Author(s): Kujala, Tuomo; Salvucci, Dario D.

Title: Modeling visual sampling on in-car displays: The challenge of predicting safety-critical lapses of control

Year: 2015

Version:

Please cite the original version:

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. https://doi.org/10.1016/j.ijhcs.2015.02.009

MODELING VISUAL SAMPLING ON IN-CAR DISPLAYS: THE CHALLENGE OF PREDICTING SAFETY-CRITICAL LAPSES OF CONTROL

Tuomo Kujala

Department of Computer Science and Information Systems,

University of Jyväskylä,

P.O. Box 35, FI-40014 Jyväskylä, Finland; e-mail: tuomo.kujala@jyu.fi

tel. +358 40 024 7392, fax +358 14 260 4400

Dario D. Salvucci

College of Computing and Informatics, Drexel University

3141 Chestnut St., Philadelphia PA, 19104; email: salvucci@drexel.edu

tel. +1 215 895 2674, fax +1 215 895 0545

Keywords: driving, distraction, in-car displays, interleaving strategy, visual search, cognitive modeling

Abstract

In this article, we study how drivers interact with in-car interfaces, particularly by focusing on understanding driver in-car glance behavior when multitasking while driving. The work focuses on using an in-car touch screen to find a target item from a large number of unordered visual items spread across multiple screens. We first describe a cognitive model that aims to represent a driver's visual sampling strategy when interacting with an in-car display. The proposed strategy assumes that drivers are aware of the passage of time during the search task; they try to adjust their glances at the display to a time limit, after which they switch back to the driving task; and they adjust their time limits based on their performance in the current driving environment. For visual search, the model assumes a random starting point, inhibition of return, and a search strategy that always seeks the nearest uninspected item. We validate the model's predictions with empirical data collected in two driving simulator studies with eye tracking. The results of the empirical study suggest that the visual design of in-car displays can have a significant impact on the probability of distraction. In particular, the results suggest that designers should try to minimize total task durations and the durations of all visual encoding steps required for an in-car task, as well as minimize the distance between visual display elements that are encoded one after the other. The cognitive model helps to explain gaze allocation strategies for performing in-car tasks while driving, and thus helps to quantify the effects of task duration and visual item spacing on safety-critical in-car glance durations.

1 INTRODUCTION

Ubiquitous computing has brought a wealth of information and entertainment to the fingertips of drivers. Although there are clear benefits to the increased availability of services and infotainment on the road, there may be serious drawbacks: in-car visual tasks increase the probability that driver's eyes wander from the road, potentially leading to unsafe situations for the driver and others. Extensive field studies have noted the statistical relationship between in-car glance durations and the probability of safety-critical incidents (see Liang et al., 2012). While the responsibility of safe driving belongs primarily with the driver, those who design and build in-car user interfaces also strive to minimize the potential of visual distraction of these interfaces.

The U.S. National Highway Traffic Safety Administration (2013) recently released testing and verification guidelines for in-vehicle electronic devices. These guidelines propose three criteria for newly developed in-car systems:

1. Individual glance durations: "For at least 21 of the 24 test participants, no more than 15 percent (rounded up) of the total number of eye glances away from the forward road scene have duration of greater than 2.0 seconds while performing the testable task one time."
2. Mean glance duration: "For at least 21 of the 24 test participants, the mean duration of all eye glances away from the forward road scene is less than or equal to 2.0 seconds while performing the testable task one time."
3. Total glance time: "For at least 21 of the 24 test participants, the sum of the durations of each individual participant's eye glances away from the forward road scene is less or equal to 12.0 seconds while performing the testable task one time."

