White Paper (2018-1)

TypeTester: A Case Study of Behavioral Data Collection Using a Smartphone Platform

Jonathan Dobres, Karola Klarl, Julia Kindelsberger, & Bryan Reimer

Abstract: Interest in leveraging smartphone technology for scientific data collection has increased significantly in recent years. Mobile platforms have now been employed to investigate a variety of physiological and behavioral phenomena. Here we add to this rapidly growing body of work, using a specially designed mobile application to collect data on text legibility using a paradigm that mirrors established laboratory methods. Lexical decision data (an established proxy for text legibility) were collected from a smartphone platform over the course of several months, ultimately resulting in a sample of trials equivalent to a moderately sized lab experiment. Two typefaces (Frutiger and Eurostile) were tested in both positive and negative polarities. Results suggest that the participant sample was highly motivated and willing to participate in periodic task probes during an engagement period of 1-2 weeks. Consistent with previous work, positive polarity text was read more easily than negative polarity, and response accuracy rose with display duration. However, no significant effects were observed for typeface (i.e., comparing Eurostile to Frutiger) under the testing conditions. Although the application successfully collected activity state and illumination for a majority of trials, sampling rates were insufficient to make comparisons along these dimensions. There was also some concern that the application framework may not present stimuli with reliable timing. Given the relatively small sample size of this initial investigation and the uncontrolled experiment setting, the results suggest that there is substantive potential for this approach as a viable platform for experimental data collection. The trade-offs inherent in a mobile data collection are substantial, and are discussed in detail for this project.

Background

Legibility and the Limits of the Lab

Legibility has long been of paramount importance to experimental psychology and human factors research, and as more reading is done at a glance from smartphone screens, the particular effects of this mode of reading have come under scientific scrutiny. Reading at a glance, in contrast to longer-form reading, may well have different

ergonomic properties, and makes different demands of the reader, especially if the reading is being done secondarily to some other task (as when reading from an in-vehicle information system, glancing at a wearable device while exercising, or checking a message on a smartphone while walking to a meeting).

In recent years, the AgeLab has undertaken a number of studies designed to examine reading at a glance. The earliest of these looked at reading in a multi-tasking environment (Reimer et al., 2014). Participants drove a driving simulator while also completing a simple menu selection task, thus necessitating that menus be read in short glances. Menu items were set in either the Frutiger or Eurostile typefaces, controlling for optical size. It was thought by typographic experts that the Frutiger typeface, with its more open letter spacing and more varied character shapes, would be more legible at a glance than the rigid and geometric Eurostile. Indeed, the results showed that menus set in Frutiger were read more quickly and more accurately than those set in Eurostile. The AgeLab has since followed up this work with studies on a variety of typographic issues. Eschewing the complexities and overhead of the driving simulator, legibility has been probed using laboratory-based methodologies, making data collection more efficient and flexible. Studies have confirmed the effect of typographic style (Dobres et al., 2016b; Dobres, Chrysler, Wolfe, Chahine, & Reimer, 2017b), and also investigated related effects such as contrast polarity (i.e., black-on-white text or the opposite) and letter size (Dobres et al., 2016b), ambient illumination (Dobres, Chahine, & Reimer, 2017a; Wolfe, Dobres, Kosovicheva, Rosenholtz, & Reimer, 2016), visual noise and aging (Wolfe et al., 2016), font weight (Dobres, Chahine, Reimer, Gould, & Zhao, 2016a; Dobres, Reimer, & Chahine, 2016c), case and compression (Sawyer, Dobres, Chahine, & Reimer, 2017), and many other related factors.

As these laboratory methods have been used to explore typographic differences at ever-greater levels of granularity, it has become clear that there are some limitations on how small a difference these experimental methods can detect. For example, while gross differences in typographic size typically produce clear and obvious differences in dependent measures, differences in font weight have been more difficult to reveal experimentally. While typefaces with a large number of stylistic differences have been explored in the lab, typefaces that are more visually similar do not separate in the experimental data as easily (see Dobres et al., 2016a for examples). Since laboratory methods typically rely on relatively small sample sizes from which to draw inferences, it is possible that the samples are too small to attain the statistical power needed to reveal potentially significant effects. Some of the typographic factors previously investigated, such as typographic style, showed small effect sizes (Dobres et al., 2016b). Power analyses suggest that to detect these effects reliably and consistently could require up to several hundred participants. Such sample sizes are beyond the bounds of most laboratory experiments.

Smartphone Science

As smartphones and smartphone-adjacent mobile technologies have come to play central roles in daily life, interest in leveraging these platforms for scientific purposes has increased dramatically. The appeal is obvious. A smartphone is essentially an internet-connected sensor array, which can both collect data passively, and/or remind the user to perform some action or "check in" with some piece of data at regular intervals. With millions of people now carrying smartphones at all times, there exists an unprecedented opportunity for *in situ* data collection relevant to any number behavioral, psychological, and medical fields (Miller, 2012; Thomas & Azmitia, 2015). Activity in this area has gained considerable momentum, with major software companies releasing development kits specific to the collection of research-quality data from smartphone platforms ("ResearchKit - Apple Developer," n.d.; "ResearchStack," n.d.)

