The Cost of Cool: Typographic Style Legibility in Reading at a Glance

Ben D. Sawyer,[1] Jonathan Dobres,[1] Nadine Chahine[2] & Bryan Reimer[1]
[1]Massachusetts Institute of Technology, AgeLab & New England University Transportation Center
[2]Monotype, Inc.

When designers typographically tweak fonts to make an interface look 'cool,' they do so amid a rich design tradition, albeit one that is little-studied in regards to the rapid 'at a glance' reading afforded by many modern electronic displays. Such glanceable reading is routinely performed during human-machine interactions where accessing text competes with attention to crucial operational environments. There, adverse events of significant consequence can materialize in milliseconds. As such, the present study set out to test the lower threshold of time needed to read and process text modified with three common typographic manipulations: letter height, width, and case. Results showed significant penalties for the smaller size. Lowercase and condensed width text also decreased performance, especially when presented at a smaller size. These results have important implications for the types of design decisions commonly faced by interface professionals, and underscore the importance of typographic research into the human performance impact of seemingly "aesthetic" design decisions. The cost of "cool" design may be quite steep in high-risk contexts.

INTRODUCTION

Typography subtends a toolbox of age-old strategies to embellish and accent the written word, but the application of these tools to rapid-update electronic systems that offer information at a glance is poorly understood. Such "glanceable" devices are often used in the course of other tasks, and so competition can arise for common structural and cognitive resources (Sawyer, Finomore, Calvo, & Hancock, 2014). This competition arises in contexts ranging from in-vehicle displays (Mehler, Reimer, Dobres, Foley, & Ebe, 2016; Reimer, Mehler, & Coughlin, 2012; Reimer et al., 2014) to walking while using smartphones (Thompson, Rivara, Ayyagari, & Ebel, 2013). Relatively little work has explored the contribution of typographic manipulations of text upon glanceable legibility, or the ease with which a reader can accurately perceive and encode text in a glance (Slattery & Rayner, 2009). What work does exist suggests that such manipulations are not without trade-offs, and can impact the amount of time needed to read and comprehend presented information.

What, if any, influence does a design decision have upon a reader's ability to rapidly access content? The size of letters has been shown to be a strong factor in legibility, both in glanceable interfaces and traditional reading at length (Bernard, Chaparro, Mills, & Halcomb, 2003; Dobres, Chahine, Reimer, Gould, Mehler, & Coughlin, 2016b; Legge, Pelli, Rubin, & Schleske, 1985). In digital interfaces, the question of size and legibility becomes somewhat more complex. Pixel density represents the limits of any display's ability to reproduce text. A variety of "smoothing" techniques may be used to mitigate this issue, each with potential downsides to legibility. However, beyond certain lower bound thresholds, this complex problem has little effect upon response time and accuracy (Hancock, Sawyer, & Stafford, 2014). There is evidence that at smaller font sizes the impact of other textual manipulations may be exacerbated. Dobres and colleagues (2016a) presented two typefaces, each at 3 mm and 4 mm letter heights. In addition to main effects of letter height, the researchers found that legibility differences between typefaces were greater when presented in smaller letters.

Applied typography in operational settings has been explored in at least two glanceable contexts: signage, and in vehicle displays. Seminal work by Reimer and associates (2014) explored the question of typography in in-vehicle interfaces, comparing menu text set in two typefaces in a dual task driving paradigm. This study showed that typeface affected both the response time and accuracy of menu selection tasks. The legibility of roadway signage has been addressed by work which birthed an entire typeface, Clearview, custom-developed to enhance roadway readability (Chrysler, Carlson, & Hawkins, 2002; Holick, Chrysler, Park, & Carlson, 2006). While there is ongoing debate as to Clearview's efficacy, the balance of evidence suggests it to be effective, especially in comparison to precursors (for a spirited argument, (see Dobres, Chrysler, Wolfe, Chahine, & Reimer, 2017b)). New operational settings are arriving daily. Studies that might have seemed fantastical only years ago, such as an investigation of the dual-task costs of climbing while using a head-mounted display (Woodham, Billinghurst, & Helton, 2016), are now run using easily obtainable commercial and military products (Sawyer et al., 2014). Work to understand best practices in typography, in terms of augmenting human performance, has arrived as a practical necessity.