As a complement to such guidelines, there are various helpful procedures (e.g., SAE-J2365, 2002) and prototyping tools (e.g., Distract-R: Salvucci, 2009) available for designers for analyzing relevant measures of driver distraction and performance, such as in-car task completion times and effects on lateral vehicle control. However, these methods are currently unable to predict arguably the most safety-relevant aspect of multitasking while driving, namely in-car glance behavior (NHTSA, 2013; Liang et al., 2012) and to provide guidance in design to create in-car user interfaces that would pass the NHTSA criteria. At least for now, designers and manufacturers must still rely on expensive and time-consuming testing with human drivers on novel in-car user interfaces. A deeper understanding of drivers' visual sampling strategies would go a long way toward more rigorous testing procedures, empirical and otherwise, to better predict and alleviate driver distraction.

In this paper, we study how drivers perform visual sampling on an in-car device interface, specifically when searching through a large number of unordered visual items (e.g., radio stations, music albums and songs, navigational points of interest) spread across multiple screens. Specifically, we study the effects of two possible layouts for its visual items: a *grid* layout with a constant number of columns and varying number of rows, and a *list* layout with a vertical list of all items. Kujala and Saariluoma (2011) found higher individual in-car glance durations by increasing the number of items per screen as well as increased glance durations for a grid-style menu layout compared to a list layout.

The results of the current work help to better understand the effects of unordered menu layout on driver glance behavior, and more generally, to elucidate possible gaze allocation strategies used by drivers when interacting with in-car displays. As such, we hope to better understand drivers' visual sampling in general and in the context of recent guidelines and tools like those mentioned above.

We begin by specifying a proposed strategy for visual sampling while driving, along with an instantiation of this strategy as a computational cognitive model developed in the ACT-R cognitive architecture (Anderson, 2007). The proposed strategy is based on several key assumptions: (1) each in-car glance begins with the driver fixating a random item on the display; (2) after encoding the current item, the driver transitions to the nearest yet-unattended item (in unordered menus), thus inhibiting return to attended items; (3) drivers monitor the passage of time during performance of the search task; (4) to the best of their ability, drivers try to limit their glances at the display to a reasonable amount of time, after which they switch back to the driving task; (5) drivers adjust their time limits for search based on their performance in the current driving environment. Given this sampling strategy and model, we describe two experiments that provide human data to elucidate these issues and to test the validity of the claims as embodied by the cognitive model.

2 VISUAL SAMPLING WHILE DRIVING: STRATEGY AND MODEL

Visual search construed most broadly is an extremely interesting and challenging problem with many aspects (Wolfe, 2007). In the context of in-vehicle interfaces, visual search can take on a more specific form in three ways. First, visual search is often constrained to a set of similarly sized items with text labels and/or icons; certainly this is not always the case (e.g., search in a navigational map), but is one common case for in-vehicle interfaces. Second, visual search often occurs across multiple screens of items: because an in-vehicle display can typically hold a very limited set of items, scrolling across screens is likely in many search scenarios. Third, the visual search is not continuous, but instead done by brief in-vehicle glances returning vision back to the road in between the glances (i.e., visual sampling). Thus, as mentioned, we focus our efforts on visual sampling in the context of a grid or list of varying number of items spread across multiple screens.

2.1 Visual-Search Strategy

We begin by proposing a core strategy for visual search while driving, borrowing a number of ideas from previous models of visual search in non-multitasking contexts. First, we assume that the visual-search task is interleaved with driving in a series of glances to the display (for search) that are interleaved with glances to the roadway (for driving). An *in-car glance* is defined here (following SAE-J2396: SAE, 2000) to begin once the gaze starts to move towards the in-car display, and to end once the gaze has returned to the road scene. Thus, an in-car glance can comprise of several fixations on the in-car display.

