian blurring—would be beneficial for legibility. It was further decided to use complex background stimuli similar to those that would be encountered in applied uses of AR displays. One common application of the responsive UX philosophy, particularly in infotainment, is the use the album art of a currently playing song to fill the background. Interface controls and information may then be overlaid over this background. As such, based upon the prior study, it was decided that 5 scrim intensities (0%, 15%, 30%, 45%, and 60%) and 4 blur intensities (SDs of 0, 1, 2, and 3 pixels) would be included for analysis. This created a total of 20 conditions, with the 0% scrim/0 pixel blur condition representing a no-filter control. It was hypothesised that a combination of scrim and blur might allow for greater legibility while retaining more information from the background image.

Experiment II methods

A total of 34 new participants between the ages of 35 and 75 were recruited and subsequently provided written informed consent. There were 18 females (mean age = 57.89) and 16 males (mean age = 57.44) in the sample.

Experiment II utilised the same equipment and software from Experiment I. Complex background stimuli were based on a pool of album art pulled from Billboard Top 40 lists (324 images total), and

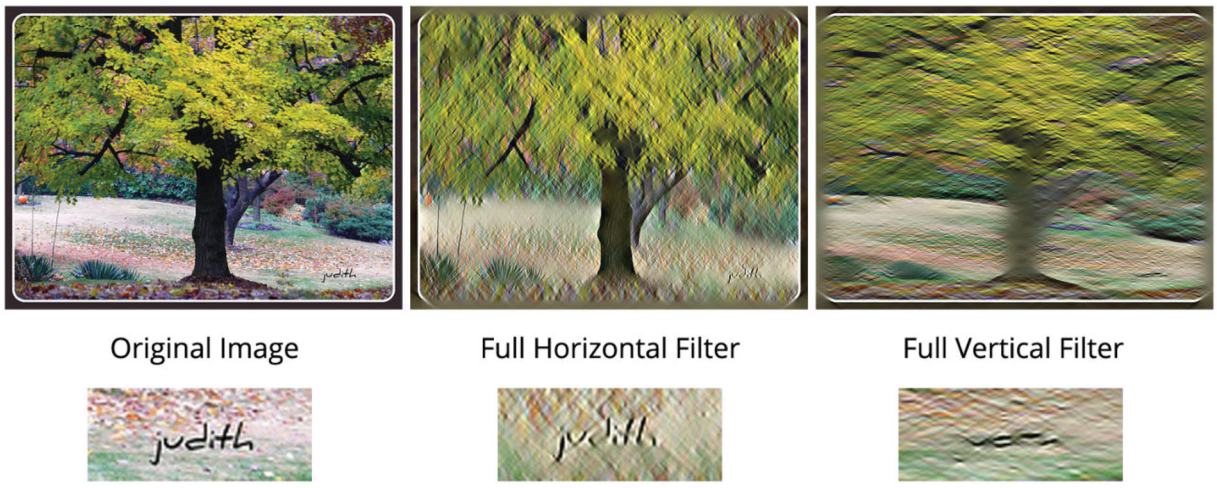

Figure 11. Examples of selectively removing all horizontal or all vertical information from one of the background images used in this study. Bottom row shows differences in preserved detail of a signature in the image's lower-right corner. This example's ImageNet ID: n13912260_7234, for reference.

standardised to a size of 600 × 600 pixels. As in Experiment I, the pool was analysed to extract statistics such as average brightness, centre area brightness, and a measure of visual complexity computed by dividing each image into 64 tiles, and computing median of the SD of pixel intensities for each tile. Images at or below the 10th percentile of brightness, or at or above the 90th percentile of brightness, were dropped from the image pool. Images with central visual complexities at or below the 25th percentile of the pool were also excluded, as were images with minimum complexity values at or below the 10th percentile. This left a total of 168 images in the pool. Each image was modified 20 times, creating versions matching the specified combinations of scrim and blur. Therefore, the underlying base images were repeated across conditions, but not within conditions.

As in Experiment I, each participant completed two 100-trial positive polarity lexical decision threshold assessment blocks. Only the second was used to inform the display time for all subsequent conditions. In the experimental portion, text stimuli were again set in Frutiger, 4 mm, white (#ffffff), and scrims were semi-transparent black (#000000). Response accuracy in each condition was again the primary measure of legibility, with greater accuracy indicating greater legibility under the conditions studied. At the conclusion of the study, participants were debriefed and paid for their time.