Smartphones have already been used as a data collection framework across a wide variety of disciplines. Smartphone-based experimental paradigms have been used to examine, for example: lexical processing (Dufau et al., 2011), mindfulness intervention (Howells, Ivtzan, & Eiroa-Orosa, 2014), linguistics (Myers, 2016), mathematical cognition (Zimmerman et al., 2016), and physical fitness interventions (Fanning et al., 2017). For extensive reviews, see (Miller, 2012; Swan, 2013; Thomas & Azmitia, 2015). Given the limitations of laboratory-based investigations of legibility discussed above, there seemed an excellent opportunity to adapt these laboratory methods to a smartphone platform. Note that while the study by Dufau *et al.* (2011) deployed a lexical decision task on a mobile platform, as the present study does, Dufau's study is markedly different in intent, presentation, and methodology. The Dufau study examined lexicality itself (i.e., whether the frequency of a word in its lexicon is related to how quickly a person can read it), had participants complete large blocks of trials at a time, and relied on participants to manually email collected data to the researchers. The present study uses the lexical decision task as a probe for text legibility (see below), deploys small numbers of trials *in situ*, and automatically centralizes data collection in real-time.

TypeTester

In the summer of 2017, the AgeLab launched a pilot study of TypeTester, a mobile application intended to conduct visual design research "in the wild". The platform sacrifices strict laboratory-level experimental control in exchange for the potential to collect large samples of data, with the hope that greater sampling would make it possible to reveal more subtle effects of visual design on perception that cannot be exposed with the smaller samples typical in lab work.

The application has three essential components: an implementation of the key legibility task, a researcher backend for configuring experiments, and a gamification element that

allowed users to track their progress and remain engaged in the data collection process (for example, tracking their number of completed trials, and maintaining contact for a weekly raffle that served as reimbursement for participating). The legibility task follows established laboratory methods for measuring legibility (see description below), while the backend grants the researcher control over the typeface used, the combination of background/foreground colors, size of the typeface, and display duration. Participants who enrolled in the pilot would receive periodic notifications containing short "challenges"—small groups of experiment trials. Participants were entered into a weekly raffle as reimbursement for continued participation. Details on how TypeTester was designed are presented in Klarl (2017).

Methods

Lexical Decision Task

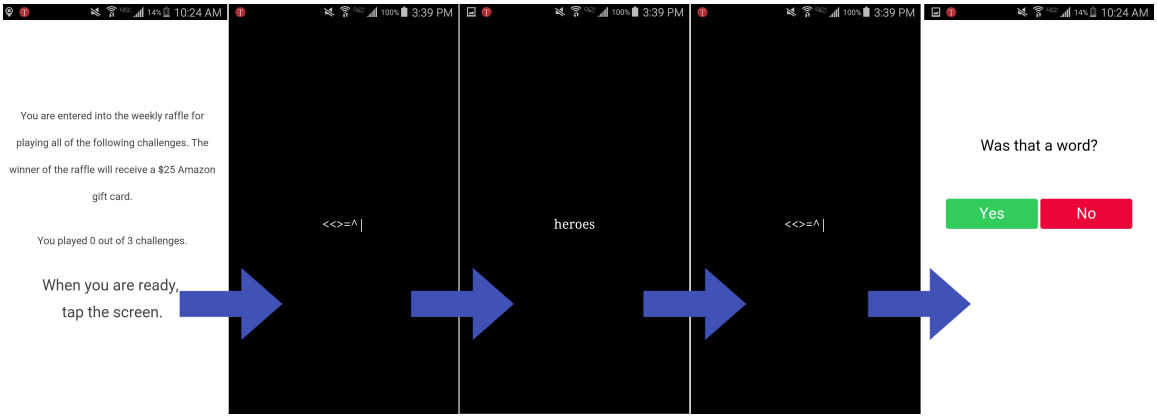


Figure 1: An illustration of a single lexical decision trial. The participant is presented with a prompt screen describing the number of trials to be completed in this "challenge". A gibberish visual mask is displayed for 200ms, followed by the target word/non-word (variable timing), and then a final mask. Finally, the user is prompted to decide whether the target was a word or non-word. The font and text sizes used in the illustration above do not represent those of the final experiment. The figure omits a brief blank screen between the prompt and the first mask.

Previous research at the AgeLab has utilized a yes-no lexical decision task (Meyer & Schvaneveldt, 1971) to probe the legibility of various typographic configurations. In this paradigm, a word or non-word is presented for some brief display duration, and the participant is then asked to indicate whether the stimulus just seen was a valid word. A schematic of the task as implemented in TypeTester is presented in Figure 1. The participant's mobile device receives a notification indicating that a new "challenge" is ready. Upon opening the application, the participant is presented with a "ready" prompt indicating the number of trials to be completed in this challenge. The participant taps to indicate that she/he is ready, and then a lexical decision trial is presented. This is comprised of an initial blank screen matching the background color of the trial's stimulus configuration, followed by a gibberish mask that is displayed for 200ms. This is followed

by the target word/non-word, which is presented for a predetermined variable time, as described below. This is immediately followed by another 200ms mask, and finally, a response screen in which the user is prompted to decide whether the stimulus was a word or non-word. Participants were not provided with feedback regarding the accuracy of their responses either during or after the completion of trials.

Challenges were delivered at random intervals throughout the time windows that participants indicated they would be willing to receive messages. The goal of this random deployment of trials was to attempt to capture data from a variety of potential multitasking situations, as well as indoor/outdoor lighting conditions and different activity states (i.e., sitting, walking, etc.).