Figure 1. The three conditions tested were height, width, and case. All possible combinations were tested, resulting in 8 conditions (2 heights × 2 widths × 2 cases). All stimuli were presented in the Frutiger typeface family.

Designers regularly manipulate the size, case and width of text (Figure 1) in response to aesthetic demands and the constraints of available space. What pleases a designer's eye may not, however, result in optimal human performance. This 'cost of cool' is foreshadowed by previous research. Literature on traditional reading indicated that text case would impact

glanceable legibility. Uppercase letters have been shown to be more legible partially due to their increased size, but this appears to come at the expense of greater letter confusion (Bouma, 1970; Pelli et al., 2007). Likewise, the effect of text width is supported by traditional reading research showing that crowding is one determinant of reading rate (Pelli et al., 2007).

The bulk of the above glanceable interface work has measured the minimum amount of time needed to read a single word with 80% accuracy during a "lexical decision" task (Meyer & Schvaneveldt, 1971). This psychophysical technique has shown good sensitivity to a variety of typographic conditions, and has been extended to investigations in a number of languages (Dobres, Chahine, & Reimer, 2017a; Dobres, Chahine, Reimer, Gould, & Zhao, 2016a) As such, it seems an appropriate tool for the present investigation, to focus upon the legibility impact of manipulations of text height, width, and case.

Based upon the pattern of results in previous literature, we forwarded four hypotheses regarding legibility, a construct we here have grounded in the above lexical decision task and associated reading time thresholds. Typographic combinations with better legibility were expected to require less time, and so lower thresholds, for accurate reading. We hypothesized that a) uppercase letters would provide greater legibility. Likewise, we hypothesize that b) non-condensed text would provide greater legibility, as compared to condensed width text. As demonstrated in a number of studies noted above, we expected c) that larger text would provide greater legibility than smaller text. Finally, we expected that smaller text would lead to greater decrements for both d) lowercase and e) condensed typeface.

METHOD

Participants

An age-diverse sample of thirty-one participants, fourteen males (M = 55.7 yrs, SD = 11.7 yrs), and seventeen females (M = 54.8 yrs, SD = 8.9 yrs), all between the ages of 31 and 72, were recruited from the greater Boston area. Participants were required to speak and understand English as their native language. We screened for self-reported good health, rejecting participants with neurological or cognitive impairment such as Parkinson's, Alzheimer's, dementia, or psychiatric illnesses, as well as physical impairment including cardiac disease, diabetes, or hospitalization in the past six months. Visual acuity was tested on site with the Snellen eye chart, and the vision of all participants was at least 20/40. Participants were compensated $40.00 for time involvement of up to 90 minutes. All participants provided informed consent, consistent with guidelines set forth by the Massachusetts Institute of Technology's Institutional Review Board.

Apparatus & Stimuli

Data were collected in a quiet room with dim illumination. Stimuli was presented using a Mac Mini running Mac OS (2.5Ghz Intel Core i5 CPU, 4GB of RAM) on a high resolution Acer monitor (21.77" × 12.24", 2560 × 1440 pixels, 60Hz refresh rate). Participants viewed this screen from a distance of approximately 70cm. No head restraints were used, though participants were reminded to attempt to maintain the intended viewing distance.