Experiment II results

Assessed display time thresholds ranged between 42 ms and 133 ms, and thresholds increased significantly with age $F(1, 32) = 8.82$, $p = 0.006$, by an average of approximately 1 ms per year across the sample. Figure 12 visualises accuracy across all conditions in the study. The dashed lines represent the accuracy level targeted during the threshold calibration stage, and should reflect a theoretical ceiling for the present data. Response accuracy was significantly affected by both blur $\chi^2(1) = 9.10$, $p = 0.003$ and scrim $\chi^2(1) = 39.6$, $p < 0.001$, and the two factors did not significantly interact $\chi^2(1) = 3.29$, $= 0.070$. The non-significant trend towards interaction likely arises because blur exerts a stronger influence on performance when scrim is weak, and vice versa.

Experiment II discussion

Scrim exerted a stronger effect on performance than blur. The effect of blur appears to plateau at 2 pixels (at the 600 pixel image size used), while scrim begins to approach the theoretical calibration point for ideal legibility in this study at 30%. The interaction, in opposition to our hypothesis, was non-significant. While accuracy for 0 blur and 45% and 60% scrim were both nominally higher than for 30% scrim, in order to maximise legibility while retaining the aesthetic and practical advantages of background visibility, the present results simply recommend a scrim of 30%. The combination of scrim of 30% and blur of 3px provide the highest performance in our sample, but at increased loss of background information.

One worthwhile consideration is that the values chosen to test here may not be equivalent between the two manipulations: it is possible that scrim merely utilised a wider range of possible values, while blur

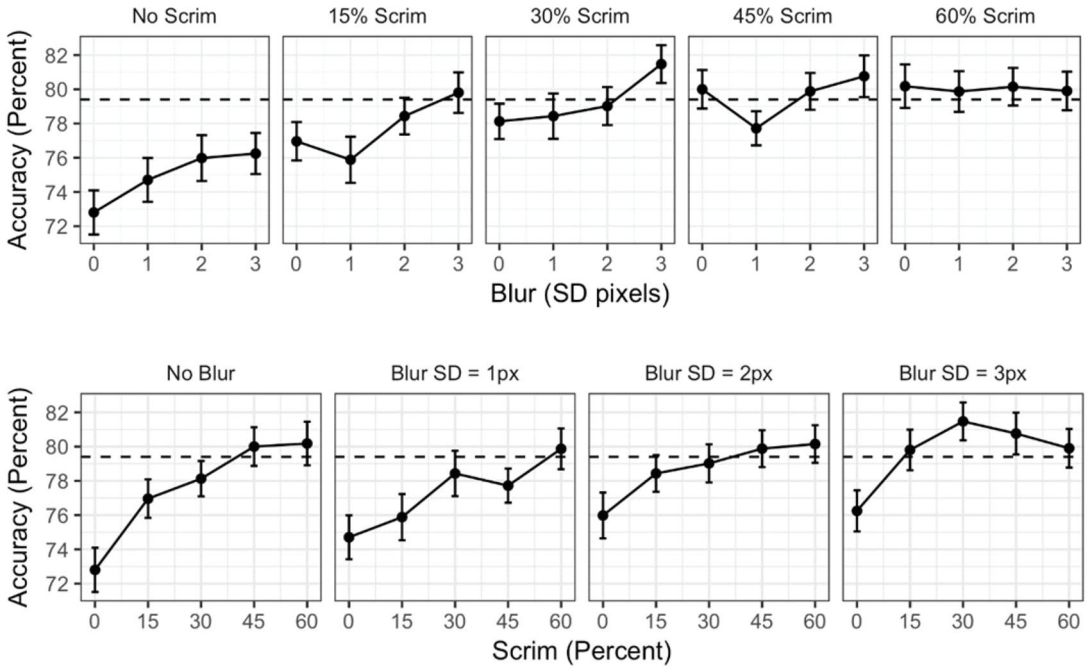

Figure 12. (Top) Blur plotted against accuracy, each panel showing a different intensity of scrim. (Bottom) The same data points, but with scrim intensity plotted against accuracy, and each panel showing a blur intensity. Dashed horizontal line is placed at 79.4% accuracy, corresponding to the theoretical calibration point for ideal legibility in this study.