Conditions

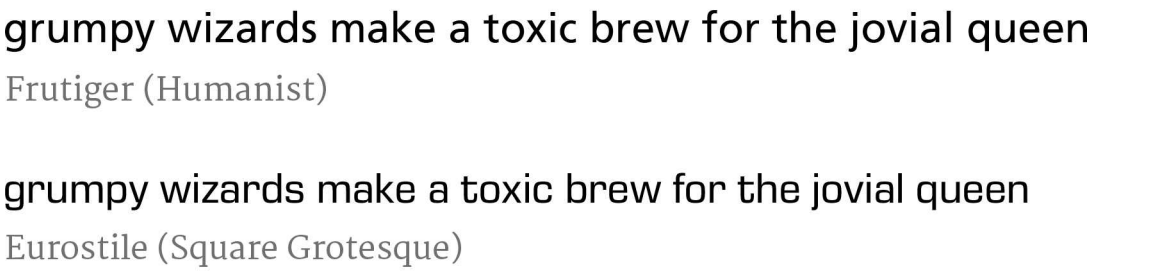


Figure 2: Pangram type samples of the two typefaces used in the present study: Frutiger (top) and Eurostile (bottom). Figure from (Dobres et al., 2016b).

The primary stimuli of this experiment were English words and non-words selected from an online orthographic database (Medler & Binder, n.d.). Words were selected such that they were relatively common in the English lexicon, and constrained to be exactly six letters in length. Six-letter non-words were generated with properties to appear similar to the word list. For further details on the creation of these lists, see (Dobres et al., 2016b). Gibberish masks were composed of randomized non-letter characters, and two different randomized masks were used for the pre- and post-stimulus masks of each trial. All word/non-word stimuli were presented in lowercase letters.

This study examined conditions previously investigated in the context of a driving simulator and under classical psychophysical paradigms (Dobres et al., 2016b; Reimer, Mehler, Dobres, Coughlin, Matteson, Gould, Chahine, & Levantovsky, 2014). The Frutiger and Eurostile typefaces were used, and both were presented in two color combinations (positive polarity black-on-white, and negative polarity white-on-black). In addition, each of these 4 combinations was presented at 5 display durations (33, 67, 100, 133, and 167ms), for a total of 20 conditions. As in previous experiments, “white” was hex color #ffffff, and “black” was hex color #000000. All masks and lexical decision stimuli were presented at the center of the device screen. Frutiger and Eurostile were presented at nominal heights of 22 points and 20 points, respectively, which equalized the

heights of their capital 'H' characters on screen. Note that here, "point" does not refer to the classic typographic size unit, but rather, a virtualized pixel. On lower-end mobile device with relatively low-resolution screens, a single "point" may correspond to a single hardware pixel. On high-resolution screens, a single "point" may be composed of several pixels. While the operating system usually treats one points' worth of pixels as a single unit for screen layout and for determining overall font sizes, various graphics routines make use of the individual pixels within each point for added sharpness. This is particularly common for text rendering algorithms.

Participants & Participation

Upon downloading the application from the Google Play store and launching it, participants were asked to affirm that they were at least 18 years of age, U.S. citizens, and native English speakers. They were then presented with a brief consent document approved by MIT's institutional review board for human subjects research. The consent document explained the type of data that was to be collected, potential risks/rewards of participating, the right to stop participation at any time, and guarantee of data anonymity. After the consent was digitally signed, participants provided basic demographic information and a valid email address (which was used only for reimbursement purposes and was stored separately from experimental data). Participants then received a small set of practice trials designed to familiarize them with the experiment task. Subsequently, challenges containing trials pertinent to the main experiment were delivered at intervals throughout the participant's chosen participation windows.

A total of 151 individuals participated in the TypeTester program through mid-September 2017. Participants ranged in age from 19 to 84, with a median age of 44. The representation of ages across the lifespan skewed slightly young, but nevertheless represents a good variety of ages (Figure 3). Among participants who disclosed a gender, 63 identified as male, 62 as female, and 1 as transgender. The remaining 25 gave no response for this question. Age did not differ significantly between genders ($p = 0.496$).

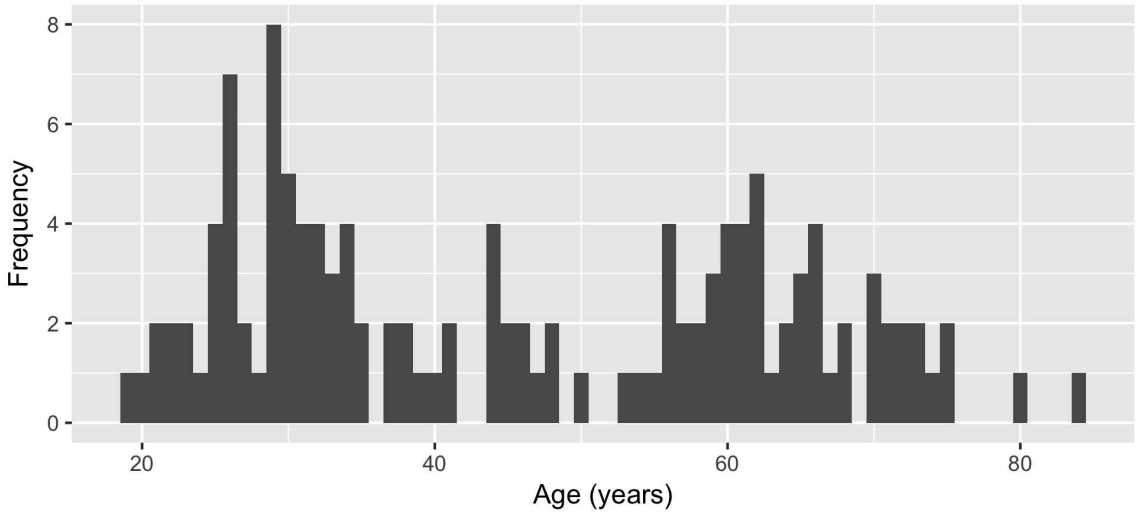


Figure 3: Age distribution for participants who disclosed a date of birth.