Each in-car glance begins with the driver fixating a random item on the display. When the driver finishes encoding the current item, we assume, following the model of Halverson and Hornof (2007), that the driver transitions to the nearest yet-unattended item; if there are multiple nearest unattended items, the driver chooses one at random. The limitation of this kind of search model is that it does not probably apply to semantic and alphabetic organizations of items (Bailly et al., 2014). Thus, here we are modeling search behaviors in unordered menus. It also assumes inhibition of return to attended items, which has been found in standard visual-search paradigms (e.g., Klein, 2000; Posner and Cohen, 1984) but has not, to our knowledge, been explored in a similar multitasking context. The central issue here is whether "markers" of attended items (e.g., "finsts": Pylyshyn, 1989) persist across multiple glances to a display—or, put another way, whether the items marked as attended will remain marked after an interleaving glance to the roadway and the associated time needed to focus on the driving task.

2.2 Interleaving Strategy

The next challenge in our understanding of visual sampling while driving concerns the timing of interleaving between the search and driving tasks. Our understanding of this process

generally follows the guidelines of the theory of threaded cognition (Salvucci and Taatgen, 2008), which assumes that each task is associated with a distinct cognitive "thread" and that these threads share cognitive resources in a balanced manner. However, this theory does not dictate one important piece of driver behavior, namely how the driver shares visual resources between the two tasks. For this purpose, we make three important assumptions: (1) that drivers are aware of the passage of time (to the best of their ability) during performance of the search task; (2) that drivers try to limit their glances to a reasonable amount of time, after which they switch back to the primary driving task; and (3) drivers adjust their time limits for search based on aspects of, and their performance in, the current driving environment.

Related to the first two assumptions, Wierwille (1993) found that drivers try to limit in-car glances within the range of 500 to 1600 milliseconds in most real-world driving environments. Related to the third assumption, Wierwille (1993) also found that drivers adapt their in-car glance durations according to the driving task demands by shortening individual glance durations with increased driving demands. More specifically, our proposed strategy posits that drivers adapt their time limit for in-car glance based on the driving environment immediately upon returning to the driving task: if the vehicle is stable and "well-placed" in the lane, driver increase the limit, under the notion that perhaps they could have done more searching; if the vehicle is unstable and/or badly displaced from the lane center, drivers decrease the limit, under the notion that the current limit was too long and resulted in a less desirable situation. The details of this process are further quantified in the model below.

2.3 Cognitive Model

The above sections provide a description of the overall strategy for visual sampling while driving; however, we desire a more rigorous formulation to facilitate testing and direct comparison to empirical data. For this purpose, we developed a computational cognitive model of these strategies in the ACT-R architecture (Anderson, 2007) using the Java ACT-R

task environment (version 1.1). ACT-R has a long history of modeling both driving (Salvucci, 2006) and complex perceptual-motor tasks more generally (summarized in Anderson, 2007). Models developed in ACT-R are specified as condition-action *production rules* that embody particular procedural skills (e.g., the skills necessary for driving and search). For our model here, ACT-R offers a number of benefits, most notably the incorporation of rigorous theories of eye movements and temporal perception (described soon), as well as the ability to run computer simulations to gather testable predictions.

The model of visual search generally follows the strategy specified earlier, instantiated in the form of ACT-R production rules that follow this process: for each in-car glance, ACT-R's spotlight of visual attention starts at a random item, then proceeds to the nearest not-yet-attended item. From these unobservable shifts of visual attention, ACT-R predicts observable eye movements using a recent theory developed for reading and related tasks (Salvucci, 2001); this theory can predict, for example, skipped fixations on short high-frequency words and multiple fixations on long low-frequency words, and thus provides a realistic mapping from attention to eye movements. To check whether or not a particular item matches the target, the model checks the first word of the item's title, and in the case of a match, continues checking the rest of the title. If the item matches the target, the model presses the item to complete the trial; otherwise, the model continues to the nearest unattended item as described earlier. Eventually, if the model has attended all the items and still has not found the target, the model locates and presses the downward scroll arrow to view the subsequent screen. We assume, following Janssen et al. (2012), that the press of a scroll button would act as a motor cue and a natural breakpoint for a task switch, and thus, our model returns eyes back to the driving environment after each change of a screen.

