This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Human–Computer Interaction on 20 December 2017, available online: https://www.tandfonline.com/doi/full/10.1080/10447318.2017.1415688

Title:

Effects of Quantity and Size of Buttons of In-Vehicle Touch Screen on Drivers' Eye Glance Behavior

Shortened Title:

QUANTITY AND SIZE OF BUTTONS ON DRIVER EYE GLANCES

Authors:

Fred Feng [a], Yili Liu [b], Yifan Chen [c]

Affiliations:

[a] University of Michigan Transportation Research Institute, Ann Arbor, MI, USA

[b] University of Michigan, Department of Industrial and Operations Engineering, Ann Arbor, MI, USA

[c] Ford Motor Company, Dearborn, MI, USA

Address:

[a] 2901 Baxter Road, Ann Arbor, MI, 48109

[b] 1205 Beal Avenue, Ann Arbor, MI, 48109

[c] 1 American Road, Dearborn, MI, 48126

Abstract:

This paper investigates the effects of the quantity and size of touch screen buttons and the task interleaving strategies on drives' eye glance behavior. An experiment was conducted on a fixed-base driving simulator with 20 participants. The participants were asked to perform a button search-and-press task on an in-vehicle touch screen while driving. A full-factorial within-subject design was used with three button quantities (4, 8, and 15) and three button sizes (14 mm, 24 mm, and 33 mm). Although a normal distribution was often assumed for the eye glance data in previous studies, our results show that the total eyes-off-road time (TEORT) and glance durations are generally not normally distributed (positively skewed) even after a log transformation. The results show that the number of buttons has an increasing effect on task completion time, TEORT, and long (2+ seconds) glances. However, in general, no such differences were found for button sizes. Further analysis shows that long glances were strongly associated with drivers completing the task with a single glance. It seems to suggest that a major cause of long glances is that drivers are reluctant to switch the task back to driving at subtask boundaries that are probably associated with high cost of interruption. These findings confirm the importance of task resumability for in-vehicle user interfaces and have implications that careful task analysis needs to be conducted in the context of multitasking. Certain subtask combinations, such as a visual search followed by pressing the search target, may discourage task interleaving and ultimately compromise driving safety.

Keywords:

Driving; Visual search; Visual display; Touch screen; Multitasking; Infotainment system

1. Introduction

The last decade has witnessed a steady trend of modern electronic technologies, such as touch screens, digital instrument clusters, and head-up displays, becoming common features in many vehicles (Becker et al. 2014). These technologies enable the vehicle cockpit to integrate a large number and variety of functions (e.g., audio, climate control, communication, navigation) into a single device, and allow the drivers to complete tasks such as browsing music, making phone calls, and finding the nearest gas station. Although these functions are designed to enhance the driving experience, they may suffer from usability problems such as requiring extended eyes-off-road operations, information overload due to cluttered displays and system complexity (Lee, 2007; Becker et al. 2014).

Driver distraction has been a significant contributing factor in road accidents. According to the National Highway Traffic Safety Administration (NHTSA), in 2015 distracted driving accounted for 3,477 fatalities and an estimated additional 391,000 injuries in the United States (National Center for Statistics, 2017). In addition, these numbers are likely under-reported due to the difficulties in identifying driver distraction during an accident investigation. A naturalistic driving study suggests distraction of secondary tasks (i.e., those tasks not necessary to driving) account for 23% of all crashes and near-crashes (Klauer et al., 2006).

Empirical studies have shown that performing visual-manual tasks while driving may degrade drivers' performance in many aspects such as steering control and lane keeping performance (Tsimhoni, et al., 2004; Peng et al., 2013; Bao et al., 2015; Pavlidis et al., 2016), headway control and braking behavior (Harbluk et al., 2002; Lansdown, 2004), and response to sudden or hazard events (Greenberg et al., 2003; Horrey & Wickens, 2004). Studies have also shown a relationship between the visual demands of in-vehicle systems and crash risk and accident occurrence

(Wierwille & Tijerina, 1998; Horrey & Wickens, 2007). The naturalistic driving study by Klauer et al., (2006) indicated a statistical association between long (2+ seconds) off-road eye glances and increased near-crash/crash risk from baseline driving. Standards and guidelines have been proposed to evaluate a variety of secondary tasks and the designs of in-vehicle systems. The Society of Automotive Engineering (SAE) Standard SAE J2364 (2004) proposes that the maximum time for drivers to complete navigation-related tasks involving visual displays and manual controls should be less than 15 seconds (referred to as the 15-Second Rule) (Green, 1999). NHTSA published a guideline for in-vehicle electronic devices with recommendations that for 85^{th} percentile of the driver sample: (1) the mean duration of off-road glances should be less than 2.0 seconds, (2) no more than 15% of total number of glances should be greater than 2.0 seconds, and (3) the total eyes-off-road time should be no greater than 12.0 seconds (referred to as the 2/12 Rule) (NHTSA, 2012).

A large body of studies has been conducted to examine the designs of displays and controls and their effects on user performance (see Moacdieh & Sarter (2015) for a literature review of display clutter). The number of items on a screen is a crucial design factor as it affects the visual search efficiency and is a major contributor to display clutter. As summarized by Wolfe (2007), the efficiency of visual search decreases with the increasing number of distractors. Visual search efficiency is also affected by the spatial distribution of the items in the visual field. As the density of the items increases, the visual search usually becomes faster (Nothdurft, 2000), as it is less likely to require eye saccades and head movements to move the new items into the foveal vision. However, the search time increases when the items are getting too close to each other, so that it prevents the identification of the individual items (Vlaskamp & Hooge, 2006). Focusing specifically on the in-vehicle systems, Yoon et al., (2015) examined how a list of quantitative measurements of vehicle instrument cluster (e.g., the quantity of icons, icon size, number of divisions, variety of colors, and text-to-graph ratio) may affect its perceived visual complexity and visual search performance. Continuing with this work, Lee et al., (2016) further examined how the perceived visual complexity of vehicle instrument cluster may affect drivers glance behavior and preferences.

The size of buttons on a touch screen is also an important design parameter. Fitts's Law (Fitts, 1954) implies that the time required to rapidly move to a target region increases with decreasing target width. Several recommendations have been given regarding the minimum button size on a touch screen (e.g., 19 mm (3/4 inch) by Monterey Technologies Inc. (1996) and 22 mm by Greenstein (1997)). The basic idea was that the button should be at least as big as the size of an adult human fingertip (typically 16-20 mm in diameter, Dandekar et al., 2003). Jin et al (2007) studied the effect of touch screen button sizes, spacing, and manual dexterity on the reaction time, accuracy and subjective preferences of older adults. They found that longer reaction times and lower accuracy were elicited with small buttons, but the increase of accuracy plateaued with larger button sizes.

Most of the above studies assume a single task condition, where they could devote all their attention resources to the task. Many efforts have also been made to investigate how humans switch attention in a dual-task condition. It is generally believed that humans utilize subtask boundaries or subgoal completions as natural break points to interleave tasks (Miyata & Norman, 1986; Payne et al., 2007; Janssen et al., 2012). Iqbal and Bailey (2005) examined the use of workload-aligned task models to predict opportune moments for interruption. They found that interrupting tasks at the predicted subtask boundaries with the lowest mental workload consistently caused less time to resume the interrupted task after completing the interrupting task. Using a series of experiments involving phone-dialing and steering tasks, Brumby et al. (2009) and Janssen et al. (2012) examined the role of priorities and cognitive and motor cues on the patterns of task interleaving. They found that drivers tended to suspend the phone-dialing task and switch to driving at both cognitive chunk boundaries (e.g., the three chunks of a telephone number 734-122-2288 in the U.S. formatting convention) and motor cues (e.g., moving the finger to a different digit after typing in a complete set of repeating digits). Lee et al. (2015) examined driver's glance patterns at task boundaries in a message reading and button pressing task. They found that drivers adopted distinctive glance patterns at the subtask boundary (i.e., pressing a button after reading a message) that was defined by a cognitive cue (i.e., end of the message) and motor cue (i.e., move the hand to the button). Lee & Lee (2017) further developed a driver task switching model that performs text reading and text entry tasks while driving. A linear combination of task structural constraints (i.e., subtask boundary) and environmental demand (i.e., accumulated uncertainty on the road) was used to model the driver's task switching behavior.

In the automotive industry, there have been many guidelines and standards on the design of traditional vehicle controls such as push buttons, toggle buttons, knobs (SAE J2119, 1993; ISO 2575, 2010; Stevens, 2002). However, there are not yet similar guidelines and standards available for in-vehicle touch screen, which is gaining popularity for in-vehicle information or infotainment systems (Harvey et al., 2011; Feng et al., 2017). Nonetheless, in recent years there has been a growing number of studies that investigated the in-vehicle touch screens design factors in the context of multitasking and driving. Kujala & Saariluoma (2011) and Kujala (2013) studied the effect of touch screen menu structure (grid- and list-layout), number of display items per screen (2, 4, 6, or 9), and scrolling style (arrow buttons, swipe, and kinetic) on drivers eye glance behavior. They found that a list-style menu led to smaller variance in glance durations compared to a grid-layout menu (especially with 9 display items), and the number of display items had a significantly increasing effect on the maximum glance durations. More recent work by Kujala & Salvucci (2015) and Salvucci and Kujala (2016) further investigated drivers' visual sampling strategies on an in-vehicle touch screen. They found that the number of menu items had a significant increasing effect on the number of off-road glances, total eyes-off-road time, both mean and maximum glance durations, and the percentage of long (2+ seconds) glances. A computational model of multitasking was also developed by the authors to quantify the balance between the structural constraints (e.g., how a task can be broken down to subtasks) and temporal constraints (i.e., how long can a driver look away from the road). It is to note that in this study the authors mainly focused on the in-vehicle task of visual search and excluded the button-pressing task by only using the data of absent target search in the screens prior to the final screen that included the search target.

In addition, Boyle et al, (2013) examined drivers' glance behavior in a text entry and reading task. They found that the drivers' mean glance duration and total eyes-off-road time (TEORT) were significantly longer for long text string lengths compared to short text string lengths. Drivers engaged in the long text entry tasks (12 characters) also had significantly longer TEORT values compared to medium (6 characters) and short (4 characters) text entry tasks. Crundall et al. (2016) examined the effect of text size (ranging from 4 mm to 9 mm) of a touch screen on driver eye glance behaviors. They found that there was no evidence of text size on the total eyes-off-road time. However, larger text sizes were associated with a high number of relatively short glances, whereas smaller text led to a smaller number of long glances.