was more constrained. It is also worth noting that this recommendation only refers to legibility, and does not investigate the question of ability to quickly and accurately obtain information from the background. When initially considering an experiment looking at both legibility and background interpretability aspects, our conversation quickly turned to the possibility of simply shrinking the region of scrim. Could we create a small 'bubble' of scrim around the text and receive similar results? It seemed very likely. It was also considered, upon reflection, that a constrained scrim region might be seen as very similar to drop-shadows, a conventional typeface manipulation. Design literature does do have some insight in this matter, suggesting that a combination of weight (bold typeface), and either outline or drop-shadow, offer legibility enhancement. How would these typeface manipulations hold up, alone and in combination?

Experiment III

Experiment III investigates a logical extension of Experiment II: the use of outline or drop-shadow in front of complex backgrounds. We should first highlight that these techniques have been considered in design contexts (Graham 2012). In terms of typeface weight, previous work has shown a small legibility decrement in the use of bold-weight typefaces on simple

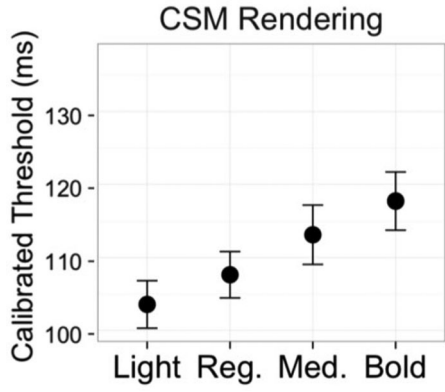

Figure 13. Trends in stimulus display time needed for accurate reading have been reported across typeface weights. Adapted from Dobres et al. (2016).

backgrounds (Dobres et al. 2016, see Figure 13; increased thresholds are worse). On the other hand, one might expect a typeface that hides more of the background to have better visibility and therefore better legibility performance. In order to understand the impact of these typeface manipulations in the absence of a 'middle layer', the present manipulated outlines and drop-shadows, presented in bold and regular typeface. Indeed, we here conceptualise outline and shadow as themselves a letter-attached, localised middle layer. We hypothesised that shadows would provide significant legibility advantages while outlines would

Figure 14. The conditions for the present study include two font weights of Neue Frutiger typeface, modified by three textual manipulations. As in previous studies, text is presented in white. In both the outline conditions and shadow conditions the interior of the displayed textual stimuli is white, and the outline or shadow is black. In all cases, textual stimuli were displayed centred on a complex album art background. For illustrative purposes, the examples here show a close-up of the text and manipulation in each upper left-hand corner. In the actual experiment, this close-up was not shown.

create too much visual noise and therefore would negatively affect legibility. This would be in contrast to current industry practice. We further hypothesised that bold typeface would degrade legibility and that it would interact with both outline and shadow to provide more substantial decrements in legibility.

Experiment III methods

A total of 48 new participants between the ages of 35 and 75 were recruited and provided informed consent. Of these, two were removed for exhibiting extreme thresholds, one failed to calibrate to threshold, one was removed for poor reaction times, one reported retinal scarring, one was interrupted by a fire drill, one reported being a non-native speaker after the completion of the experiment, and one was removed for non-completion. This left a total of 40 participants in the analysis sample, 20 female (mean age = 58.8) and 20 male (mean age = 56.25). Each provided written informed consent prior to participating in the experiment.

Complex background stimuli in Experiment III consisted of the same album art used in Experiment II. Each participant underwent two initial blocks of threshold assessment, with textual stimuli set in Frutiger at a 4 mm capital letter height in white (hex: #ffffff) against a plain black background (hex: #000000) and the threshold from the second run was used as the display duration for all subsequent conditions. Response accuracy in each condition was again the primary measure of legibility, with greater accuracy indicating greater legibility under the conditions studied. The task was again lexical decision, but in Experiment III text stimuli were set in either Neue Frutiger Regular or Neue Frutiger Bold, at a 4 mm capital letter height, in white (#ffffff). Our textual manipulation conditions, outline and drop-shadow, were displayed at default width in black (#000000), modifying the white characters. At the conclusion of the experimental portion of the study, participants were debriefed and paid for their time (Figure 14).