Stimuli for the experiment were displayed in all possible combinations of the two heights, two widths, and two cases, for a total of eight categories of stimuli. The lexical decision task used (see Fig. 2) involved a forced choice in which a simple yes or no decision was made as to whether a text string was either a word or a pseudoword (i.e., a text string which was pronounceable but not a word, such as "mindle"). First, a fixation rectangle would appear on screen for 1000 ms, followed by a 200 ms mask. Thereafter, the stimuli were presented with variable timing, followed by another 200 ms mask and a 5-second response interval. Participants indicated word or pseudoword by pressing one of two keys on the computer's numeric keypad.

The variable timing of the stimuli resulted from the use of an adaptive staircase procedure to control lexical decision task difficulty, adjusting according to the responses of the participant. The experiment followed a "one up, three down" rule, in which three consecutive correct responses would result in reduced duration, whereas a single incorrect response would result in increased duration (Leek, 2001; Levitt, 1971). This rule ensured staircase convergence upon the stimulus duration corresponding to approximately 80% accuracy, a number used in previous glanceable legibility efforts (Dobres, Chahine, & Reimer, 2017a; Dobres, Chahine, Reimer, Gould, & Zhao, 2016a; Dobres, Chahine, Reimer, Gould, Mehler, & Coughlin, 2016b; Dobres, Chrysler, Wolfe, Chahine, & Reimer, 2017b; Dobres, Reimer, Mehler, Chahine, & Gould, 2014) Regardless of the total number of correct or incorrect responses, stimulus duration did not drop below 16.7 ms, the lower limit of the screen, nor exceed 1000 ms.

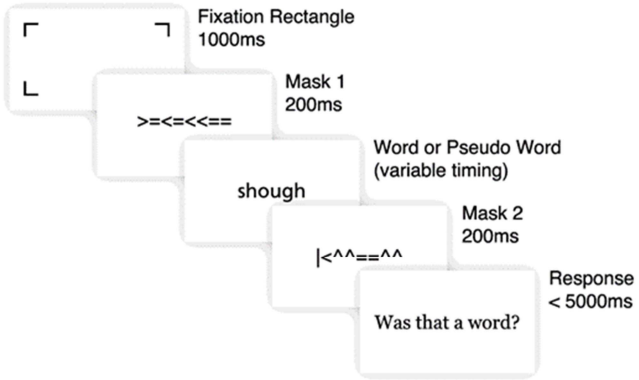

Figure 2. In the lexical decision task stimuli were presented between two masks. Participants were asked to determine if each stimulus was a word or 'pseudoword', a pronounceable non-word. Task difficulty was controlled by increasing or reducing the duration during which the stimulus was presented on screen.

Procedure

Before arriving, participants completed a basic demographic survey. Upon arrival, participants completed informed consent, were allowed to ask any questions, and were administered a vision test, as described above. They were then seated in front of the lexical decision task, where they were permitted to use eyeglasses or other forms of optical correction they felt appropriate. Participants initially received

a series of practice trials, in which stimulus duration was set to 1000 ms, and were only allowed to move forward to the main experiment after five consecutive correct responses.

In the main experiment, every condition was presented in a block of 100 consecutive trials. Condition order was randomized per participant to counteract habituation and learning effects. Simulus duration was initiated at 800 ms for three trials, before being reduced to 600 ms for three additional trials, and finally 200 ms for three final trials. This nine trial "controlled descent" was followed by staircase adaptation control per the rules described above in "Apparatus and Stimuli". Upon completion of the experimental portion, participants completed a brief post study questionnaire, were debriefed, and were paid for their participation.

Analysis

The present analysis focuses upon stimulus duration thresholds. Percent correct, controlled by the staircase design and therefore equal across conditions, is therefore not analyzed. Stimulus duration thresholds for every participant were calculated as the median of the final 20 trials of each condition. The study was analyzed as a 2 height (3mm, 4mm) × 2 width (uncondensed, condensed) × 2 case (uppercase, lowercase) ANOVA. Additionally, a Friedman rank sum test was performed upon the stimuli order to assure adequate counterbalancing.