The objective of this paper is to investigate the effects of two basic design parameters of touch screen user interfaces, namely the quantity and size of buttons, on drives' eye glance behavior in the context of driving. An experiment was conducted in a driving simulator, in which drivers were asked to search for and then press a specific button on a touch screen. Nine designs were tested with varying quantity of buttons and button sizes. The drivers' task operation and eye glance behavior were recorded from video cameras and manually extracted after the experiment. A number of metrics were examined that included task completion time, total eyes-off-road time, number of glances used, glance duration, and the occurrence of long glances.

It is worth to note that one of the most relevant work to this paper is a study by Large et al. (2017). In this study a series of experiments were conducted to develop a model that predicts the visual demand (total glance time, mean glance duration, and number of glances) elicited by in-vehicle touch screen designs with varying button size (from 6 mm to 24 mm), numbers of buttons (from 1 to 36), and structure conditions (alphabetical or unstructured). However, with a main goal of developing a prediction model, the study only examined the typical values of total glance time, mean glance duration, and number of glances, rather than the distributions of the metrics and particularly long glance durations, which are arguably most safety-relevant and have a stronger relation to crash risk (Horrey and Wickens, 2007). In addition, in the experiment by Large et al., (2017) the drivers were specifically asked to keep hands on the steering wheel until they have located the target. This was reasonable as their goal was to develop a model that incorporates the Fitts's Law and Hickman-Hyman Law, and to do so they needed to ask the drivers to separate the searching and pointing component of the task. However, in our study no such requirement was given to the drivers, as we aim to mimic the real-world interactions when requirements on task execution at a level of arm and hand movement were rarely given.

2. Methods

2.1. Participants and experimental setup

Twenty participants were recruited for the experiment (10 male and 10 female, half between 20 and 39 years old, and half between 40 and 69 years old). All the participants were employees from an automotive company in the United States and had their driver license for at least one year. The time the participants spent during the experiment counted as their company work time.

The experiment was conducted in a fixed-base driving simulator (shown in Figure 1). The front road scene was projected on a flat screen in front of the cockpit. The participants were asked to drive the simulator on a virtual highway and maintain a speed of about 60-70 miles per hour (97-113 km/h). The virtual highway is a square loop with two lanes in one direction. The participants were asked to stay in the left lane in which there is no other vehicle. To avoid the additional factor of the participant's initial right-hand position when performing the button search-and-press tasks, they were asked to always put both hands on the steering wheel except when performing the tasks. The driving environment was set as daytime.

INSERT FIGURE 1 HERE

An 8-inch (diagonal size) touch screen was mounted in the center console area of the cockpit. The touch screen is resistive, mono-touch with no tactile feedback, and has a resolution of 800-by-480 pixels. A driver-facing video camera was used to capture the drivers' eye glance behavior. Another camera facing the center console area was used to capture the drivers' operation on the touch screen. Both cameras have a frame rate of 30 frames per second. After the experiment, the drivers' eye glance locations (either on or off the road) during the button search-and-press tasks were manually coded from the video by a human data reducer. At the start of the task, multiple buttons in a grid layout were presented on the touch screen (from a black screen). Each button has a unique text label generated from a pool of common acronyms for media sources (e.g., MP3, FM1). The goal for the drivers was to search for the target button which was labeled with "USB", and then press the target button. The target button was always the "USB" button for the entire experiment, and it was always included in each task trial. The order of the buttons in each design was randomly generated before the experiment so that the drivers could not have expectations of the target button's location prior to each task trial. However, the same randomly generated designs were used for all subjects.

2.2. Experimental design

There were three independent variables in the experiment: (1) parked or driving, as described in detail later, (2) quantity of buttons (4 buttons in a 2x2 layout, 8 buttons in a 2x4 layout, or 15 buttons in a 3x5 layout), and (3) button sizes (small (66 pixels or 14 mm side length), medium (108 pixels or 24 mm), or large (150 pixels or 33 mm)). The three layout configurations (2x2, 2x4, and 3x5) was chosen to resemble the common grid layout in many production in-vehicle infotainment systems with a landscape touch screen (e.g., Chevrolet MyLink (n.d.)). The medium button (24 mm) was set to be larger than the recommended minimal values (e.g., 19 mm by Monterey Technologies Inc. (1996)). The large button (33 mm) was set to be about double the size of the medium button, while the small button (14 mm) was set to be smaller than the recommended minimal values.

A full-factorial within-subject design was used for the experiment. The nine designs (= 3 levels of quantity of buttons x 3 levels of button size) are illustrated in Figure 2. A total of four trials were used for each design combination. That gives a total of 72 trials (= 9 designs x 4 trials x 2 (parked or driving)) for each participant.

INSERT FIGURE 2 HERE

The labels for the buttons were selected from common acronyms of media sources. For the 2x2 layout, the four labels are "AM1", "FM1", "USB", and "DVD". For the 2x4 layout, the same labels from the 2x2 design were used with the addition of “AM2”, “FM2”, “MP3”, and “CD”. For the 3x5 layout, the same labels from the 2x4 design were used with the addition of “AM3”, “FM3”, “TV”, “SD”, “Tape”, “AV In”, and “Scan”. The size of the labels was considered legible on the touch screen with a font size of 20 pixels (or 4.5 mm) and was kept consistent for all buttons in all conditions. All participants responded prior to the experiment data collection that they had no problem reading the button labels from their seated position. The horizontal and vertical spacing between buttons was also kept consistent in all designs (both at 4 pixels, or 0.9 mm). The default background color of the buttons was grey (hex color code: #404040). Once a button is pressed, its background color switches to dark green (hex color code: #1B402C) to provide a visual confirmation that the button was successfully pressed. The designs were created in HTML language and interpreted by a web browser in the full-screen mode.

In the parked condition, the participants were instructed to perform the button search-and-press tasks while the simulator is parked on roadside. In the driving condition, the participant was asked to drive the simulator, and at certain points, they were verbally instructed by the experimenter to perform the button search-and-press task when they believe it is safe to do so. The experimenter only instructed the participants to start the task when the vehicle was in the straight section of the road (i.e., not when the vehicle is entering, negotiating, or leaving a curve).

The dependent variables in this study include task completion time and a number of eye glance behavior measures that include total eyes-off-road time, number of off-road glances (simply referred to as “glances” for the rest of the paper) in each task trial, glance duration, and the occurrence of long glances. The task completion time is a common metric in usability testing that assesses the efficiency of a system (Rubin & Chisnell, 2008). In this study, the task completion time in the parked condition is the duration from the onset of the buttons on the touch screen to the time when a button is pressed. Note in the parked condition, the participants were instructed to always look at the touch screen. The definition of task completion time in the driving condition was borrowed from Tsimhoni and Green (2001), which defines it as the duration from the beginning of the first glance at the device to the end of the last glance during a task trial. An off-road glance duration was defined as from the time when the driver's eyes start to move away from the road to the time when the eyes move back to the road. Note this is slightly more conservative compared to the glance duration definition by SAE J2396 (2000) and ISO 15007 (2014), in which only the prior gaze transition was included. We used our definition because we were only interested in the time durations when the drivers' eyes were off the road, rather than differentiating off-road glances at different locations, in which the subsequent gaze transition shall not be included as it would be counted twice when summing up the glance durations. The total eyes-off-road time is the cumulative time when the driver's eyes were away from the road when performing the task (i.e., the summation of all the off-road glance durations during the task). Note only the glances that initiated prior to pressing the button were included. This was to exclude the additional glances occasionally made by the drivers after pressing the button presumably for confirming the button activation. Studies have shown that besides the total eyes-off-road time, the individual glance duration, especially the long glances are particularly related to road crashes and near-crashes (NHTSA, 2012). Horrey and Wickens (2007) have suggested that compared with the average glance durations, the tail end (i.e., larger values) of the glance duration distribution is more related to crash risks. Different threshold values for long glances have been used in previous studies and guidelines (e.g., 1.6 s by Wierwille, 1993b and Horrey & Wickens, 2007, and 2.0 s by AAM, 2006 and NHTSA 2012). In this study, we chose 2.0 s as the definition threshold for long glances.

2.3. Procedure

Once the participants arrived at the laboratory, they were first asked to complete the consent form. Then they were asked to sit in the driving simulator. They were asked to adjust the seat position to make sure they can reach the touch screen with the arm only (i.e., without having to lean the torso). The participants were given an introduction using slides to ensure that they understood the tasks they were about to perform. This was followed by a practice session for both the button search-and-press task (at least three times) and the simulator driving (about five minutes). Then the data collection started. Counterbalancing was used for controlling the order effects of the within-subject design. 10 (50%) participants started with the parked condition, while 10 (50%) started with the driving condition. The order of the 36 trials (= 9 designs x 4 replications) was randomly placed in both the parked and driving condition. However, the same order was used for all participants.

3. Results

3.1. Task completion time

The task completion time corresponding to the nine designs in both the parked and driving conditions is shown in Figure 3. As can be seen, the data seem to be positively skewed (i.e., with a longer tail to larger values). The normality tests (Shapiro-Wilk, same hereafter) show that in general, the task completion time in each group did not follow a normal distribution. In the parked condition, the data violate normality in all designs (all $p < 0.05$) except the three 2x2 designs (small buttons: $p = 0.51$; medium buttons: $p = 0.65$; and large buttons: $p = 0.17$) and the 2x4 design with large buttons ($p = 0.21$). In the driving condition, the data violate normality in all designs (all $p < 0.05$). After applying a natural log transformation to the task completion time, the data became normally distributed for most of the groups ($p > 0.05$). However, the data still violate normality for the 3x5 design with large buttons in the parked condition ($p < 0.05$), and for the 2x4 design with medium buttons, and 3x5 design with small or medium buttons in the driving condition (all $p < 0.05$).