Experiment III results

Among the remaining age and gender balanced sample, assessed display time thresholds ranged between 33 ms and 283 ms across participants. Neither age nor gender had significant impact upon thresholds, $F(1, 39) = 8.82$, $p = 0.006$. The present analysis assessed accuracy of two typeface weights (regular, bold) x three typeface manipulations (none, outline, shadow) using a repeated-measures analysis of variance (ANOVA). Multivariate tests revealed a main effect of typeface weight Wilk's $\lambda = .62$, $F(1, 9) = 5.59$, $p = 0.04$. A main effect of typeface manipulation was found Wilk's $\lambda = 0.06$, $F(2, 8) = 0.06$, $p < 0.01$. These main effects are best interpreted in light of a significant interaction between typeface weight and typeface manipulation Wilk's $\lambda = 0.41$, $F(2, 8) = 5.76$, $p < .01$. The interaction describes the distribution of the typeface weight effect, such that bold typeface strongly outperforms regular typeface under the 'outline' manipulation, but makes no significant impact under either 'no manipulation' or the 'shadow' manipulation (Figure 15).

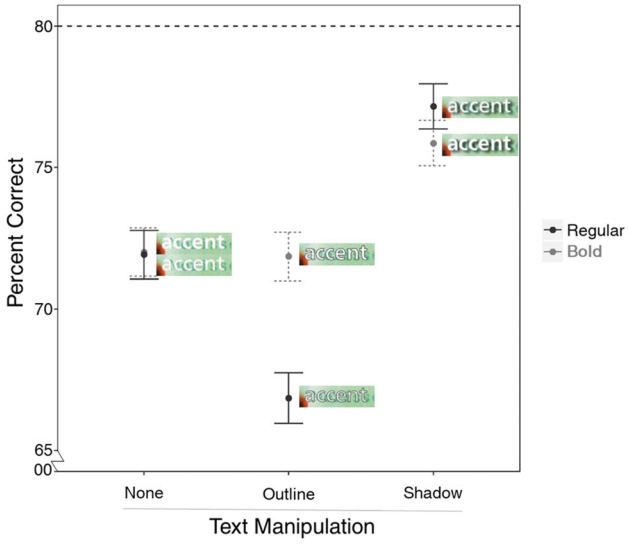

Figure 15. Two font weights (regular, bold) plotted against the three textual manipulations (none, outline, shadow). Percent correct here is a measure of legibility. Bold typeface strongly outperforms regular weight under the 'outline' manipulation, but makes no significant impact under either no manipulation or the shadow manipulation. Indeed, the clear design recommendation is, in situations where performance matters, to use shadow to increased legibility of text on complex backgrounds. The dashed horizontal line at 80% accuracy represents a theoretical calibration point for ideal legibility in this study, representing a 'ceiling' for the present data.

Experiment III discussion

The pattern at hand clearly shows shadow to have been a superior typeface manipulation for improving legibility of the Neue Frutiger typeface used in the present experiment. Outline, conversely, and in agreement with our hypothesis, was, at best, only as legible as non-manipulated text. Indeed, outlined text presented in regular Neue Frutiger typeface was the worst performing condition in the entire experiment. Our hypothesis regarding an interaction between typeface manipulation and weight was only partially upheld. Typeface without manipulation was not strongly impacted by typeface weight. Previous work (Dobres et al. 2016) had shown a small effect (see Figure 13). When combined with outline, greater weight greatly improved performance in line with our hypothesis, albeit only in the face of the substantial decrements imposed by the outline itself. When combined with shadow, greater weight in fact resulted in a small trend towards performance decrement.

These data carry clear design implications for practitioners displaying text over complex backgrounds. In such situations, legibility can be best achieved through the use of drop-shadows, rather than outlines. When using drop-shadows, bolding exerts such a small decrement effect as to be considered unimportant in terms of performance. If, perhaps for aesthetic reasons, outline is preferred, it is important to use a bolded typeface to ameliorate some, but not all, of the losses in legibility. Finally, on complex backgrounds, text presented in a bold typeface will not be significantly more legible. As an aesthetic decision this may be advisable, but it will not boost performance.