Trial-level data indicate that the majority of responses were collected within the first month of launch (Figure 4). The median participation window was 6.2 days (75% of users stopped participating after 10 days). Few trials were collected on weekends, suggesting that most users retained the default notification settings, which exclude weekends. Most participants contributed a relatively large number of trials. Half of all participants completed at least 97 trials, and 80% of participants completed at least 93 trials (Figure 5).

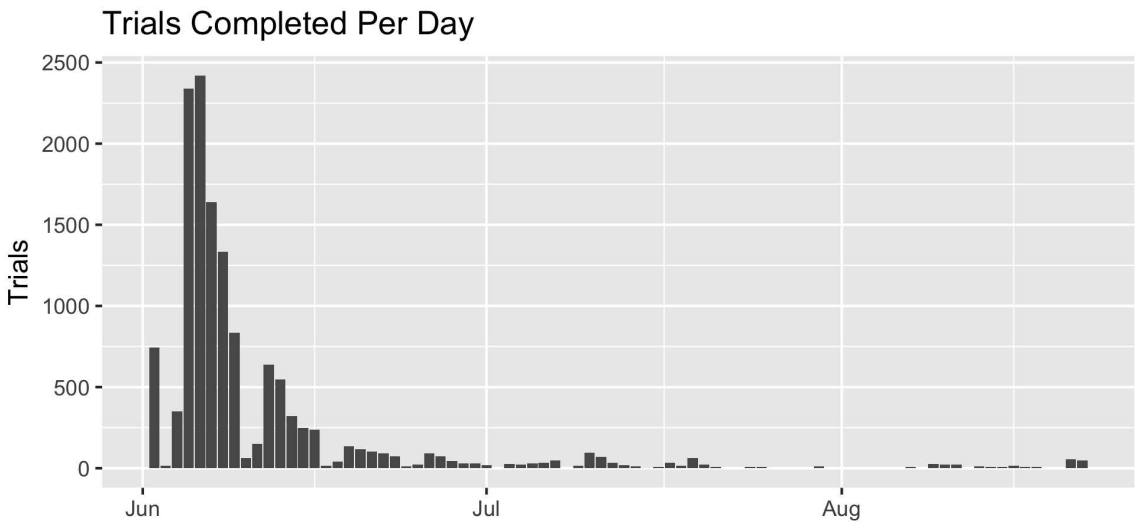


Figure 4: Trials completed per day since initial launch.

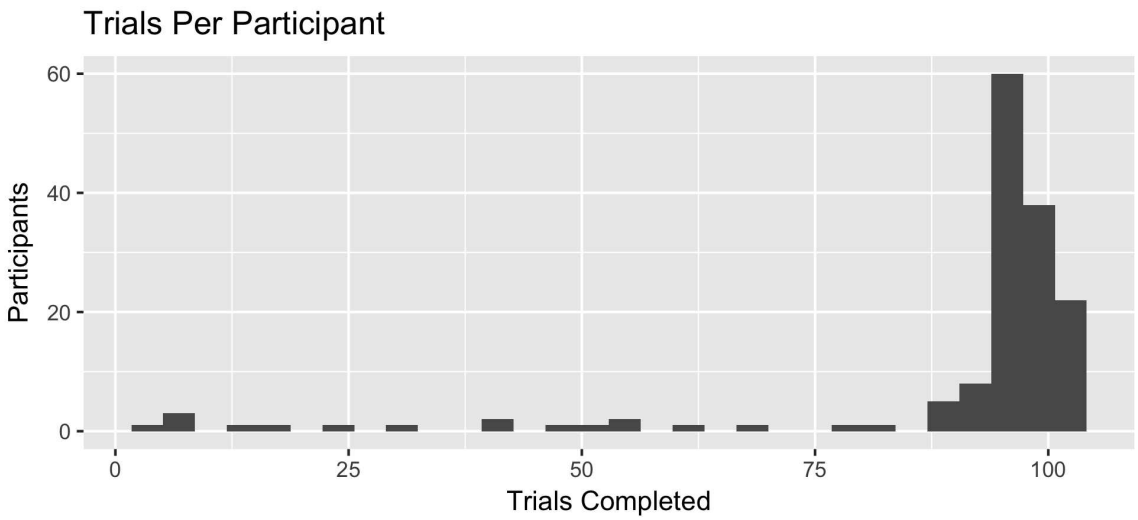


Figure 5: Distribution of trials completed per participant, suggestive of a high degree of engagement among those who agreed to participate.