INSERT FIGURE 3 HERE

Levene's test for homogeneity of variance shows that both the original and log-transformed data do not have equal variance among the groups (all $p < 0.001$). Since both the normality and homogeneity of variances assumption of ANOVA (analysis of variance) were violated, even for the log-transformed data, a rank-based nonparametric Aligned Rank Transform procedure (Wobbrock et al., 2011) was used. The procedure enables the use of ANOVA after alignment and ranking to examine both the main and interaction effects of the experiment factors. The results show that there was no significant three-way interaction among the factors of driving, quantity of buttons, and button size ($F(4, 1422) = 0.709$, $p = 0.586$, $\eta^2 = 0.002$), and no significant two-way interaction between the quantity of buttons and button size ($F(4, 1422) = 1.048$, $p = 0.381$, $\eta^2 = 0.003$). However, there was significant two-way interaction between the parked/driving condition and both the quantity of buttons ($F(2, 1422) = 18.548$, $p < 0.001$, $\eta^2 = 0.025$) and button size ($F(2, 1422) = 3.358$, $p = 0.035$, $\eta^2 = 0.005$). Simple main effects analysis show that the task completion time in the driving condition was significantly longer compared to the parked condition for any quantity of buttons or button size (all $p < 0.001$). In the parked condition, the designs with more buttons were associated with significantly longer task completion time ($F(2, 711) = 194.533$, $p < 0.001$, $\eta^2 = 0.354$). Post hoc tests (Games-Howell for unequal variances, same hereafter) further show that the task completion time was significantly longer from the 2x2 design to the 2x4 design, and then to the 3x5 design (all $p < 0.001$). The effect of the button size on the task completion time in the parked condition was not significant ($F(2, 711) = 1.884$, $p = 0.153$, $\eta^2 = 0.005$). In the driving condition, the designs with more buttons were associated with significantly longer task completion time (F(2, 711) = 187.202, $p < 0.001$, $\eta^2 = 0.345$). Post hoc tests further show that the task completion time got significantly longer from the 2x2 design to the 2x4 design, and then to the 3x5 design (all $p < 0.001$). The designs with smaller buttons were associated with significantly longer task completion time (F(2, 711) = 6.845, $p < 0.01$, $\eta^2 = 0.019$). Post hoc tests further show that the task completion time in small-button design was significantly long compared to both the medium-button design ($p < 0.01$) and large-button design ($p < 0.05$). However, there was no significant difference between the medium- and large-button design ($p = 0.503$).

The geometric mean of the task completion time was also examined (shown in Figure 4) as arguably a better estimate of human task time compared to the arithmetic mean or median (Sauro & Lewis, 2010). The geometric mean was calculated by transforming the mean of the log-transformed data back to the original scale by exponentiation. Figure 4a shows the significant effects that driving elicited longer task completion time regardless of the number of buttons, and more buttons elicited longer task completion time in both the parked and driving condition. It also shows that 3x5 design had the largest increase of task completion time from parked to driving, while the 2x2 design had the least increase. Figure 4b shows that in the parked condition the button size did not have a significant effect on task completion time. However, in the driving condition, the task completion time with small buttons increased significantly more compared to the medium- or large-button design. The geometric means, arithmetic means, and median of task completion time and the total eyes-off-road time are shown in Table 1.

INSERT FIGURE 4 HERE

INSERT TABLE 1 HERE

3.2. Total eyes-off-road time

The total time when the drivers' eyes were off the road was extracted for the nine designs in the driving condition. The total eyes-off-road time is shown in Figure 5, grouped by the nine designs. Similar to the task completion time, it could be observed that the data also seem to be positively skewed. The normality tests show that the data violate normality in all designs (all $p < 0.05$) except the two 2x2 designs (small buttons: $p = 0.53$; medium buttons: $p = 0.13$). A natural log transformation was applied to the task completion time. However, the data still violate normality for the 2x4 design with medium buttons and 3x5 design with small buttons (both $p < 0.05$).

INSERT FIGURE 5 HERE

The same Aligned Rank Transform procedure was applied. The results show that there was no significant two-way interaction between the quantity of buttons and button size ($F(4, 711) = 0.558$, $p = 0.693$, $\eta^2 = 0.003$). However, significant main effects were found for both the quantity of buttons ($F(2, 711) = 203.477$, $p < 0.001$, $\eta^2 = 364$) and button size ($F(2, 711) = 7.502$, $p < 0.01$, $\eta^2 = 0.021$). Post hoc tests further show that the total eyes-off-road time was significantly longer in the 3x5 design compared to both the 2x2 and 2x4 design, and the total eyes-off-road time was significantly longer in the 2x4 design compared to the 2x2 design (all $p < 0.001$). In terms of button size, the total eyes-off-road time was significantly longer in the small-button design compared to both the medium-button design ($p < 0.01$) and the large-button design ($p < 0.05$). However, there was no significant difference between the medium- and large-button design ($p = 0.768$).

3.3. Number of glances

As expected it may take the drivers more than one glance to complete the task. Figure 6 shows the percentage of task trials in each of the nine designs that were completed with one, two, and more than two glances. Fisher's exact tests show that there was a significant association (at the 1%

significance level) between the quantity of buttons and the number of tasks that were completed with multiple glances. However, there was no significant association between the button size and the number of tasks that were completed with multiple glances. Specifically, for the 2x2 design, about 73-80% of the trials were completed with a single glance, and the rest completed with two glances; For the 2x4 design, about 51-59% of the trials were completed with a single glance, about 35-39% with two glances, and about 6-13% with three or more glances; For the 3x5 design, about 33-43% of the trials were completed with a single glance, about 34-48% with two glances, and about 20-29% with three or more glances.

INSERT FIGURE 6 HERE

3.4. Glance duration

The duration of the glances each time the drivers moved their eyes off the road was shown in Figure 7, grouped by the nine designs. The normality tests show that the data violate normality in all designs (all $p < 0.05$) except the three 2x2 designs with small (/medium/large) buttons ($p = 0.43$ (/0.27/0.30)) and the 2x4 design with medium buttons ($p = 0.08$). After applying a natural log transformation, the data still violate normality except the 2x4 design with small buttons ($p = 0.54$) and medium buttons ($p = 0.07$), and the 3x5 design with medium buttons ($p = 0.15$) and large buttons ($p = 0.22$).

INSERT FIGURE 7 HERE

The same Aligned Rank Transform procedure was applied. The results show that there was no significant two-way interaction between the two design factors ($F(4, 1134) = 0.464$, $p = 0.762$, $\eta^2 = 0.002$). The main effect of quantity of buttons was not significant ($F(2, 1134) = 2.836$, $p = 0.059$, $\eta^2 = 0.005$). Post hoc tests show that there was no significant difference in glance duration between the 2x4 design and both the 3x5 design ($p = 0.283$) and 2x2 design ($p = 0.772$). The main effect of button sizes was significant ($F(2, 1134) = 7.267$, $p < 0.01$, $\eta^2 = 0.013$). Post hoc tests show that the glance duration was significantly longer in the small-button design compared to the large-button design ($p < 0.05$). However, there was no significant difference between the medium-button design and either the small-button design ($p = 0.207$) or the large-button design ($p = 0.745$).

The long glances (defined as a single glance that lasted longer than 2 s, from NHTSA's 12/2 Rules (NHTSA, 2012)) were also highlighted in Figure 7 as red pentagrams above the 2-second horizontal line. From a total of 720 task trials in the driving condition, a total of 223 long glances were observed. The occurrence of long glances in each design was also summarized in Table 1. Notably, 13% (/ 30% / 46%) of the trials in the 2x2 (/ 2x4 / 3x5) design had long glances, and 34% (/ 28% / 27%) of the trials in the small (/ medium / large) button design had long glances. Fisher's exact tests show that there was a significant association between the quantity of buttons and the number of task trials with long glances (all $p < 0.001$), but no such association between the button size and the number of trials with long glances.

Table 2 summarizes the findings of task completion time and eye glance metrics, presented in terms of the impact of the quantity of buttons and button size.

INSERT TABLE 2 HERE

3.5. Long glances and the number of glances

In this section, we examined the relationship between the occurrence of long glances and the number of glances used to complete the task. This analysis aimed to get some insights into the probable causes of long glances associated with drivers' task interleaving strategies. First, the glances were separated by whether the tasks were completed with a single glance or multiple glances. Figure 8 shows the distributions (aggregated for all nine designs) of the glance durations of task trials that were completed with a single glance (termed "single-glance trials" hereinafter) and task trials that were completed with multiple glances (termed "multiple-glance trials" hereinafter).

INSERT FIGURE 8 HERE

It was noticeable that the glance durations in the single-glance trials and multiple-glance trials have different distributions. The normality tests show that both distributions violate normality (both $p < 0.001$). The Kruskal-Wallis test (the nonparametric equivalent of one-way ANOVA) shows that the glance durations of the single-glance trials were significantly longer compared to the multiple-glance trials ($\chi2(1) = 413.058$, $p < 0.001$). When focused specifically at long glances, single-glance trials accounted for 74% (164) of a total of the 233 long glances, although they only accounted for 57% (411) of a total of the 720 task trials. On the other hand, multiple-glance trials accounted for only 26% (59) of the long glances, although they accounted for 43% (309) of the total task trials. It seems that long glances were overrepresented in the single-glance trials. Indeed, Fisher's exact test shows a significant association between the single-glance trials and long glances ($p < 0.001$).

INSERT FIGURE 9 HERE

Figure 9 shows the glance duration of the nine designs, each separated by the single-glance trials and multiple-glance trials. For the 2x2 designs, all (28 out of 28) long glances were from single-glance trials, which accounted for 78% (186) of the task trials; for the 2x4 designs, 91% (62 out of 68) long glances were from single-glance trials, which accounted for 56% (134) of the task trials; and for the 3x5 designs, 58% (69 out of 120) long glances were from single-glance trials, which accounted for 38% (91) of the task trials. Fisher's exact tests show a significant association between the single-glance trials and long glances in each of the case (all $p < 0.01$). It was also to note that 15% (/46% / 76%) of the single-glance trials in the 2x2 (2x4 / 3x5) designs resulted in long glances.