The model of the secondary task was then integrated with an existing model of driver behavior (Salvucci, 2006). The interleaving of search and driving follows the strategy

mentioned earlier, and through its instantiation in ACT-R, benefits from ACT-R's embedded theory of time perception (Taatgen et al., 2007). The theory of time perception posits that internally, time perception acts like a metronome that ticks slower as time progresses, with additional noise drawn from a logistic distribution—the end result being predictions that match well to the abilities and limitations of real human time perception. The initial number of "ticks" used by the model for an in-car glance was set to start with a cautious strategy, near the lower limit of Wierwille's (1993) visual sampling model—17 ticks, which corresponds to roughly 500 milliseconds (although noise in the model may change this interval slightly). Whenever the model begins an in-car glance, it starts its mental timer, and continues to check whether the time has reached or surpassed the current limit; when it has, the model switches back to the primary driving task.

As for human drivers, the model driver can adapt its time threshold according to the demands of the driving environment. When the model returns to the driving task after a search glance, it estimates the stability of the driving in relation to speed and lane position. Vehicle stability is measured in the model as a function of the vehicle's lateral position in the lane and its lateral (side-to-side) velocity; as detailed in Salvucci (2006), there are two parameters that control the estimate of stability, namely thresholds for lateral position and lateral velocity respectively. If the driving is stable, the model increases the time limit by 1 tick; if not, the model resets the limit to its initial value, representing a decision to revert back to a safe interval (see Salvucci, Taatgen, & Kushleyeva, 2006, for a similar approach). The 1-tick increase here corresponds to the most cautious increase—initially small, but increasingly large because of increasing noise in estimates of longer time periods (Taatgen et al, 2007).

Figure 1 provides an overview of the model's general flow of processing including branches at decision points. The boxes roughly correspond to the core ACT-R production rules that control behavior, and the arrows represent typical control flow from one rule to the

next. It should be noted that under the integrated theory of time perception (Taatgen et al., 2007), the internal cognitive timer is updated subconsciously (i.e., is not actively performed or controlled by the driver directly). The noise in the timing process also occurs automatically and can result in slightly different perceptions of time for different trials. Nevertheless, the act of checking and acting upon the running timer is indeed under active driver control, and is noted in the figure as part of the checks for whether the time has reached the desired limit.

In running model simulations, the model driver was given the goal to drive at 80 km/h on the center lane of a three-lane road and to follow a simulated car that kept a constant speed of 80 km/h, following the NHTSA testing guidelines (2013). We estimated several parameters of the model to achieve the best overall fit to the empirical data of the experiments: two parameters that account for visual encoding times (*emma-enc-fac*=.009, default .006, *emma-enc-exp*=1.0, default .4), and the driver's stability factor (*stabilityFactor*=2.0). The stability factor affects the driver's threshold for vehicle instability, above which the driver avoids switching away from driving as described above (Salvucci, 2009). It was observed that drivers did tolerate a fair amount of deviation in vehicle's lateral position but that lane excursions were rare.