Discussion

Our data from Experiments I, II, and III provide a window into legibility considerations related to the display of typeface over complex backgrounds, as they exist in AR applications. They additionally speak to the efficacy of the various 'middle layer' and typographical manipulation approaches designers deploy in these circumstances. It is worth noting that our results, overlaying text on static images, may not generalise to dynamic backgrounds (as exist in many real-world AR applications, such as vehicular HUDs and HMDs). Future work should investigate this question. Natural environments can be more complex than the stimuli we used as backgrounds (e.g. road scenes, which themselves vary considerably in their contents as a function of environment, with urban scenes being more visual complex than highway scenes). Our album

covers, specifically, included text and other symbology. But, of course, so do many naturalistic scenes. Indeed, stimuli representative of text or symbology presented over video or 3-D environments, in head-up displays (HUDs), or over dynamic graphical user interface (GUI) may themselves show differences, and human factors and engineering psychology centred efforts to determine these differences would be a wise investment for any group strongly invested in a particular context. For example, as dynamic interfaces enter the vehicle and other time critical contexts, it will be important to understand how dynamic background elements and the middle layers that mediate them affect legibility at a glance. This work should be done with an eye not only to the basic vision science, but also with an applied understanding of how various factors may holistically impact human performance beyond the legibility task itself in real-world contexts.

More specifically, it remains an open question as to why shadow, in general, has such a strong impact compared to outline. One possibility has to do with an underlying cause of poor legibility of text over complex backgrounds: each letter contains within it information encoded in an outline, and outlines in the background compete with this information. Adding an outline to the text itself may only add interference, by increasing the complexity of the overall discrimination task. A drop-shadow, on the other hand, selectively places a semitransparent layer between the letter and the background. The efficacy of this approach, notably with a much larger semitransparent layer, can be seen Experiments I and II. Such a scrim may be efficacious even when it is very small, attached only to part of a letter. Indeed, a new and interesting question arises from this line of thought: how small can a scrim, or a drop-shadow, be and yet remain effective? The advantage of smaller scrims is in retaining situational awareness of the background is also worth a deeper investigation, on the basis that less coverage of background elements distorts less environmental information.

For that matter, based on our results in Experiments I and II, we believe there is substantially more work to be done in understanding why Gaussian blur performed so well. Necessarily, this includes questions of how the semi-transparent layer technique used in the present work (blending towards the image's mean colour) compares to the more common techniques of blending towards white or black. Opportunities in the Fourier domain are especially intriguing, and we readily acknowledged that the filters developed for Experiment 1 are relatively crude. There are undoubtedly more refined ways of controlling the brightness and contrast of the filtered images, and more targeted ways of filtering out background information that competes with foreground text. For example, consider the possibilities of filtering for text-like spatial frequencies in combination with orientation, rather than orientation alone. Indeed, Gaussian blur may perform so well simply because, in the Fourier domain, the manipulation better removes power associated with text. For a similar reason, choosing intensity levels for each middle layer technique proved challenging. There is little actionable guidance for these types of information spaces, and therein lies a research opportunity. There is, of course, a considerable body of research on the many interactions between contrast and visual perception, but little of it is directly applicable here. There is also some recent work on the effects of additive blur and noise on text legibility (Wolfe et al. 2016), but this focuses on the foreground, not the background, our focus in this study. This, then, brings up a particular limitation of Experiments I & II: it is difficult to compare scrim and blur effects precisely because it is difficult to quantify whether one (likely scrim) simply used a wider range of values, relative to some unknown perceptual baseline, than the other. Indeed, quantifying this question alludes to the larger and ongoing questions of fundamental understanding of human sensation and subsequent perception.

A number of fascinating questions and potential applications remain uninvestigated in this work. Further research using the approaches of this paper, might provide interesting information regarding which typographic components facilitate visual acquisition of textual information, both on a letter basis and in aggregate. We especially remain intrigued by the possibility of a more optimal, if less traditional, middle ground between middle layers as employed in Experiment I and III, and dropshadows as employed in Experiment III. For example, one might imagine shifting the spatial extent of the middle layer modification in some fashion, potentially in multiple levels (e.g. X pixels/0.X° visual angle from the centre of the word), while considering word orientation. It would be interesting to examine the influence of spatial extent of the filtered middle layer region around the word on legibility, which would allow us to determine a trade-off curve between enhancing legibility and obscuring the background, and such future work might also include an investigation of text polarity in this context. Beyond such hybrids, digital typefaces often have default variants for common typographic