4. Discussion

This paper investigated the effects of two basic design parameters of touch screen user interfaces, namely the quantity of buttons and their size, on drivers' eye glance behavior. A driving simulator experiment was conducted, in which drivers were asked to perform a button search-and-press task on an in-vehicle touch screen while driving. A number of metrics were examined that include task completion time, TEORT, glance duration, and the occurrence of long glances. Although a normal distribution was often assumed for the eye glance data in previous related studies during the data analysis, our results show that the time completion time, total eyes-off-road time, and glance duration are generally not normally distributed (even after a log transformation) but positively skewed (i.e., with a longer tail to larger values). For this reason, a nonparametric rank-based procedure (Aligned Rank Transform) was applied to the data before the use of ANOVA. The reason for choosing this method over the traditional nonparametric tests (e.g., Kruskal-Wallis test) is that the Aligned Rank Transform procedure enables us to examine not only the main effects but also the interaction effects.

Task completion time: The results show that the designs with a larger number of buttons were associated with longer task completing time in both the parked and driving conditions. This is consistent with the majority of previous findings that the visual search efficiency decreases with an increasing number of distractors (e.g., Wolfe, 2007). The significant two-way interaction between driving and the quantity of the buttons shows such an effect was even stronger when driving was involved (Figure 4a). This result reconfirms the necessity of testing the usability of in-vehicle systems in the context of driving, rather than in a single task condition where users can devote all their attention to the task. The button size did not show a significant effect on the task completion time in the parked condition. This may be because the effect of the button size on the time for the hand to press the button as predicted by Fitts's law was simply overwhelmed by the large variance of the time associated with searching the target button. However, there was a significant two-way interaction between driving and the button size. The task completion time associated with small buttons was particularly worse compared to the medium- or large-button design (Figure 4b). In this study, the small buttons were designed with a size of 14 mm, which was the only one of the three sizes that are smaller than the recommended minimum button size (e.g., 22 mm by Greenstein (1997)). This result clearly reconfirms the previous findings and guidelines that the user interface designers should be cautious with small buttons especially in the context of driving. It is also interesting to see the null effect of medium buttons (24 mm) versus large buttons (33 mm) on the task completion time.

Total eyes-off-road time (TEORT): The data of TEORT are similar to the task completion time in the driving condition, as shown in the summary statistics in Table 1. This suggests that the button search-and-press task in this experiment relies heavily on drivers' visual attention. The designs with a larger number of buttons were associated with a larger TEORT. This is consistent with previous findings (e.g., Kujala & Salvucci, 2015; Large et al., 2017). The small-button designs were associated with significantly larger TEORT compared to medium- and large-button designs. However, there was no significant effect between the medium- and large-button designs. In fact, we found no significant effect between the medium- and large-button designs in all the eye glance metrics including TEORT, glance duration, and the occurrence of long glances (see Table 2). The maximum TEORT recorded in the entire experiment was 11.8 s, which is just under the 12.0 s from the NHTSA's 2/12 Rule. It seems this value can also be estimated using the maximum task completion time in the single task condition (7.1 s), which roughly represents the maximum time needed to complete the task without driving. When driving is involved, if we assume a driver keeps the mean off-road glance duration less than 1.6 s (upper bound value from Dingus et al., 1989), a task that needs 7.1 s to complete requires five glances (rounded up). Each glance requires two eye saccades from and back to the road. If each transition takes 0.5 seconds (upper bound value from SAE J2396, 2000), the maximum TEORT can be estimated as 12.1 (= 7.1 + 5 x 2 x 0.5) second, which is close to the observed maximum TEORT of 11.8 s.

Use of multiple glances: The results show that with a larger number of buttons, the drivers tended to use multiple glances to complete the task. 60% (/45% / 25%) of trials in the 3x5 (/2x4 / 2x2) design were completed with multiple glances. This was expected as the task completion time data shows the tasks may take more than a few seconds to complete even without driving (for the 3x5 designs the geometric means of the task completion time are all longer than 2 s in the parked condition). To perform the task while driving, the drivers may have to temporarily suspend the task and switch the visual attention back to the road, make corrective maneuvers if needed, and then make another off-road glance and resume the secondary task.

Glance duration: Unlike the TEORT, the analysis shows that in general there were no significant effects of the two factors on the duration of individual glances (except two cases of (a) 2x2 vs 3x5 and (b) small- vs large-buttons, both of which have a small effect size of $\eta^2 = 0.01$). The geometric mean of the glance duration ranges from 1.30 to 1.56 s (see Table 1). These results are consistent with previous findings that drivers try to keep their mean off-road glance durations in between fairly static limits (e.g., 0.6-1.6 s from Wierwille, 1993a).

Long glances: Long glances were further analyzed given their demonstrated association with crash risk (Klauer et al., 2006; Horrey & Wickens, 2007). The result shows that the number of buttons had a significantly increasing effect on the occurrence of long glances. In the 3x5 designs, 40% (large buttons) to 54% (small buttons) of the task trials included long glances. In the 2x4 design, the percentages are still substantial ranging from 24% to 35%. The only seemingly "acceptable" designs are the two 2x2 design with medium or large buttons, in which 6% (large buttons) and 14% (medium buttons) of the task trials included long glances. The results imply that particular caution needs to be exercised when designing an interface with a large number of buttons for such tasks. The design should avoid displaying more than 8 unfamiliar items (2x4 design) as it seems to already induce too many long glances for safe driving. A total of 4 items (2x2 design) seem to be close to the acceptable maximum number of items for such tasks. These results are in line with the study by Kujala and Salvucci (2015) that the 6-item list menu was the only design (among the 6-, 9-, and 12-item grid and list menu designs in their experiment) that would pass the NHTSA over-2-second glance criterion. The results are also in line with the study by Kujala and Saariluoma (2011) that moving from 6 to 9 display items significantly increases the maximum glance duration. Also interestingly, there is a similar effect of text size on an in-vehicle display in a study by Crundall et al. (2016), which found that larger text sizes were associated with a higher number of relatively short glances, whereas smaller text led to a higher number of long glances.

Task interleaving strategies: To get some insights into the probable causes of long glances, we further analyzed the relationship between the long glances and the number of glances used in each task trial. The results show a significant association between the occurrence of long glances and the task trials that were completed with a single glance. Single-glances were used in 57% of all task trials but accounted for 74% of all long glances. It suggests that the driver's tendency of using a single glance rather than breaking it down to multiple glances was a major cause of long glances. To analyze the drivers' task interleaving strategies (which determines their eye glance behavior), the button search-and-press task can be decomposed to a subtask of searching for the target button, and a subsequent subtask of reaching and pressing the located target button. The subtask boundary is associated with an explicit motor cue of moving the hand towards the target button. However, previous works show that the task interleaving can be affected by both the task structural constraints and temporal constraints (i.e., accumulated uncertainty of the roadway) (Kujala & Salvucci, 2015; Lee & Lee, 2017). The influence of the task structural constraints is related to the task resumability or cost of interruption (Noy et al., 2004; Iqbal & Bailey, 2006; Burns et al., 2010).

In our experiment with the button search-and-press task, there seems to be a relatively high cost of interruption at the subtask boundary. To temporarily suspend the in-vehicle task and resume it at a later time, it requires the drivers' additional mental resource to (1) retain the target-button location in the working memory during driving, (2) retrieve the target button location before the next off-road glance, (3) direct the visual attention to the retrieved button location on the screen, and (4) re-identify the button label and confirm it is the target. The driver may also potentially have to redo the search in case the previously found target button cannot be successfully located. With these additional costs of task switching at the subtask boundary, it seems the drivers prefer to continue with the subtask of reaching and pressing the target button within the same glance, even though it may lead to a long (2+ seconds) glance. This finding further confirms the importance of task resumability and the necessity of careful task analysis in the context of multitasking when designing in-vehicle user interfaces. Certain subtask combinations (e.g., in our case, a visual search subtask followed by a reaching subtask that requires the location of search target) that are associated with low resumability may lead to extended eyes-off-road time and ultimately compromise the driving safety, as the drivers are inclined not to switch tasks at these subtask boundaries.

As briefly described in the Introduction section, our study is different from a relevant study by Large et al., (2017) in that no additional instructions were given to the drivers in terms of hand movement, while in Large's experiments, the drivers were specifically instructed to keep hands on the steering wheel until they have located the target button. In our experiment, we observed that various strategies seem to be used in terms of the time sequence of executing the searching and pointing subtasks. On one end of the spectrum, the drivers may move the hand from the steering wheel to the target button with a quick and smooth movement, suggesting that the target button may have already been located prior to the initiation of the hand movement. On the other end of the spectrum, the drivers may start moving the hand towards the touch screen immediately after the start of the first glance. Then the driver had his/her finger hovered over the screen while seemingly still searching for the target button. There were also cases in between where the driver's hand made probable pauses when moving towards the touch screen, suggesting the driver may have used more than one pre-programmed movements of the hand and was conducting the visual search between the movements. Since the hand movements were not directly measured and hard to quantify from the video data, they were not processed in this study. However, these observations do seem to suggest that Fitts's Law may not be directly applicable in this case, as the target location may still be unknown at the early phase of the hand movements. In the future, we may further investigate how these types of task execution strategies are related to the interface and task designs.

It is also worth to note that the task in this study is a rather simple task that involved only a single button pressing. In most production in-vehicle systems it is not uncommon to require the drivers to press a series of buttons or to have multiple interactions. For example, to change the radio station it may require the driver first to go to the entertainment screen from the home screen, then select the media source from a list of options (e.g., FM, AM), and then select a preset radio station (or even type in the radio frequency on a number pad). It is challenging to design an in-vehicle system that not only supports a large number of functions and complex interactions but also is safe to use while driving. Nonetheless, the findings from this study have implications for user interface designers to tackle some of the challenges by designing subtask boundaries and cues that encourage task interleaving, designing the interactions and tasks that improve task resumability, and considering innovative interaction styles such as haptic cues (Burnett & Porter, 2001), gesture controls (Jeong & Liu, 2017), and multimodal interactions (Chen et al., 2016).

There are several limitations in this study. First, the experiment was conducted in a fixed-base driving simulator. The lack of physical movement and vibration of the vehicle may potentially affect the driver's performance of the button-press task. Humans use both visual and proprioceptive feedback to execute body movement such as programming arm trajectories (Ghez et al., 1990). In a real-world driving environment, a driver's proprioceptive system may be negatively affected by the whole-body vibrations from a moving and vibrating vehicle (Gauthier et al., 1981). However, this effect was not assessed in our study. There are also other factors that were not accounted for in this study but may potentially affect the driver' task performance and eye glance behavior. Such factors include driver characteristics (e.g., skill and comfort level of driving the simulator, risk-taking attitudes, and manual dexterity), driving environment (e.g., vehicle speed, information density of the roadway (Senders et al., 1967)), and the mounting position of the display (Horrey & Wickens, 2004).