RESULTS

The non-parametric Friedman rank sum test of differences among counterbalanced orders returned $\chi^2 = 9.25$, which was non-significant ($p = .24$), indicating adequate counterbalancing.

The three-way ANOVA indicated a significant main effects of case, $F(1, 30) = 95.24$, $p < 0.01$, such that uppercase stimuli were detected at a lower calibrated threshold. These data suggest that, as suggested in our first hypothesis, (a) uppercase typeface is in general more legible in glanceable reading. There was also a significant main effect of width, $F(1, 30) = 6.77$, $p = 0.01$, such that uncondensed stimuli were detected at a lower calibrated threshold. This suggests that, in congruence with our second hypothesis, (b) condensed typeface is in general less legible in glanceable reading. There was also a significant main effect of size, $F(1, 30) = 35.85$, $p < 0.01$, such that the larger, 4 mm stimuli were detected at a lower calibrated threshold. This upholds our third hypotheses (c), as well as previous findings (Dobres, Chahine, Reimer, Gould, Mehler, & Coughlin, 2016b) that smaller typeface is less legible in glanceable reading.

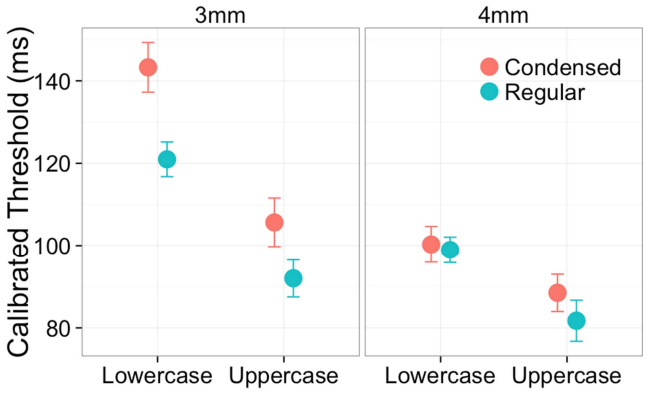

Figure 3. Significant main effects of case, width, and size are joined by a significant interaction between case and size such that differences between upper and lowercase typeface stimuli were greater in the smaller 3 mm typeface. Bars represent within-subject standard error.

The three-way ANOVA further indicated a significant interaction between case and size (see Fig. 3), $F(1, 30) = 6.79$, $p = 0.01$, such that lowercase stimuli were detected at a much higher calibrated threshold relative to uppercase stimuli when presented at the smaller 3 mm size. This upholds our fourth hypothesis (d), and suggests that the effects of case are exacerbated when typeface size is smaller, or conversely that the larger uppercase lettering may have helped to reduce some of the differences in letter height. No further interactions were found to be significant, and no significant pattern was seen for our final hypothesis (e).

Two of the conditions in the present study—the lowercase, regular width fonts at 3mm and 4mm letter heights—are identical to conditions tested in an earlier study (Dobres, Chahine, Reimer, Gould, Mehler, & Coughlin, 2016b). Statistical comparisons between the thresholds measured in the present study and the earlier one indicate no significant differences between threshold assessments ($t(57) = 1.23$, $p = .22$ and $t(56) = 0.91$, $p = .37$ for comparisons between the 3mm and 4mm samples, respectively). This suggests that the methodology described here maintains acceptable validity, even across different (though demographically similar) participant samples and hardware/software setups.

DISCUSSION

The present pattern of results can be seen as strong evidence that typographic manipulation is a source of variability in human performance related to glanceable reading. Specifically, these data uphold our first hypothesis, replicating past efforts (Dobres, Chahine, Reimer, Gould, Mehler, & Coughlin, 2016b), and showing size to be one such contributing factor. These data also support our second and third hypotheses, showing for the first time that case and text width are contributing factors to human performance in glanceable reading. Consider that the differences in thresholds found for lowercase lettering mean that lowercase text required 26% more time for accurate reading, while condensed text required 11.2% more time. Statistical significance aside,

these manipulations resulted in effects of a size likely to be impactful to real-world operational environments.