Data Reduction & Metrics

Data were analyzed in R (R Core Team, 2018). Trials with a response time greater than 5 seconds were excluded from analysis (and from the statistics presented above), constituting 1.6% of the raw data. The 5-second cutoff was chosen for consistency with lab standards, which grant the participant no more than 5 seconds for a response. Median response time in the remaining sample was 1.2 seconds. While this is considerably slower than response times observed in laboratory conditions (which are on the order of 100-300ms), it is to be expected that a sporadic finger tap to a touch screen would be slower than a highly practiced key press in a lab.

These data comprise 13,619 trials, roughly equivalent to a sample size of about 34 participants in a 4-condition laboratory study. Notably, trial distribution was random and somewhat uneven: 3,583 Eurostile Negative, 2,483 Frutiger Negative, 2,570 Eurostile Positive, and 4,983 Frutiger Positive.

The primary metric of this pilot is response accuracy, that is, how often a correct response was made per each condition. Trial responses were averaged per participant and condition to produce individual-level accuracy measures. An alternative approach that disregards participant identification and treats each condition as a large "bucket of trials" produced similar statistical results. The individual-level approach is therefore used for greater consistency with laboratory results, which would typically be calculated per participant. Note that trials were randomly and unevenly distributed to each participant, and it cannot be guaranteed that each participant saw every condition, or even a relatively equal number of trials per condition.

Activity state was determined from estimates reported by the Android operating system. Based on accelerometer and gyroscopic data, the system provided estimates for various activity states on a scale of 0-100. Possible activity states included: in vehicle, on bicycle, on foot, walking, running, tilting, still, and unknown. A confidence level was reported for each of these possibilities per trial. The state with the highest confidence rating is taken as the activity on that trial.

Results

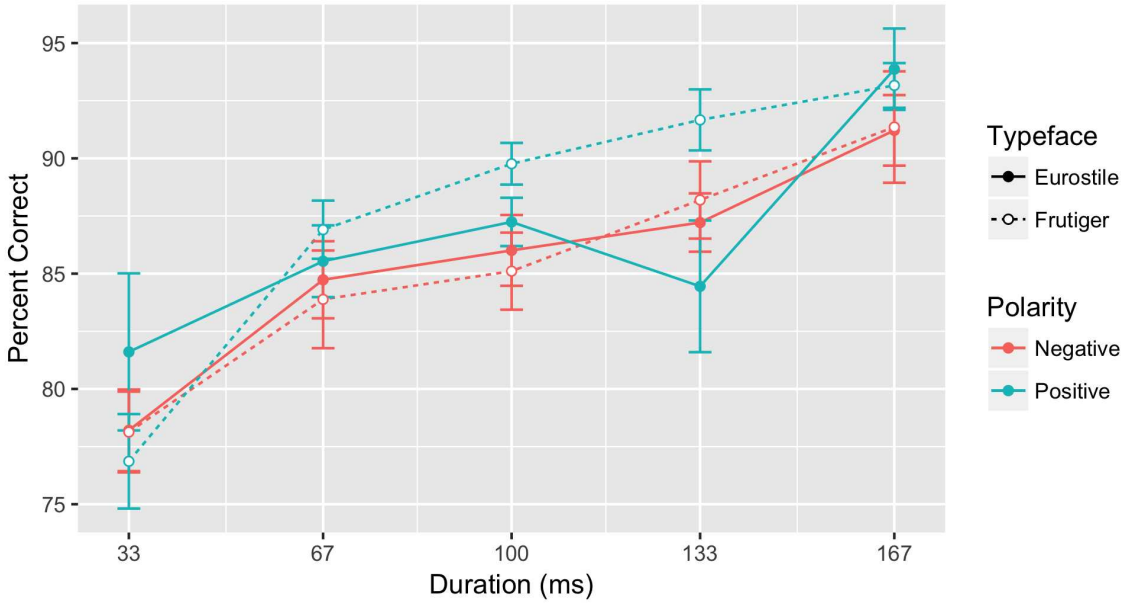


Figure 6: Mean performance accuracy in the 4 conditions of interest at each display duration (error bars are ±1 mean-adjusted SEM).

Figure 6 shows response accuracy across all conditions examined in this study. An ANOVA that included age group (younger than 44/44 and older), gender, typeface, polarity, and display duration as predictors found significant main effects of display duration ($X^2 = 95.7$, $p < 0.001$), and polarity ($X^2 = 5.34$, $p = 0.021$), and a statistical trend for age ($X^2 = 2.07$, $p = 0.150$; when considered as a linear predictor instead of a group, this effect weakens somewhat). No other significant main effects or interactions are evident. Notably, the effect of typeface (Frutiger or Eurostile) on response accuracy shows no evidence of statistical significance or of a trend toward significance ($X^2 = 0.06$, $p = 0.800$). For clarity, a plot that aggregates across typeface is included below (Figure 7).