Secondly, a constant and small spacing between the buttons (0.9 mm, edge to edge distance between two adjacent buttons) was used for all designs, rather than to keep the separation of the button labels constant. The choice was made to avoid the large spacing for the small- and medium-button designs (which would occur to accommodate the large-button designs if the button label separation was kept constant), and to resemble a more realistic touch screen design, in which small spacing between buttons was commonly used in a grid layout (e.g., Chevrolet MyLink (n.d.)), probably due to limited screen space and aesthetic reasons. However, in the tested designs, the button label separation is associated with button size. Small-button designs are associated with the smallest text separation. And the texts get further apart by increasing the button size. This difference in text separation may affect the capability of the drivers to simultaneously encode multiple items during the visual search, and consequently affect the number of fixations and glances needed to find the target. Our results show that in general the designs with different button size, especially between medium and large buttons, had a null effect on the task completion time and all the eye glance metrics. However, we were not able to further examine whether it was related to the button size, text separation, or a combination of both.

Thirdly, this work focused on two basic design parameters of the quantity and size of buttons. There are other important design factors for in-vehicle touch screen systems that may be worth investigating, such as the display organization (e.g., logical or conceptual grouping of items, Wickens & Carswell, (1995)), local and global crowding effect (Vlaskamp & Hooge, 2006), task-relevance of items (i.e., whether an item is needed for the task at hand, Alexander et al., (2008)), and display personalization (Normark, 2015). In addition, the button labels in the experiment were randomly arranged in each task trial. However the buttons in real-world production systems are usually in fixed locations and arranged in a certain structure (e.g., alphabetical order, the frequency of use). This enables the users to obtain knowledge about the button layout over time through learning. This top-down knowledge may potentially facilitate the visual search task by pinpoint or at least narrow down the search area. However, this learning effect was not examined in this work. The findings of this study may only be applied as a worst case scenario representing novice users who do not have any pre-knowledge or understanding of the button layout.

5. Conclusion

This paper investigates the effects of the quantity and size of touch screen buttons and the task interleaving strategies on drives' eye glance behavior. The results show that the number of buttons has an increasing effect on task completion time, TEORT, and long glances. However, in general, no such differences were found for button sizes. Further analysis indicates that long glances were strongly associated with drivers completing the task with a single glance. It seems to suggest that a major cause of long glances is that drivers are reluctant to switch the task back to driving at subtask boundaries that are probably associated with high cost of interruption. These findings confirm the importance of task resumability for in-vehicle user interfaces and have implications that careful task analysis needs to be conducted in the context of multitasking. Certain subtask combinations, such as a visual search followed by pressing the search target, may discourage task interleaving and ultimately compromise driving safety.

References

AAM (Alliance of Automobile Manufacturers). (2006). Statement of Principles, Criteria, and Verification Procedures on Driver Interactions with Advanced In-Vehicle Information and Communications Systems, Draft Version 2.1 (with updates), Driver Focus-Telematics Working Group.

Alexander, A. L., Stelzer, E. M., Kim, S.-H., & Kaber, D. B. (2008). Bottom-up and top-down contributors to pilot perceptions of display clutter in advanced flight deck technologies. *In Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting*, 1180-1184.

Bao, S., Guo, Z., Flannagan, C., Sullivan, J., Sayer, J. R., and LeBlanc, D., (2015). Distracted driving measures: A spectral power analysis. *Journal of the Transportation Research Record*, (2518), 68-72

Becker, S., Hanna, P., and Wagner, V., (2014). Human Machine Interface Design in Modern Vehicles, *Encyclopedia of Automotive Engineering*, Online publication, John Wiley & Sons.

Boyle, L.N., Lee, J.D., Peng, Y, Ghazizadeh, M., Wu, Y., Miller, E., and Jenness, J. (2013). Text reading and text input assessment in support of the NHTSA visual-manual driver distraction guidelines. Report No. DOT HS 811 820. Washington, D.C. NHTSA.

Brumby, D. P., Salvucci, D. D., & Howes, A. (2009). Focus on driving: How cognitive constraints shape the adaptation of strategy when dialing while driving. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1629-1638). ACM.

Burnett, G. E., & Porter, J. M. (2001). Ubiquitous computing within cars: designing controls for non-visual use. *International Journal of Human-Computer Studies*, 55(4), 521-531.

Burns, P., Harbluk, J., Foley, J. P., & Angell, L. (2010). The importance of task duration and related measures in assessing the distraction potential of in-vehicle tasks. *In Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 12-19). ACM.

Chen, Y., Tonshal, B., Rankin, J., & Feng, F. (2016). Development of an Integrated Simulation System for Design of Speech-Centric Multimodal Human-Machine Interfaces in an Automotive Cockpit Environment. *In ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. Paper No. DETC2016-59309.

Chevrolet MyLink (n.d.), Welcome to connectivity, Retrieved from http://www.chevrolet.com/owners/mylink-vehicle-technology

Crundall, E., Large, D. R., & Burnett, G. (2016). A driving simulator study to explore the effects of text size on the visual demand of in-vehicle displays. *Displays*, 43, 23-29.

Dandekar, K., Raju, B. I., & Srinivasan, M. A. (2003). 3-D finite-element models of human and monkey fingertips to investigate the mechanics of tactile sense. *Journal of Biomechanical Engineering*, 125(5), 682-691.

Dingus, T. A., Hulse, M. C., Antin, J. F., & Wierwille, W. W. (1989). Attentional demand requirements of an automobile moving-map navigation system. *Transportation Research Part A: General*, 23(4), 301-315.

Feng, F., Liu, Y., and Chen, Y., (2017). A Computer-Aided Usability Testing Tool for In-Vehicle Infotainment Systems, *Computers & Industrial Engineering*. 109, 313-324.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6), 381.

Gauthier, G. M., Roll, J. P., Martin, B., & Harlay, F. (1981). Effects of whole-body vibrations on sensory motor system performance in man. *Aviation, Space, and Environmental Medicine*, 52(8), 473-479.

Ghez, C., Gordon, J., Ghilardi, M. F., Christakos, C. N., & Cooper, S. E. (1990). Roles of proprioceptive input in the programming of arm trajectories. *In Cold spring harbor symposia on quantitative biology* (Vol. 55, pp. 837-847). Cold Spring Harbor Laboratory Press.

Green, P. (1999). Estimating compliance with the 15-second rule for driver interface usability and safety. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 43(1), 987-991

Greenberg, J., Tijerina, L., Curry, R., Artz, B., Cathey, L., Kochhar, D., ... & Grant, P. (2003). Driver distraction: Evaluation with event detection paradigm. *Transportation Research Record: Journal of the Transportation Research Board*, (1843), 1-9.

Greenstein, J.S. Pointing Devices. in M. Helander, T.L., P. Prabhu (ed). Handbook of Human-Computer Interaction, North-Holland, Amsterdam. (1997), 1317-1348.

Harbluk, J. L., Noy, Y. I., & Eizenman, M. (2002). The impact of cognitive distraction on driver visual behaviour and vehicle control (No. TP# 13889 E).

Harvey, C., Stanton, N. A., Pickering, C. A., McDonald, M., & Zheng, P. (2011). In-vehicle information systems to meet the needs of drivers. *International Journal of Human-Computer Interaction*, 27(6), 505-522.

Horrey, W., & Wickens, C. D. (2004). Driving and side task performance: the effects of display clutter, separation, and modality. *Human Factors*, 46(4), 611-24.

Horrey, W., & Wickens, C. D. (2007). In-vehicle glance duration: distributions, tails, and model of crash risk. Transportation Research Record: *Journal of the Transportation Research Board*, (2018), 22-28.

International Organization for Standardization, (2010). ISO 2575: Road vehicles - Symbols for controls, indicators and tell-tales.

International Organization for Standardization, (2014). ISO 15007-1: Road vehicles - Measurement of driver visual behaviour with respect to transport information and control systems - Part 1: Definitions and parameters.

Iqbal, S. T., & Bailey, B. P. (2005, April). Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. *In CHI'05 Extended Abstracts on Human Factors on Computing Systems* (pp. 1489-1492). ACM.

Iqbal, S. T., & Bailey, B. P. (2006). Leveraging characteristics of task structure to predict the cost of interruption. *In Proceedings of the SIGCHI conference on Human Factors in computing systems*, 741-750.

Janssen, C. P., Brumby, D. P., & Garnett, R. (2012). Natural break points: the influence of priorities and cognitive and motor cues on dual-task interleaving. *Journal of Cognitive Engineering and Decision Making*, 6(1), 5-29.

Jeong, H., & Liu, Y. (2017). Effects of touchscreen gesture's type and direction on finger-touch input performance and subjective ratings. *Ergonomics*, 1-12.

Jin, Z. X., Plocher, T., & Kiff, L. (2007). Touch screen user interfaces for older adults: button size and spacing, *Proceedings of the 4th International Conference on Universal Access in Human Computer Interaction: Coping with Diversity*, 933-941.

Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweeks, J. D., & Ramsey, D. J. (2006). The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. Report No. DOT HS 810 594.

Kujala, T. (2013). Browsing the information highway while driving: three in-vehicle touch screen scrolling methods and driver distraction. *Personal and Ubiquitous Computing*, 17(5), 815-823.

Kujala, T., & Saariluoma, P. (2011). Effects of menu structure and touch screen scrolling style on the variability of glance durations during in-vehicle visual search tasks. *Ergonomics*, 54(8), 716-732.

Kujala, T., & Salvucci, D. D. (2015). Modeling visual sampling on in-car displays: The challenge of predicting safety-critical lapses of control. *International Journal of Human-Computer Studies*, 79, 66-78.

Lansdown, T. C., Brook-Carter, N., & Kersloot, T., (2004). Distraction from multiple in-vehicle secondary tasks: vehicle performance and mental workload implications. *Ergonomics*, 47(1), 91-104.