The significant interaction between typeface size and case shows that design factors can influence one another, as in this case the taller capital letters assisted readers in quickly distinguishing stimuli presented at a smaller typeface size. Consider, for example, an interface where the user may enlarge or reduce the text size. The present data suggest that this seemingly minor degree of control could have unintended consequences. Reduced lettering height could lead to interactions with other design decisions, and surprising emergent issues. More generally, the customizability of operational interfaces is called into question, as end-users may choose combination they consider aesthetically pleasing, but are ultimately operationally detrimental. Dual use devices, such as smartphones, might revert to special high-legibility display modes in operational settings, in order to preserve both choice and performance.

At the same time, knowledge of these trade-offs provides a pathway to solutions. For example, while 3mm text generally had inferior legibility compared to 4mm text, it is notable that 3mm uppercase had very similar reading time thresholds compared to 4mm lowercase text. A close visual inspection of the text samples shows that, coincidentally for the Frutiger typeface family, most prominent lowercase text elements are roughly ¾ the size of their uppercase counterparts. In other words, 3mm uppercase text has a similar visual size to 4mm lowercase text. We do not suggest that smaller uppercase text has absolutely equivalent legibility compared to larger lowercase text (uppercase letters are more prone to crowding and letter confusion), but setting smaller text in uppercase may help ameliorate some of the degradation in legibility.

Indeed, the present pattern of results when considered alongside previous efforts investigating glanceable legibility, seems strongly to indicate the need for a science of *managing consequence* in design decisions. Especially in time constrained operational settings, it seems likely that certain dimensions of freedom in interface design exact a cost. While it may be desirable to end-users and front-line designers to produce eye-catching, configurable, "cool" interfaces, this goal must be considered alongside the need to deliver information to operators engaged in demanding tasks with as little interference as possible. Therefore, there is a need to better understand the etiology of typographic tools used to embellish and accent the written word in the context of costs and benefits in glanceable interface.

The data collected here are an excellent starting point, but are not without limitations. This work was conducted in laboratory conditions, whereas the applied settings we hope to address exist in a far noisier "real world". While we hope to generalize to operational settings, the present results were acquired in a single task protocol. It is extremely likely that performing these tasks in a dual task paradigm would significantly increase calibrated thresholds, and exacerbate differences between conditions. Certainly, ad hoc real-world experiments of this kind are being carried out every day by users of thousands of in-hand and in-vehicle devices experiencing every conceivable operational setting. There is, therefore, a strong need for tools to study text discrimination outside of the laboratory and in complex multitask settings.

The goal, implicit in both this discussion and the data at hand, is to provide designers with an understanding of the performance trade-offs inherent in the aesthetic choices they make. Certainly, there are many products that are simultaneously beautiful to look at and a burden to use in a timely manner. As an increasing amount of reading happens in glanceable interface, and the prevalence of such glanceable interface in complex operational tasks continues to grow, it is imperative that designers have tools to understand how to produce interface that is beautiful, functional, and promotes optimal human performance. Interface you cannot take your eyes off is unquestionably suboptimal in some settings. In the end, designers and users alike want interface optimal to the task at hand. With the right tools and understanding, designers will certainly find ways to work within the degrees of freedom they have to produce interface that is both cool to look at, and cool to use.

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 Andrew Sipperly, and Anthony Pettinato in collecting the experimental data for this work.

REFERENCES

Bernard, M. L., Chaparro, B. S., Mills, M. M., & Halcomb, C. G. (2003). Comparing the effects of text size and format on the readibility of computer-displayed Times New Roman and Arial text. *International Journal of Human-Computer Studies*, *59*(6), 823–835. http://doi.org/10.1016/S1071-5819(03)00121-6

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

Chrysler, S. T., Carlson, P. J., & Hawkins, H. G. (2002). *Nighttime Legibility of Ground-Mounted Traffic Signs As a Function of Font, Color, and Retroreflective Sheeting Type* (No. FHWA/TX-03/1796-2) (pp. 1–76). College Station, TX: Texas Transportation Institute.