manipulations, but whether these presets are optimal for glanceable legibility in unknown and likely amenable to optimisation on a per-typeface basis or in a more generalisable form. The outline, shadow, and bold variants we test in Experiment III indeed are unlikely to be the 'optimal' variants that would result from such work. There are also extensions in terms of the vision science component of the present work that we feel would be fruitful. Depth cues, which may be generated intentionally in many mixed reality systems or unintentionally, for example in windshield HUDs, may provide affordances to legibility. Indeed, many of the examples in this paper include such a depth component. Where depth and switching between layers occurs, and implicit question of divided attention arises which we have alluded to in this work, but here call our specifically and a direction for additional research. We suggest this be investigated broadly, as there is evidence that depth information may enhance salience in dimensions beyond legibility (see Greenlee et al. 2015). Further, we suspect that some glanceable reading may be possible using using peripheral vision, and so the questions of central and peripheral vision should be considered, both as their own interesting direction and in conjunction with questions of depth (and see Wolfe et al. 2019, for evidence that peripheral vision can support other visual tasks, even under conditions of distraction).

Our present results are informative in and of themselves, and reveal several interesting relationships between text, background, and middle layer modifications. These findings have direct relevance to graphic design for augmented reality applications. Indeed, they provide a foundation for broader understanding of design trade-offs and best practices in glanceable reading over complex backgrounds. The future of reading at speed, and of displaying type in-line with rich environments, presently appears to be ever-growing, and in fact we look with the present results towards a time when reading on the static page is a far more niche activity. In considering this future, it is important to reflect that writing has been with humans for at least 9000 years (Li et al. 2003). Earthen walls, clay tablets, bound hides, and stone are each a more complex background than the sterile canvases of today. The 500 year old art of typography, therefore, would seem to have resilient roots. Writing, as a technology, may prove surprisingly robust to a future of being displayed over photos, live scenery, virtual worlds, and scenes presently unimaginable. Designers and researchers facing those challenges to come will need to rethink the basic tenants of design, interface, and moving information from screen to mind. We here provide a first step towards exploring and understanding this increasingly common and practical design problem, the balance between removing background information and providing foreground legibility.

Acknowledgments

Support for this publication was provided through the Clear Information Presentation (Clear-IP) consortium at MIT. Clear-IP is a collaborative effort focussed on the study of legibility and related topics, primarily supported to date by Monotype Imaging and Google. Studies I and II were supported by Clear-IP and study III was independently supported by Monotype. The views and conclusions being expressed as those of the authors and may not necessarily represent those of individual sponsoring organisations.

Disclosure statement

Nadine Chahine was employed by Monotype at the time portions of this work were performed.

Ben D. Sawyer was employed by The Massachusetts Institute of Technology at the time portions of this work were performed.

References

Caird, J. K., K. A. Johnston, C. R. Willness, M. Asbridge, and P. Steel. 2014. "A Meta-Analysis of the Effects of Texting on Driving." *Accident Analysis & Prevention* 71: 311–318. doi: 10.1016/j.aap.2014.06.005.

Debernardis, S., M. Fiorentino, M. Gattullo, G. Monno, and A. E. Uva. 2014. "Text Readability in Head-Worn Displays: Color and Style Optimization in Video versus Optical See-Through Devices." *IEEE Transactions on Visualization and Computer Graphics* 20 (1): 125–139. doi:10.1109/TVCG.2013.86.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. "Imagenet: A Large-Scale Hierarchical Image Database." Presented at the In CVPR.

Dobres, J., N. Chahine, and B. Reimer. 2017. "Effects of Ambient Illumination, Contrast Polarity, and Letter Size on Text Legibility under Glance-like Reading." *Applied Ergonomics* 60: 68–73. doi:10.1016/j.apergo.2016.11.001.

Dobres, J., N. Chahine, B. Reimer, D. Gould, B. Mehler, and J. F. Coughlin. 2016. "Utilising Psychophysical Techniques to Investigate the Effects of Age, Typeface Design, Size and Display Polarity on Glance Legibility." *Ergonomics* 59 (10): 1377–1391. doi:10.1080/00140139.2015.1137637.

Dobres, J., B. Wolfe, N. Chahine, and B. Reimer. 2018. "The Effects of Visual Crowding, Text Size, and Positional Uncertainty on Text Legibility at a Glance." *Applied Ergonomics* 70: 240–246. doi:10.1016/j.apergo.2018.03.007.