Large, D. R., Burnett, G., Crundall, E., van Loon, E., Eren, A. L., & Skrypchuk, L. (2017). Developing predictive equations to model the visual demand of in-vehicle touchscreen HMIs. *International Journal of Human-Computer Interaction*, 1-14.

Lee, J. D. (2007). Technology and teen drivers. *Journal of Safety Research*, 38(2), 203-213.

Lee, J. Y., & Lee, J. D. (2017). Modeling the Effect of Subtask Boundaries on Driver Glance Behavior. *National Institutes of Health Consensus Development Conference Summaries* (June 2017).

Lee, J. Y., Gibson, M., & Lee, J. D. (2015). Secondary task boundaries influence drivers' glance durations. *In Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 273-280). ACM.

Lee, S. C., Hwangbo, H., & Ji, Y. G. (2016). Perceived Visual Complexity of In-Vehicle Information Display and Its Effects on Glance Behavior and Preferences. International Journal of Human-Computer Interaction, 32(8), 654-664.

Miyata, Y., & Norman, D. A. (1986). Psychological issues in support of multiple activities. User centered system design: New perspectives on human-computer interaction, 265-284.

Moacdieh, N., & Sarter, N. (2015). Display clutter a review of definitions and measurement techniques. *Human Factors*: 57(1), 61-100.

Monterey Technologies, Inc. Resource Guide for Accessibility: Design of Consumer Electronics. Draft Submitted to: EIA-EIF Committee on Product Accessibility, A Joint Venture of the

Electronic Industries Association and the Electronic Industries Foundation Washington, DC 20006 (1996)

National Center for Statistics and Analysis. (2017). Distracted Driving 2015. (Traffic Safety Facts Research Note. Report No. DOT HS 812 381). Washington, DC: National Highway Traffic Safety Administration.

Neisser, U., Novick, R., & Lazar, R. (1964). Searching for novel targets. *Perceptual and Motor Skills*, 19, 427-432.

NHTSA (2012). Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices. Washington, DC: National Highway Traffic Safety Administration.

Normark, C. J. (2015). Design and evaluation of a touch-based personalizable in-vehicle user interface. *International Journal of Human-Computer Interaction*, 31(11), 731-745.

Nothdurft, H. (2000). Salience from feature contrast: variations with texture density, *Vision Research*, 40(23), 3181–3200.

Noy, Y. I., Lemoine, T. L., Klachan, C., & Burns, P. C. (2004). Task interruptability and duration as measures of visual distraction. *Applied Ergonomics*, 35(3), 207-213.

Pavlidis, I., Dcosta, M., Taamneh, S., Manser, M., Ferris, T., Wunderlich, R., Akleman, E., & Tsiamyrtzis, P. (2016). Dissecting Driver Behaviors Under Cognitive, Emotional, Sensorimotor, and Mixed Stressors. *Scientific Reports*, 6.

Payne, S. J., Duggan, G. B., & Neth, H. (2007). Discretionary task interleaving: Heuristics for time allocation in cognitive foraging. *Journal of Experimental Psychology General*, 136(3), 370.

Peng, Y., Boyle, L. N., & Hallmark, S. L. (2013). Driver's lane keeping ability with eyes off road: Insights from a naturalistic study. *Accident Analysis & Prevention*, 50, 628-634.

Rubin, J., & Chisnell, D. (2008). Handbook of usability testing: howto plan, design, and conduct effective tests. John Wiley & Sons.

Salvucci, D. D., & Kujala, T. (2016). Balancing Structural and Temporal Constraints in Multitasking Contexts. *In CogSci 2016: Proceedings of the 38th Annual Conference of the Cognitive Science Society*, ISBN 978-0-9911967-3-9. Cognitive Science Society.

Sauro, J., & Lewis, J. R. (2010). Average task times in usability tests: what to report? *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2347-2350). ACM.

Senders, J. W., Kristofferson, A. B., Levison, W. H., Dietrich, C. W., & Ward, J. L. (1967). The attentional demand of automobile driving. Highway research record, (195).

Society of Automotive Engineers, (1993). SAE J2119: Manual Controls for Mature Drivers.

Society of Automotive Engineers, (2000). SAE J2396: Definitions and Experimental Measures Related to the Specification of Driver Visual Behavior Using Video Based Techniques.

Society of Automotive Engineers, (2004). SAE J2364: Navigation and Route Guidance Function Accessibility While Driving.

Stevens, A., Quimby, A., Board, A., Kersloot, T., & Burns, P. (2002). Design guidelines for safety of in-vehicle information systems. TRL Limited.

Tsimhoni, O., & Green, P. (2001). Visual demand of driving and time execution of display intensive in-vehicle tasks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 45, 1586-1590.

Tsimhoni, O., Smith, D. & Green, P. (2004). Address entry while driving: Speech recognition versus a touch-screen keyboard. *Human Factors*, 46(4), 600-610.

Vlaskamp, B. N., & Hooge, I. (2006). Crowding degrades saccadic search performance. *Vision Research*. 46 (3), 417-25.

Wickens, C. D., & Carswell, C. M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors*, 37, 473-494.

Wierwille, W. W. (1993a). Demands on driver resources associated with introducing advanced technology into the vehicle. *Transportation Research Part C: Emerging Technologies*, 1(2), 133-142.

Wierwille, W. W. (1993b). Visual and manual demands of in-car controls and displays. In B. Peacock & W. Karwowsk (Eds.), *Automotive Ergonomics*, London: Taylor and Francis, 299-320.

Wierwille, W. W., & Tijerina, L. (1998). Modelling the relationship between driver in-vehicle visual demands and accident occurrence, *Vision in Vehicles* - VI, 233-244.

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The aligned rank transform for nonparametric factorial analyses using only anova procedures. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 143-146). ACM.

Wolfe, J. M. (2007). Guided Search 4.0: Current progress with a model of visual search. In W. Gray (Ed.), *Integrated Models of Cognitive Systems*. New York: Oxford, 99-119.

Yoon, S. H., Lim, J. H., & Ji, Y. G. (2015). Perceived visual complexity and visual search performance of automotive instrument cluster: a quantitative measurement study. *International Journal of Human-Computer Interaction*, 31(12), 890-900.

Figure 1. A button search-and-press task on a touch screen in a fix-based driving simulator

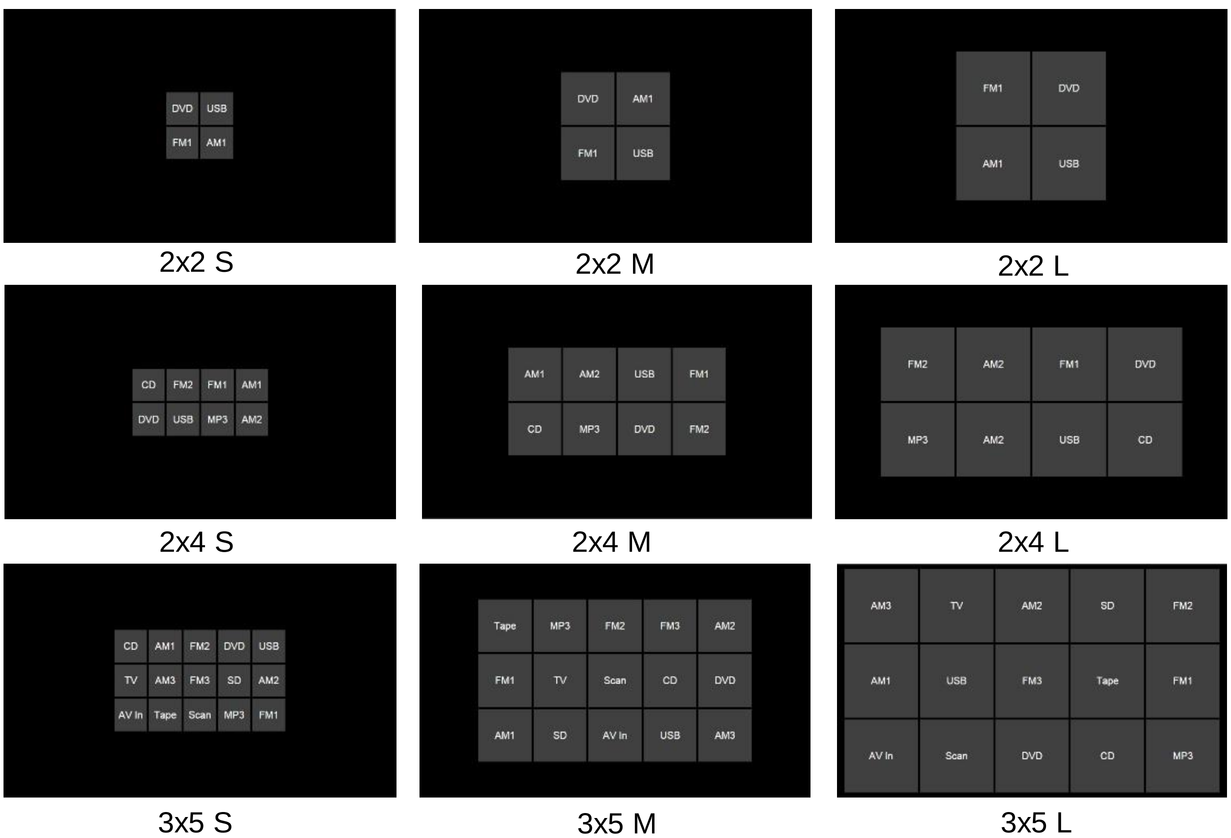

Figure 2. Nine designs with three levels of quantity of buttons and three levels of button sizes. Note four trials were used for each design with randomized label order in a parked or driving condition. This figure only shows one variant of each design.