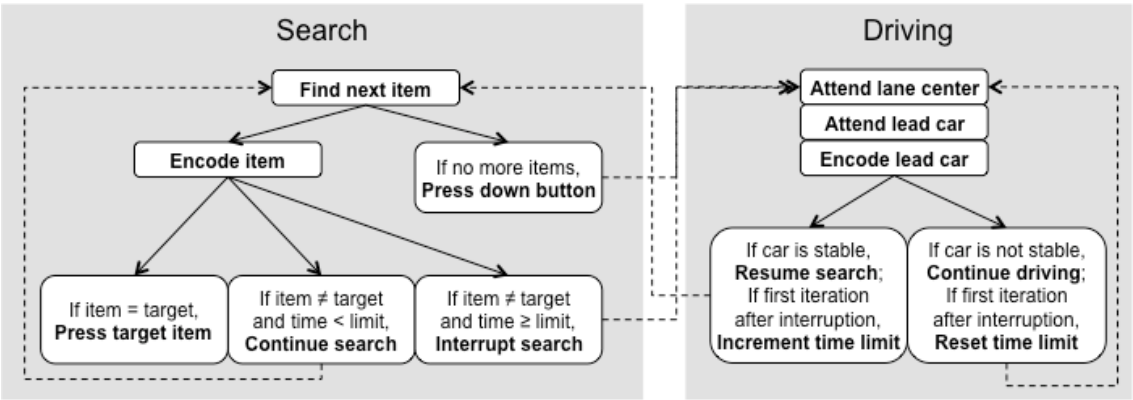

Figure 1. Schematic overview of the model's flow of processing.

The resulting model of search and driving behavior runs in simulation and generates predictions of task performance. Most critically for purposes here, the model generates predictions of in-car glance durations, as derived by the movement of visual attention dictated by the model's search and task interleaving mechanisms as well as by the predictive model of eye movements built into the ACT-R architecture, as mentioned. In the next two sections, we report the results of two empirical studies and, for each, compare these results with the predictions of the full model to better understand how our proposed strategies correspond to human behavior in the empirical tasks.

3 EXPERIMENT 1: MENU STRUCTURE AND ITEMS PER SCREEN

The first experiment examined the effects of varying the menu layout and the number of the items in the in-car menu, using the *grid* and *list* layouts as well as the number of menu items (6, 9 and 12 items per screen, i.e., 60, 90 and 120 menu items per 10-screen menu) as the critical variants. This section describes the experiment and results, and also compares the empirical results to those of the ACT-R model to better understand how human behavior matches the visual-search and interleaving strategies described in the previous section.

3.1 Research Method

The experiment followed a within-subject 2x3 design. There were two different menu structures in the in-car search tasks, Grid and List, as well as three different sizes of item sets per screen (6, 9 and 12 items), corresponding to item sets of 60, 90, and 120 items in total per 10 screen menus.

3.1.1 Participants

A total of 12 volunteers were recruited via student mailing lists of University of Jyväskylä. The sample included 6 women and 6 men between the ages of 22 and 34. They all had a valid driving license and either 20,000 km or 2 years of driving experience; these criteria served to

mitigate the effect of low driving experience on visual sampling efficiency (Wikman et al., 1998). All the participants had normal or corrected-to-normal vision. The experiments were conducted in Finnish with fluent Finnish speakers. Participants received a movie ticket as compensation for participation in the study.

3.1.2 Environment and Tools

The medium-fidelity fixed-base driving simulator used in the study is located at the driving simulator laboratory of the Department of Computer Science and Information Systems in the University of Jyväskylä (see Figure 2). The virtual driving scene was projected on three screens with a resolution of 1280 x 1024 pixels each. The front screen, positioned on a distance of roughly 135 centimeters from the participants' eyes, measured 170 x 64 cm and the two side screens measured 110 x 64 cm. The left screen was roughly 130 centimeters whereas the right side screen was roughly 150 centimeters from the participant's eyes. The corresponding visual angles subtended by the front driving scene were: horizontal 72.2 degrees, vertical 27.2 degrees. The visual angles subtended by the left screen were: 48.5 degrees horizontal, 28.2 degrees vertical and by the right screen: 42.0 degrees horizontal, 24.4 degrees vertical. The visual angle for the participant between the driving screens and the 22" interactive display was roughly 37 degrees. The size of the interactive screen on the upper part of the display was 640 by 380 pixels (18.0 x 10.7 cm, 8.2"), the distance form the participant's eyes was about 75 centimeters and the corresponding visual angles subtended by the screen were: horizontal 13.7 degrees, vertical 8.2 degrees. The simulator was equipped with a Logitech G25 force-feedback steering wheel, accelerator, and brake. The distance from participant to the screens was fixed but the positions of the pedals and the steering wheel were adjustable.

Figure 2. The medium-fidelity driving simulator and the experimental setup from the participant's point of view.