Field, D. J. 1987. "Relations between the Statistics of Natural Images and the Response Properties of Cortical Cells." *Journal of the Optical Society of America A* 4 (12): 2379–2394. doi:10.1364/JOSAA.4.002379.

Gabbard, J. L., I. J. E. Swan, D. Hix. 2006. "The Effects of Text Drawing Styles, Background Textures, and Natural Lighting on Text Legibility in Outdoor Augmented Reality." *Presence: Teleoperators and Virtual Environments* 15 (1): 16–32. doi:10.1162/pres.2006.15.1.16.

Graham, L. 2012. *Basics of Design: Layout & Typography for Beginners*. Clifton Park, NY: Cengage Learning.

Greenlee, Eric T., Gregory J. Funke, Joel S. Warm, Victor S. Finomore, Robert E. Patterson, Laura E. Barnes, Matthew E. Funke, and Michael A. Vidulich. 2015. "Effects of Stereoscopic Depth on Vigilance Performance and Cerebral Hemodynamics." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57 (6): 1063–1075. doi:10.1177/0018720815572468.

Hancock, P. A., B. D. Sawyer, and S. Stafford. 2015. "The Effects of Display Size on Performance." *Ergonomics* 58 (3): 337–354. doi:10.1080/00140139.2014.973914.

Harding, T. H., J. S. Martin, and C. E. Rash. 2005. "Using a Helmet-Mounted Display Computer Simulation Model to Evaluate the Luminance Requirements for Symbology." In *Presented at the Defense and Security Symposium*, edited by C. E. Rash and C. E. Reese, Vol. 5800, 159. Bellingham: SPIE. doi:10.1117/12.602123.

Harding, T. H., J. S. Martin, and C. E. Rash. 2007. "The Legibility of HMD Symbology as a Function of Background Local Contrast." In *Presented at the Defense and Security Symposium*, edited by R. W. Brown, C. E. Reese, P. L. Marasco, and T. H. Harding, Vol. 6557, 65570D. Bellingham: SPIE. doi:10.1117/12.719657.

Hoffman, D. M., A. R. Girshick, K. Akeley, and M. S. Banks. 2008. "Vergence–Accommodation Conflicts Hinder Visual Performance and Cause Visual Fatigue." *Journal of Vision* 8 (3): 33–33. doi:10.1167/8.3.33.

Huey, E. B. 1908. "The Psychology and Pedagogy of Reading." The Macmillan Company.

ISO. 1993. *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTST). Part 3: Visual Display Requirements*. ISO 9241-3. Geneva: ISO.

Javal, E. 1878. "Essai sur la physiologie de la lecture." *Annales d'Ocilistique* 80: 97–117.

Kress, B., and T. Starner. 2013. "A Review of Head-Mounted Displays (HMD) Technologies and Applications for Consumer Electronics." In *Presented at the SPIE Defense, Security, and Sensing*, edited by A. A. Kazemi, B. C. Kress, and S. Thibault , Vol. 8720, 87200A. Bellingham: SPIE. doi:10.1117/12.2015654.

Legge, Gordon E., David H. Parish, Andrew Luebker, and Lee H. Wurm. 1990. "Psychophysics of Reading. XI. Comparing Color Contrast and Luminance Contrast." *Journal of the Optical Society of America A* 7 (10): 2002–2010. doi:10.1364/JOSAA.7.002002.

Li, X., G. Harbottle, J. Zhang, and C. Wang. 2003. "The Earliest Writing? Sign Use in the Seventh Millennium BC at Jiahu, Henan Province, China." *Antiquity* 77 (295): 31–44. doi:10.1017/S0003598X00061329.

Medler, D. A., and J. R. Binder. 2005. "MCWord: An On-line Orthographic Database of the English Language."

Milgram, Paul, and Fumio Kishino. 1994. "A Taxonomy of Mixed Reality Visual Displays." *IEICE Transactions on Information and Systems* 77 (12): 1321–1329.

Milgram, P., H. Takemura, A. Utsumi, and F. Kishino. 1995. "Augmented Reality: A Class of Displays on the Reality-Virtuality Continuum." In *Telemanipulator and Telepresence Technologies*, edited by M. R. Stein, Vol. 2351, 282–293. Bellingham: International Society for Optics and Photonics.