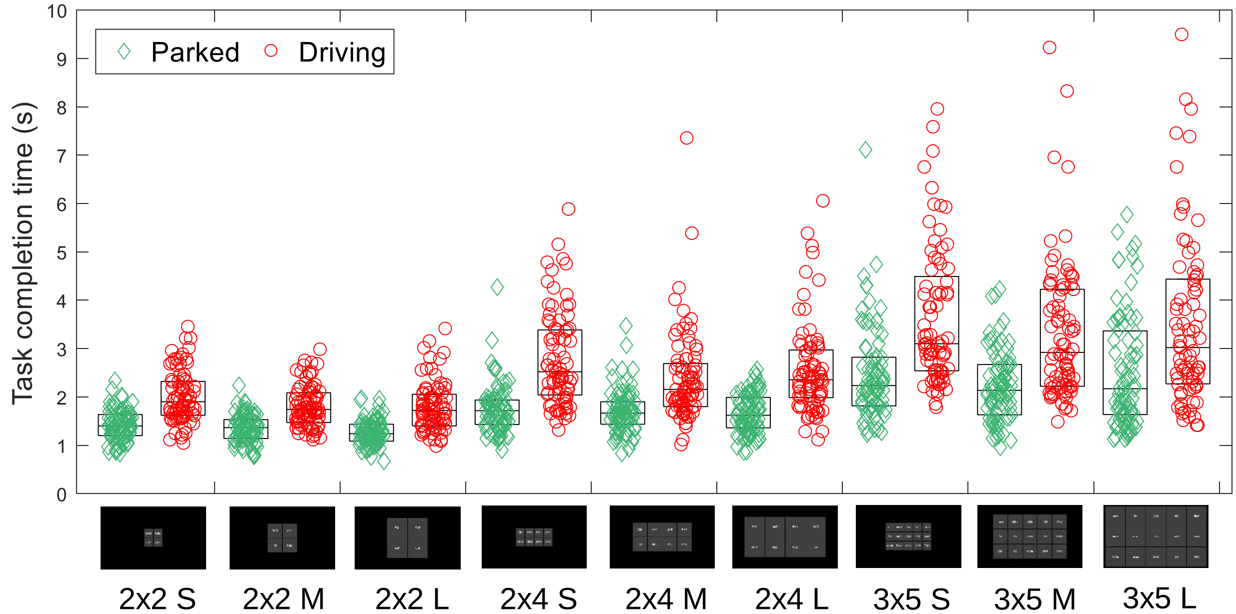

Figure 3. Task completion time grouped by the nine designs and parked/driving. On each box, the central mark indicates the median, and the bottom and top edges indicate the 25th and 75th percentiles, respectively.

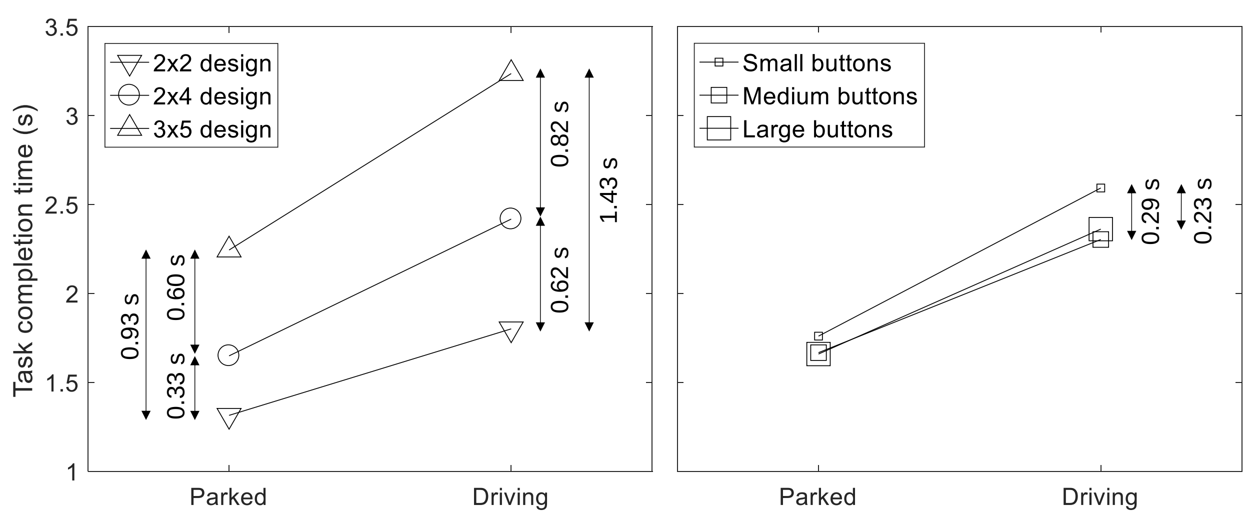

Figure 4. The effect of driving and the two design parameters on task completion time (geometric means)

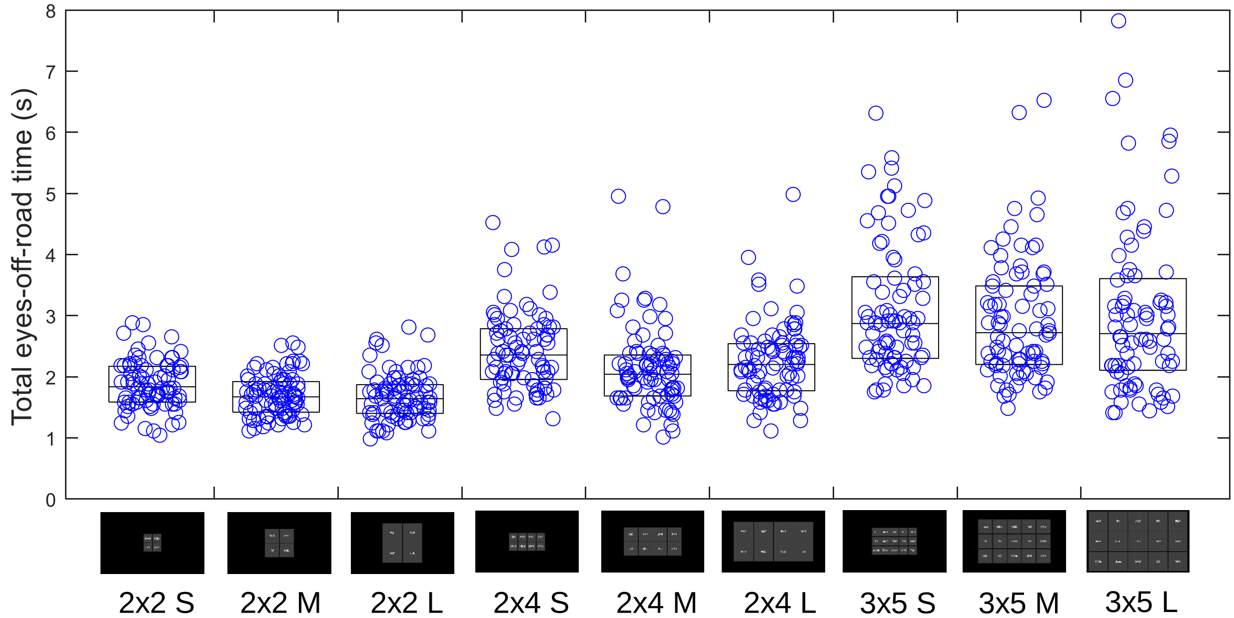

Figure 5. Total eyes-off-road time for the nine designs. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively.

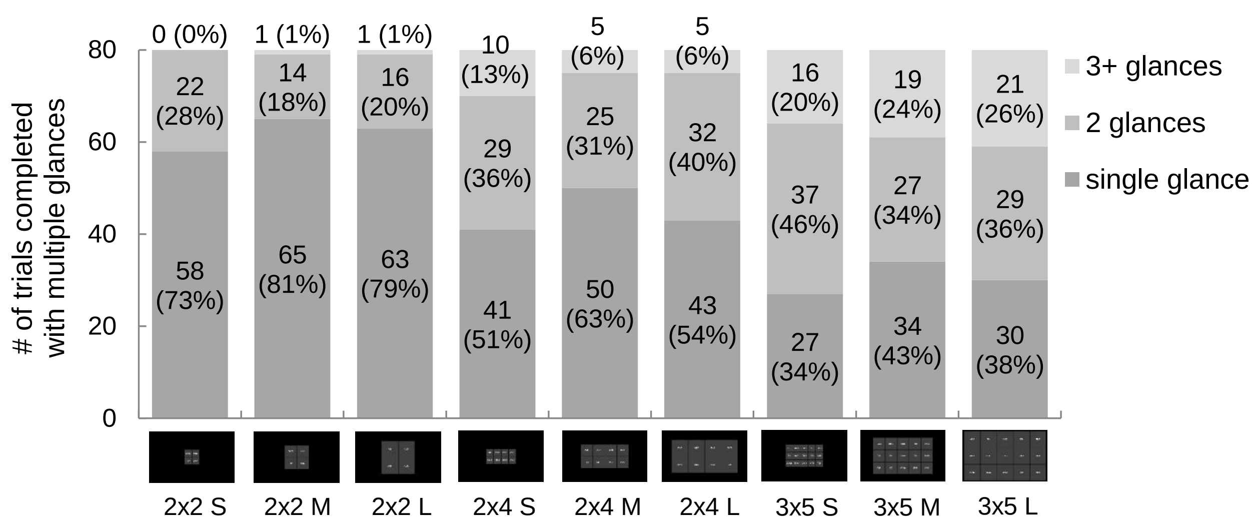

Figure 6. Number of task trials with single and multiple glances

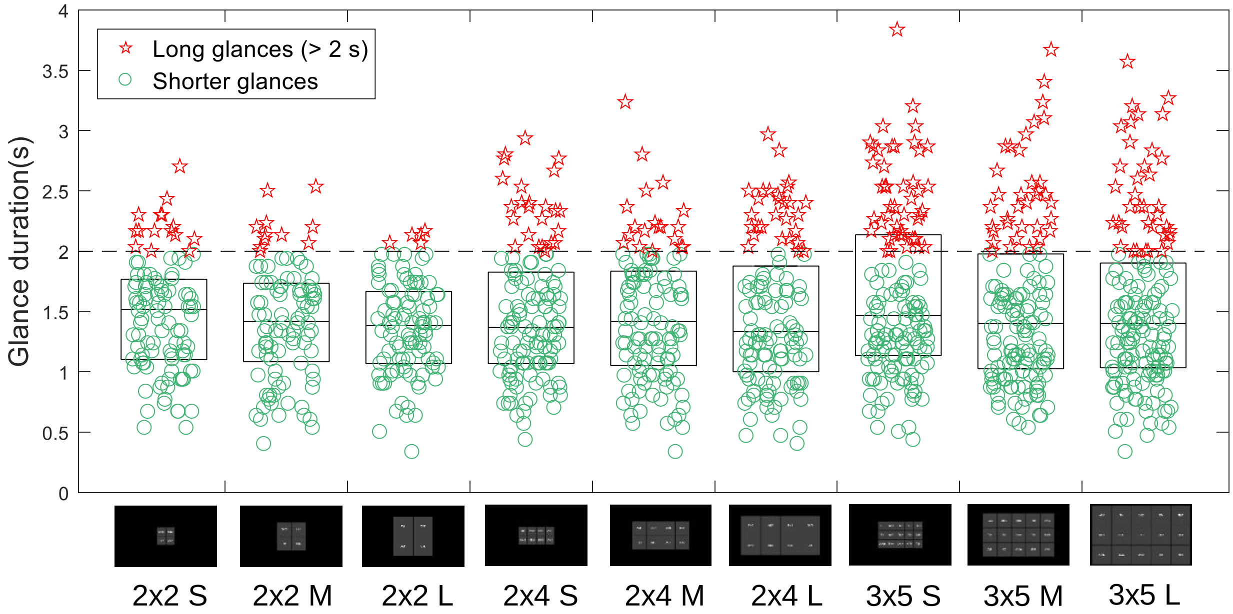

Figure 7. Glance duration for the nine designs. Note in this figure each data point represents one glance. Since the drivers may take more than one glance in a trial, this figure is essentially different from Figure 5, in which each data point represents the total eyes-off-road time of one trial.