The driving simulation software was provided by Eepsoft (http://www.eepsoft.fi). In the experiment, drivers navigated a virtual environment with a three-lane empty straight highway road and an instructed speed limit of 80 km/h. The virtual environment included a heads-up display (HUD) speedometer, RPM gauge, and rear- and side-view mirrors. The virtual car's transmission was set to automatic. Driving log data were logged and saved at 10 Hz. The research equipment included a head-mounted Dikablis eye-tracking system with a 50 Hz sampling rate and a laptop for controlling the secondary-task display, namely a 3M M2256PW (22") capacitive multi-touch display. In this first experiment, an additional eye tracking system—a SMI RED remote eye-tracking system with 500 Hz sample rate—was attached to the top of the touch screen to record more detailed data on participants' eye movements on the display.

The search tasks were performed on the screen with six different layouts, illustrated in Figure 3. The font size was identical in all text labels in all conditions. The search tasks simulated a situation in which the participant searches for a certain song in an in-car music player. The distraction effects of searching music tracks while driving are some of the most

studied topics in distraction research (e.g., Jeon et al, 2015; Lee et al., 2012; Chisholm et al., 2007; Salvucci et al., 2007). The song titles were artificial and were generated with an online song name generator (http://www.songname.net/). The song lists were unordered and the positions of the song titles on each screen were varied. Although an alphabetical ordering might be used in some cases on real systems, there are many real-world situations in which alternative orderings would be used (e.g., sorting points of interest by proximity to the driver); we focus here on the more general case in which there is no predetermined ordering for the items, thus minimizing potential effects of previous knowledge and practice.