Neve, J., and A. Jenniskens. 1994. *Low Vision* (G. C. Woo, Ed.). New York, NY: Springer. http://doi.org/10.1007/978-1-4612-4780-7

Peirce, J. W. 2008. "Generating Stimuli for Neuroscience Using PsychoPy." *Frontiers in Neuroinformatics* 2: 1–8. doi:10.3389/neuro.11.010.2008.

Petkov, N., and M. A. Westenberg. 2003. "Suppression of Contour Perception by Band-Limited Noise and Its Relation to Nonclassical Receptive Field Inhibition." *Biological Cybernetics* 88 (3): 236–246. doi:10.1007/s00422-002-0378-2.

Reimer, Bryan, Bruce Mehler, Jonathan Dobres, Joseph F. Coughlin, Steve Matteson, David Gould, Nadine Chahine, and Vladimir Levantovsky. 2014. "Assessing the Impact of Typeface Design in a Text-Rich Automotive User Interface." *Ergonomics* 57 (11): 1643–1658. doi:10.1080/00140139.2014.940000.

Roethlein, B. E. 1912. "The Relative Legibility of Different Faces of Printing Types, Vol. 3. Clark University Press.

Rusch, M. L., M. C. Schall, Jr, P. Gavin, J. D. Lee, J. D. Dawson, S. Vecera, and M. Rizzo. 2013. "Directing Driver Attention with Augmented Reality Cues." *Transportation Research Part F: Traffic Psychology and Behaviour* 16: 127–137. doi:10.1016/j.trf.2012.08.007.

Sanford, E. C. 1888. "The Relative Legibility of the Small Letters." *The American Journal of Psychology* 1 (3): 402–435.

Sawyer, B. D., J. Dobres, N. Chahine, and B. Reimer. 2017. "The Cost of Cool: Typographic Style Legibility in Reading at a Glance." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61 (1): 833–837. doi:10.1177/1541931213601698.

Sawyer, B. D., J. Dobres, N. Chahine, and B. Reimer. 2020. "The Great Typographic Bake-Off: Comparing Legibility at a Glance." *Ergonomics* 63 (4): 391–398.

Sawyer, B. D., V. S. Finomore, A. A. Calvo, and P. A. Hancock. 2014. "Google Glass: A Driver Distraction Cause or Cure?" *Human Factors: The Journal of the Human Factors and Ergonomics Society* 56 (7): 1307–1321. doi:10.1177/0018720814555723.

Skrypchuk, L., P. Langdon, B. D. Sawyer, and P. J. Clarkson. 2019. "Unconstrained Design: improving Multitasking with in-Vehicle Information Systems through Enhanced Situation Awareness." *Theoretical Issues in Ergonomics Science* 21 (2): 183–219. doi:10.1080/1463922X.2019.1680763.

Skrypchuk, L., P. Langdon, B. D. Sawyer, A. Mouzakitis, and P. J. Clarkson. 2019. "Enabling Multitasking by Designing for Situation Awareness within the Vehicle Environment." *Theoretical Issues in Ergonomics Science* 20 (2): 105–128. doi:10.1080/1463922X.2018.1485984.

Wallace, S., R. Treitman, J. Huang, B. D. Sawyer, and Z. Bylinskii. 2020a. "Accelerating Adult Readers with Typeface: A Study of Individual Preferences and Effectiveness. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1–9).

Wallace, S., R. Treitman, N. Kumawat, K. Arpin, J. Huang, B. Sawyer, and Z. Bylinskii. 2020b. "Towards Readability Individuation: The Right Changes to Text Format make Large Impacts on Reading Speed." *Journal of Vision* 20 (10): 17–17.

Weiser, M. 1993. "Ubiquitous Computing." *Computer Magazine*. 26 (10): 71–72. doi:10.1109/2.237456.

Wolfe, B., J. Dobres, A. Kosovicheva, R. Rosenholtz, and B. Reimer. 2016. "Age-Related Differences in the Legibility of Degraded Text." *Cognitive Research: Principles and Implications* 1 (1): 13. doi:10.1186/s41235-016-0023-6.

Wolfe, B., B. D. Sawyer, A. Kosovicheva, B. Reimer, and R. Rosenholtz. 2019. "Detection of Brake Lights While Distracted: Separating Peripheral Vision from Cognitive Load." *Attention, Perception, & Psychophysics* 81 (8): 2798–2716. doi:10.3758/s13414-019-01795-4.