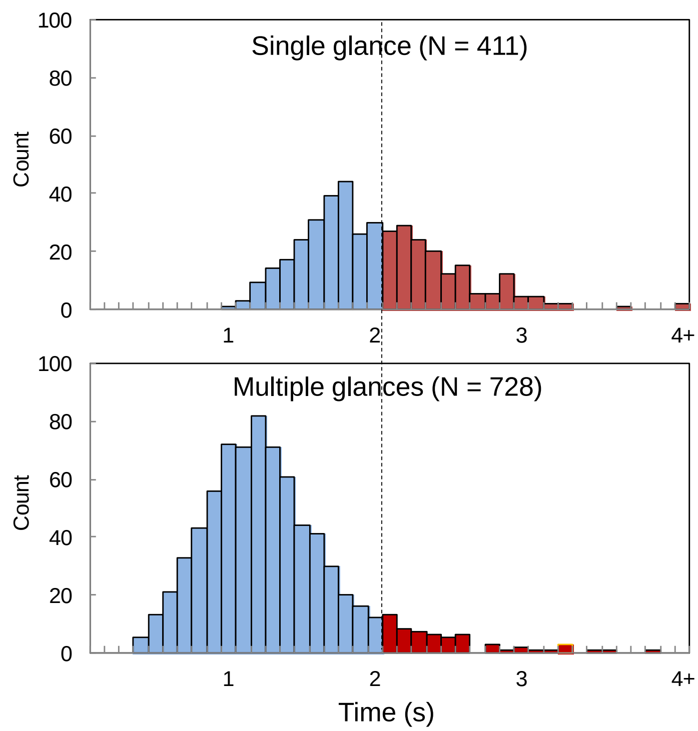

Figure 8. Distributions of glance durations of single-glance trials (top figure) and multiple-glance trials (bottom figure). Histogram bin width: 0.1 s. The vertical line is the 2-second threshold that separates the long glances.

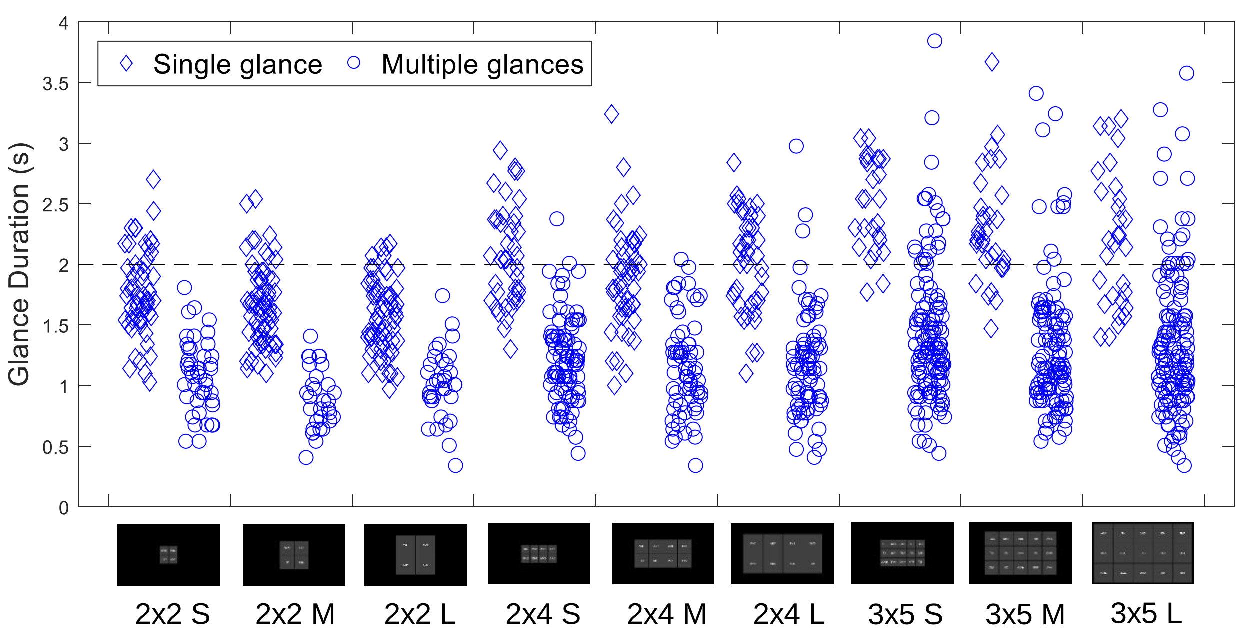

Figure 9. Glance duration separated by tasks completed in single or multiple glances.

Table 1. Summary statistics of task completion time and eye glance metrics

		2x2 layout			2x4 layout			3x5 layout		
		S	M	L	S	M	L	S	M	L
Task completion time - parked (in s)	GM*	1.39	1.32	1.25	1.69	1.66	1.60	2.33	2.12	2.29
	50th	1.40	1.37	1.24	1.72	1.67	1.62	2.24	2.14	2.17
	85th	1.73	1.66	1.52	2.13	2.09	2.12	3.57	3.07	3.71
	95th	1.95	1.82	1.77	2.60	2.57	2.32	4.31	3.51	4.99
Task completion time - driving (in s)	GM	1.93	1.76	1.72	2.62	2.25	2.40	3.44	3.09	3.18
	50th	1.90	1.74	1.72	2.52	2.16	2.36	3.10	2.92	3.02
	85th	2.69	2.26	2.21	3.87	3.17	3.16	5.18	4.52	5.23
	95th	2.97	2.66	2.85	4.76	3.89	4.77	6.91	6.03	7.69
TEORT (in s)	GM	1.83	1.66	1.63	2.34	2.05	2.18	2.99	2.77	2.79
	50th	1.84	1.67	1.64	2.36	2.04	2.20	2.87	2.72	2.71
	85th	2.30	2.09	2.02	3.00	2.69	2.77	4.52	3.89	4.41
	95th	2.59	2.24	2.52	3.91	3.26	3.49	5.37	4.69	6.24
Glance duration (in s)	GM	1.39	1.32	1.30	1.38	1.36	1.35	1.56	1.41	1.37
	50th	1.52	1.42	1.38	1.37	1.42	1.33	1.47	1.40	1.40
	85th	1.98	1.91	1.80	2.06	2.04	2.20	2.51	2.28	2.22
	95th	2.24	2.18	2.03	2.53	2.31	2.50	2.90	2.87	2.87
Trials w/ long glance		19%	14%	6%	30%	24%	35%	54%	45%	40%

*GM: geometric mean

Table 2. Summary of findings of task completion time and eye glance behavior

	Quantity of buttons			Button size		
	2x2 → 2x4	2x4 → 3x5	2x2 → 3x5	S → M	M → L	S → L
Task completion time (parked)	↑ (***, <u>*0.19*</u>)	↑ (***, <u>*0.15*</u>)	↑ (***, <u>*0.46*</u>)	n.s.	n.s.	n.s.
Task completion time (driving)	↑ (***, <u>*0.20*</u>)	↑ (***, <u>*0.12*</u>)	↑ (***, <u>*0.46*</u>)	↓ (***, <u>*0.03*</u>)	n.s.	↓ (*, <u>*0.01*</u>)
Total eyes-off-road time	↑ (***, <u>*0.22*</u>)	↑ (***, <u>*0.13*</u>)	↑ (***, <u>*0.48*</u>)	↓ (***, <u>*0.03*</u>)	n.s.	↓ (*, <u>*0.01*</u>)
Glance duration	n.s.	n.s.	↑ (*, <u>*0.01*</u>)	n.s.	n.s.	↓ (*, <u>*0.01*</u>)
Tasks w/ multiple glances	↑ (**)	↑ (**)	↑ (**)	n.s.	n.s.	n.s.
Tasks w/ long glances	↑ (***)	↑ (***)	↑ (***)	n.s.	n.s.	n.s.

*: $p < .05$, **: $p < .01$, ***: $p < .001$, n.s.: not significant, <u>*0.19*</u>: partial eta squared (η^2).

Biographies

Fred Feng, Ph.D., is a research fellow at the University of Michigan Transportation Research Institute. His research interests include human-machine interactions, user interface designs, and human performance modeling with applications of road traffic safety, driver behaviors, and the safety of vulnerable road users.

Yili Liu, Ph.D., is Arthur F. Thurnau Professor of Industrial and Operations Engineering at the University of Michigan at Ann Arbor. His areas of research and teaching include computational cognitive modeling, cognitive engineering and human factors, and aesthetic and cultural ergonomics.

Yifan Chen, Ph.D., is the manager of Mobility Research at Ford Motor Company. He leads a team to develop innovative technologies for further mobility needs. Previously, he had worked in mobile computing, HMI simulation, and CAD/CAE. Dr. Chen received his Ph.D. in Naval Architecture and Marine Engineering from the University of Michigan, Ann Arbor in 1993.