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

Dobres, J., Chahine, N., Reimer, B., Gould, D., & Zhao, N. (2016a). The effects of Chinese typeface design, stroke weight, and contrast polarity on glance based legibility. *Displays*, *41*(C), 42–49. http://doi.org/10.1016/j.displa.2015.12.001

Dobres, J., Chahine, N., Reimer, B., Gould, D., Mehler, B., & Coughlin, J. F. (2016b). 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. *Transportation Research Record: Journal of the Transportation Research Board*, *2624*, 1–8. http://doi.org/10.3141/2624-01

Dobres, J., Reimer, B., Mehler, B., Chahine, N., & Gould, D. (2014). A Pilot Study Measuring the Relative Legibility of Five Simplified Chinese Typefaces Using Psychophysical Methods. *the 6th International Conference* (pp. 1–5). Seattle, WA: ACM.

http://doi.org/10.1145/2667317.2667339
Hancock, P. A., Sawyer, B. D., & Stafford, S. (2014). The effects of display size on performance. *Ergonomics*, *58*(3), 337–354. http://doi.org/10.1080/00140139.2014.973914
Holick, A. J., Chrysler, S. T., Park, E. S., & Carlson, P. J. (2006). *Evaluation of the Clearview Font for Negative Contrast Traffic Signs* (No. FHWA/TX-06/0-4984-1) (pp. 1–130). College Station, TX: Texas Transportation Institute.
Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, *63*(8), 1279–1292.
Legge, G. E., Pelli, D. G., Rubin, G. S., & Schleske, M. M. (1985). Psychophysics of reading--I. Normal vision. *Vision Research*, *25*(2), 239–252.
Levitt, H. (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477. http://doi.org/doi:10.1121/1.1912375
Mehler, B., Reimer, B., Dobres, J., Foley, J., & Ebe, K. (2016). Additional Findings on the Multi-Modal Demands of "Voice-Command" Interfaces (Vol. 1, pp. 2016–01–1428–20). Presented at the SAE 2005 World Congress & Exhibition, 400 Commonwealth Drive, Warrendale, PA, United States: SAE International. http://doi.org/10.4271/2016-01-1428
Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234.
Pelli, D. G., Tillman, K. A., Freeman, J., Su, M., Berger, T. D., & Majaj, N. J. (2007). Crowding and eccentricity determine reading rate. *Journal of Vision*, *7*(2), 1–36. http://doi.org/10.1167/7.2.20
Reimer, B., Mehler, B., & Coughlin, J. F. (2012). *An Evaluation of Typeface Design in a Text-Rich Automotive User Interface* (No. 2012-12) (pp. 1–36). Cambridge.
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
Sawyer, B. D., Finomore, V. S., Calvo, A. A., & Hancock, P. A. (2014). Google Glass. *Human Factors*, *56*(7), 1307–1321. http://doi.org/10.1177/0018720814555723
Slattery, T. J., & Rayner, K. (2009). The influence of text legibility on eye movements during reading. *Applied Cognitive Psychology*, *24*(8), 1129–1148. http://doi.org/10.1002/acp.1623
Thompson, L. L., Rivara, F. P., Ayyagari, R. C., & Ebel, B. E. (2013). Impact of social and technological distraction on pedestrian crossing behaviour: an observational study. *Injury Prevention*, *19*(4), 232–237. http://doi.org/10.1136/injuryprev-2012-040601
Woodham, A., Billinghurst, M., & Helton, W. S. (2016). Climbing With a Head-Mounted Display. *Human Factors*, *58*(3), 452–461. http://doi.org/10.1177/0018720815623431