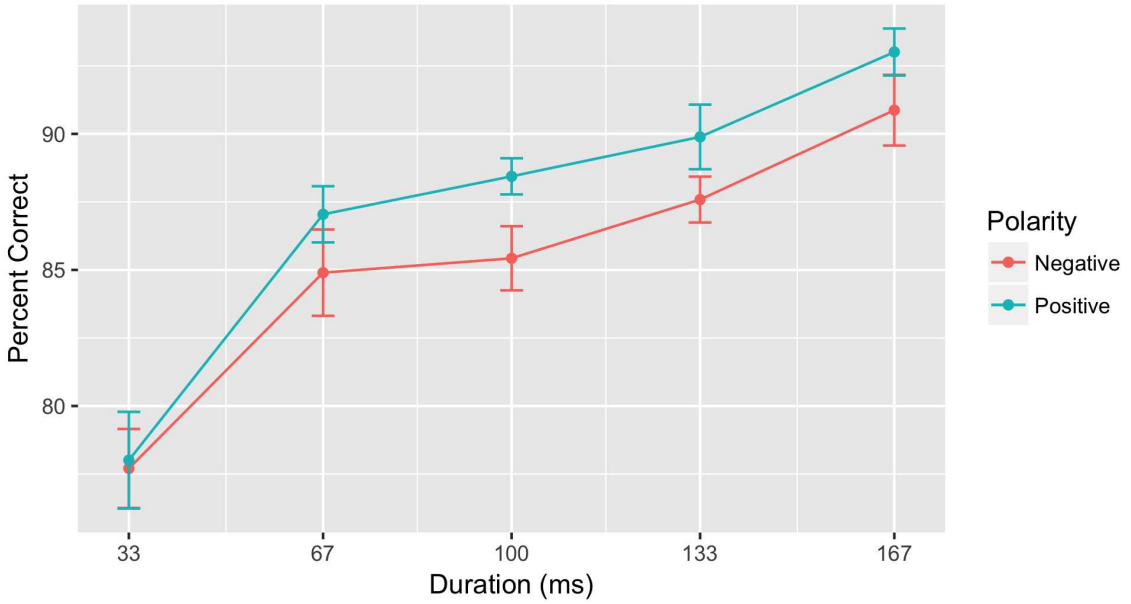
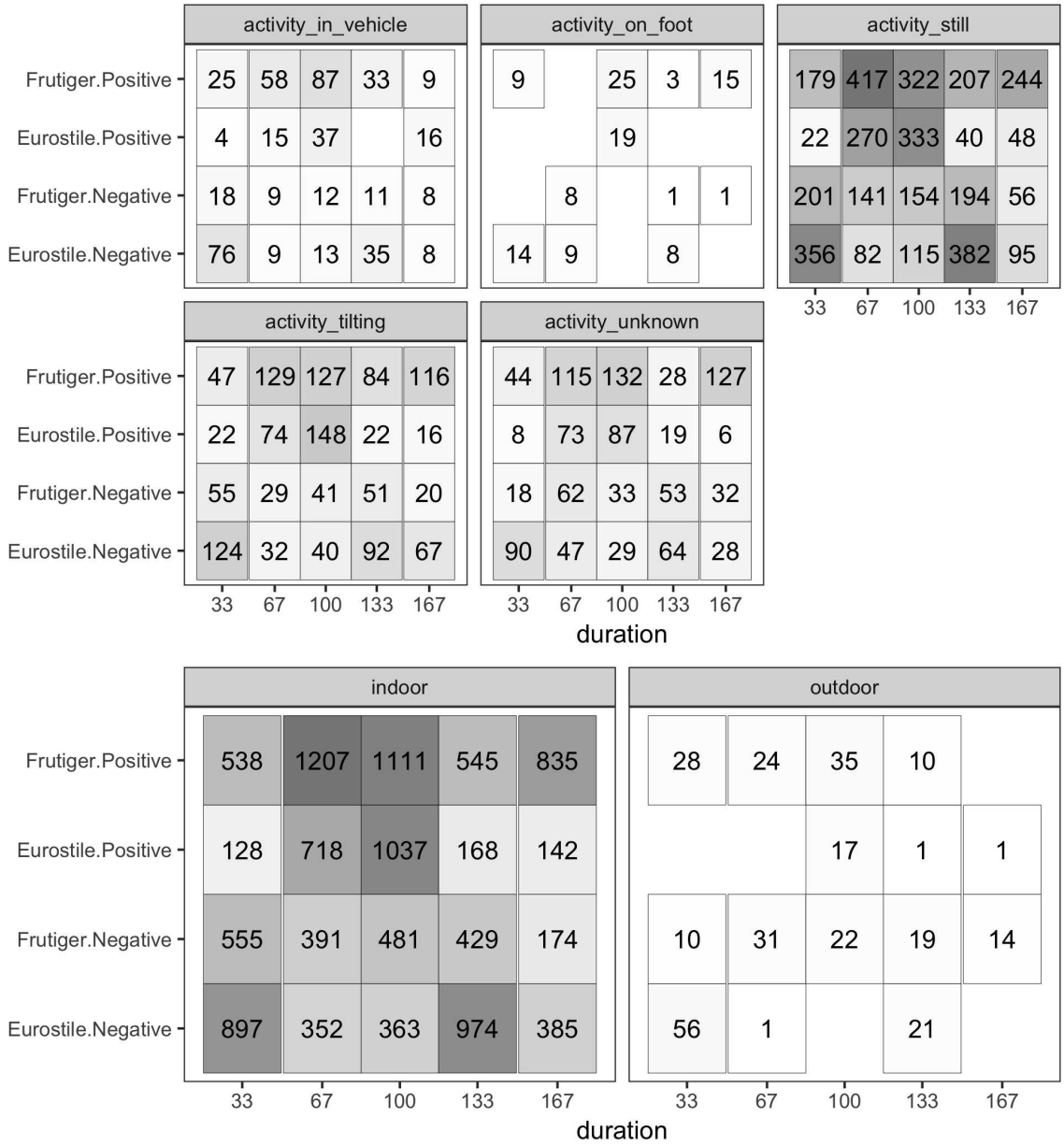


Figure 7: Mean performance accuracy for each display duration and polarity, aggregated across typefaces (per participant). Labeling as in Figure 6.

Although the data do follow a typical psychophysical curve, with accuracy rising along with display duration (as we would expect, as longer durations allow more reading time), the results are curious. The data suggest that 80% response accuracy would be achieved quite early in the curve, at speeds faster than 67ms. This is much faster than the 80% accuracy threshold seen in earlier experiments, which typically required 80-140ms depending upon condition. However, it is difficult to say whether the size of the fonts used here is optically larger than those seen in the lab, which may affect these measures.

Finally, the use of a smartphone-based platform allows for the collection of environmental data, such as the user's activity state (stationary, walking, in a vehicle, etc.) and ambient illumination. The application successfully collected illumination data for 86% of trials, though it should be noted that many of these values, while non-zero (zero indicates "unknown"), are suspiciously low. Likewise, activity state was estimated with sufficient confidence (i.e., "activity unknown" was not the estimate with the highest probability) on 84% of trials.

Unfortunately, illumination and activity states are quite lopsided in this sample. The vast majority of trials were conducted in a "stationary" or "tilting" state. Similarly, the overwhelming majority of trials were conducted under indoor lighting (here, indoor lighting is liberally defined as less than 2000 lux; the two categories are usually easy to separate, with typical indoor lighting being around 500 lux, and outdoor lighting ten times that). The tables below give counts for each state and condition:



Given the sparseness of trials conducted in non-stationary states, and especially of trials conducted outdoors, it would be statistically unsound to attempt any meaningful comparisons of these dimensions.

Discussion

These data suggest that a mobile data collection platform for visual design issues has some potential. Even with relatively little in the way of resources and promotion, enough data were collected to represent a typical laboratory-based sample size. Participants spanned a wide age range and were gender-balanced. Participation rates suggest that the sample pool was highly motivated, and that most people were willing to participate for 1-2 weeks. Primary results demonstrate a significant effect of typeface polarity consistent

with previous research. However, there was no evidence that the pilot was able to separate differences between typefaces (indeed, across all trials, accuracies were 86.6% for Frutiger and 86.1% for Eurostile). A secondary analysis that included only trials with a response time of 1 second or less (attempting to capture only highly attentive responses) does not produce results different from those seen in the general sample. While the experiment controlled for differences in size between the two typefaces by requesting appropriately scaled sizes for each, it is impossible to control the experiment-wide display size across a panoply of digital devices and casual viewing distances. It is very likely that fonts appeared optically larger in this study than in previous laboratory-based work. This may explain why this experiment failed to detect an effect of typeface, especially given that previous studies have shown that typeface differences become more pronounced at smaller sizes (Dobres et al., 2016b).

The significant effect of display duration is consistent with expected psychophysical patterns, and in this sense is reassuring. However, the high response accuracy observed even for very brief presentation times (67ms) is a cause for concern. This suggests that the smartphone platform may not be reliable enough to provide accurate stimulus timing. It is also possible that the probable larger display size of the stimuli, as described above, made the task substantially easier. If timing is the culprit, this concern might be overcome simply by accepting the lower timing accuracy and choosing wider duration gaps.

Although the platform successfully recorded activity state and illumination in the majority of cases, it appears that participants were rarely willing to participate in TypeTester challenges when outdoors or when not stationary. One has to wonder if pausing one's movements indoors and then completing the TypeTester task is due to an active effort to avoid environmental characteristics outdoors, or is more a byproduct of overall phone use behavior. Likewise, the borderline effect of age suggests that a larger sample would be beneficial. If a much larger sample could be obtained (perhaps ten times as many responses), other patterns may begin to emerge. As discussed above, one of the great advantages of these types of distributed data collection platforms is their ability to gather very large amounts of data. In this case, however, TypeTester was able to collect a number of responses equivalent to a moderately sized laboratory sample. The major limiting factor in this case seems to be that TypeTester was only made available for the Android operating system, and was not compatible with the popular iOS/iPhone platform. It is worth pointing out, however, that once the platform was deployed, data collection was entirely passive. No significant staff time was required for data collection, reimbursement costs were minimal ($25/week), and participants could contribute data without needing to travel to a specific location or otherwise take time out of their regular daily schedules. These advantages should not be dismissed. Initiatives such as TypeTester are in their infancy, and we believe that they show substantial potential for scientific research in the future.

References

Dobres, J., Chahine, N., & Reimer, B. (2017a). Applied Ergonomics. *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*, 1–15. 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., & Chahine, N. (2016c). The Effect of Font Weight and Rendering System on Glance-Based Text Legibility (pp. 91–96). Presented at the the 8th International Conference, New York, New York, USA: ACM Press. http://doi.org/10.3758/BF03195482

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F. X., et al. (2011). Smart Phone, Smart Science: How the Use of Smartphones Can Revolutionize Research in Cognitive Science. *PLoS ONE*, *6*(9), e24974. http://doi.org/10.1371/journal.pone.0024974.g001

Fanning, J., Roberts, S., Hillman, C. H., Mullen, S. P., Ritterband, L., & McAuley, E. (2017). A smartphone "app-"delivered randomized factorial trial targeting physical activity in adults. *Journal of Behavioral Medicine*, *40*(5), 712–729. http://doi.org/10.1093/geronb/gbn032

Howells, A., Ivtzan, I., & Eiroa-Orosa, F. J. (2014). Putting the "app" in Happiness: A Randomised Controlled Trial of a Smartphone-Based Mindfulness Intervention to Enhance Wellbeing. *Journal of Happiness Studies*, *17*(1), 163–185. http://doi.org/10.1016/j.jpain.2009.07.015

Klarl, K. (2017). *Legibility at a Glance–Concept, Design, and Evaluation of a Mobile Platform for Conducting Behavioral Research*. (Master's Thesis). University of Augsburg. Augsburg, Germany.

Medler, D. A., & Binder, J. R. (n.d.). MCWord. Retrieved from http://www.neuro.mcw.edu/mcword/

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234.

Miller, G. (2012). The Smartphone Psychology Manifesto. *Perspectives on Psychological Science*, *7*(3), 221–237. http://doi.org/10.1021/ac201587a

Myers, J. (2016). Meta-megastudies. *The Mental Lexicon*, *11*(3), 329–349. http://doi.org/10.1016/j.neuropsychologia.2015.08.027

R Core Team. (2018). *R*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reimer, B., Mehler, B., Dobres, J., Coughlin, J. F., Matteson, S., Gould, D., et al. (2014). Assessing the impact of typeface design in a text-rich automotive user interface. *Ergonomics*, *57*(11), 1643–1658. http://doi.org/10.1080/00140130903464358

ResearchKit - Apple Developer. (n.d.). ResearchKit - Apple Developer. Retrieved November 5, 2017, from https://developer.apple.com/researchkit/

ResearchStack. (n.d.). ResearchStack. Retrieved November 5, 2017, from http://researchstack.org/

Sawyer, B. D., Dobres, J., Chahine, N., & Reimer, B. (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. http://doi.org/10.1037/h0031564

Swan, M. (2013). The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. *Big Data*, *1*(2), 85–99. http://doi.org/10.1089/big.2012.0002

Thomas, V., & Azmitia, M. (2015). Tapping Into the App. *Emerging Adulthood*, *4*(1), 60–67. http://doi.org/10.1111/cdep.12040

Wolfe, B., Dobres, J., Kosovicheva, A., Rosenholtz, R., & Reimer, B. (2016). Age-related differences in the legibility of degraded text. *Cognitive Research: Principles and Implications*, 1–13. http://doi.org/10.1186/s41235-016-0023-6

Zimmerman, F., Shalom, D., Gonzalez, P. A., Garrido, J. M., Alvarez Heduan, F., Dehaene, S., et al. (2016). Arithmetic on Your Phone: A Large Scale Investigation of Simple Additions and Multiplications. *PLoS ONE*, *11*(12), e0168431. http://doi.org/10.1371/journal.pone.0168431.t003

Authors

Jonathan Dobres, Ph.D.

Dr. Dobres is now a Senior Data Scientist at Sonos. While at the AgeLab, his research was primarily concerned with how the properties of digital text affected legibility and performance on concurrent tasks, as well as the visual and cognitive demands associated with tasks while driving. He received a BA, MA, and PhD in Psychology from Boston University, where his research examined how visual perception changes over time with training.

Karola Klarl

During the development of this work at the AgeLab, Karola Klarl was a visiting student from Munich, Germany. She received her Bachelor's degree in Computer Science from the Technische Universität München, Germany. Her Master's degree she received with honors as part of the Elite Master Programme Software Engineering which is offered within the framework of the Elite Network of Bavaria by the University of Augsburg, the Ludwig-Maximilians-Universität München and the Technische Universität München.

Julia Kindelsberger

Julia Kindelsberger was a Software Engineering graduate student visiting the AgeLab from the Technical University of Munich, Germany when she assisted on this project. Her research interests included green mobile application development and performance measurements during her Bachelor's at TUM. At MIT, she worked on human-centered-artificial intelligence with a focus on human-machine-interaction in semi-autonomous vehicles.

Bryan Reimer, Ph.D.

Bryan Reimer is a Research Engineer in the Massachusetts Institute of Technology AgeLab and the Associate Director of the New England University Transportation Center. His research seeks to develop new models and methodologies to measure and understand human behavior in dynamic environments utilizing physiological signals, visual behavior monitoring, and overall performance measures. Dr. Reimer leads a multidisciplinary team of researchers and students focused on understanding how drivers respond to the increasing complexity of the operating environment and on finding solutions to the next generation of human factors challenges associated with distracted driving, automation and other in-vehicle technologies. He directs work focused on how drivers across the lifespan are affected by in-vehicle interfaces, safety systems, portable technologies, different types and levels of cognitive load. Dr. Reimer is a graduate of the University of Rhode Island with a Ph.D. in Industrial and Manufacturing Engineering.

About the AgeLab

The Massachusetts Institute of Technology AgeLab conducts research in human behavior and technology to develop new ideas to improve the quality of life of older people. Based within MIT's Engineering Systems Division and Center for Transportation & Logistics, the AgeLab has assembled a multidisciplinary team of researchers, as well as government and industry partners, to develop innovations that will invent how we will live, work and play tomorrow. For more information about AgeLab, visit agelab.mit.edu.

Massachusetts Institute of Technology > **AgeLab** > 77 Massachusetts Ave, E40-275, Cambridge, MA 02139
Phone: 617.253.0753 > fax: 617.258.7570 > agelab@mit.edu > agelab.mit.edu