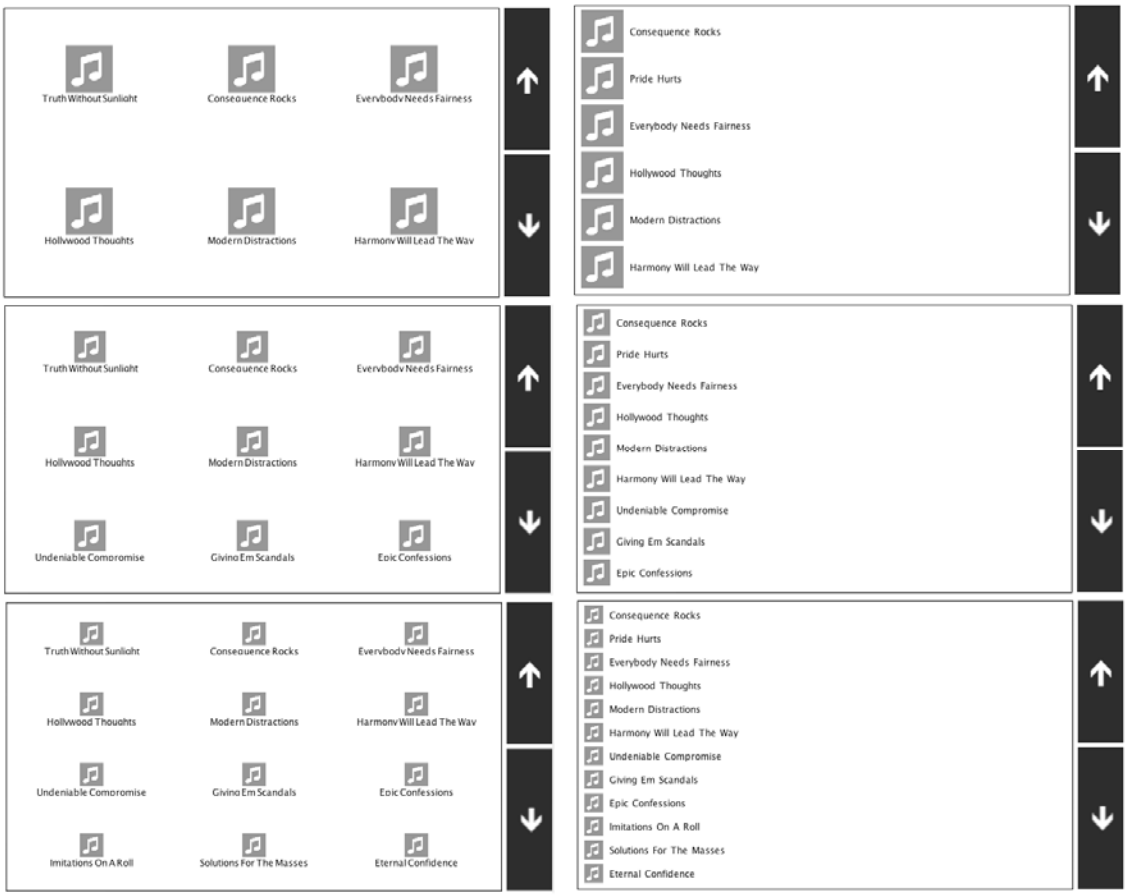

Figure 3. The in-car search displays with the Grid and List designs, with 6, 9 or 12 items per screen and the scroll buttons on the right (Experiment 2: scroll buttons left).

3.1.3 Procedure

After collection of demographic data, all participants completed a practice driving session that used the same driving environment as the experiment and lasted until the participant felt comfortable with the task (on average about 5 minutes per practice). Participants also did a practice session on multitasking (i.e., performing the search task while driving). In the experiment there were 6 blocks, one for each condition. Blocks were clustered by menu layout (grid or list), with half the participants starting with the list layout, and half with the grid layout in order to control unwanted learning effects (see Appendix 1 for an example). For the first three blocks, a random order of number of items per screen (6, 9, 12) was chosen for each participant. This order of blocks was repeated for the second set of blocks. Within each block, there were three trials. The first trial was considered a practice trial and only the other two were analyzed in order to further mitigate unwanted learning effects. The target items were always located at the same serial position from the beginning of the menu between the trials, which meant that the target item was located on different screens in the menus but the participant had to inspect the same number of items in the menu (from List-6 to List-9 and so on) in order to reach the target (see Appendix 1).

Participants were instructed to follow a speed limit of 80 km/h, following the NHTSA (2013) driving scenario and the simulation model. To encourage the participants prioritize the driving task, they were informed that the 6 most accurate participants in the driving task would be rewarded with an additional movie ticket. Driving task accuracy was instructed as a function of the total duration of lane excursions (where more time outside the lane being equivalent to lower accuracy). The accuracy of lane keeping was assessed by how many times the HUD speedometer (see Figure 2) crossed a white lane marking. Participants were also instructed to beware of unexpected events and react as they would in real situations. There were no unexpected events in the experiment, because the NHTSA (2013) driving

scenario does not include these. The participants may have ignored the instruction early after a few trials without unexpected events but with this instruction we wanted to encourage the participants to observe the road ahead in a more natural manner than merely observing the speedometer and the lane position. There were no time limits given for the completion of the trials. Participants were told not to hurry completing the search tasks and to take their time and prioritize the driving task. The driving task proved to be relatively easy and lane excursions were found to be rare. This also suggests that the participants were successful in prioritizing the driving task regarding the lane keeping as instructed and that they selected a cautious strategy for multitasking as assumed in our model.

3.1.4 Data Analysis

The in-car glance durations were scored manually frame-by-frame from the overlaid gaze and eye videos following the SAE-J2396 (SAE, 2000) definition. For the model predictions, the glance durations were calculated in a similar manner based on the predicted eye movements in the simulated task environment. Only the first five screens of the last two trials were scored for in-car glance durations, in order to have a data set of absent target search with four button presses. The first trial was intended for practice; to reduce unwanted learning effects in the data. The sixth screen included the search target for the second 12-item trials, and this is why we selected only the five first screens in trials 2 and 3 per condition (5+5 screens) under analyses. In order to compare the fit between the model's predictions and the data, correspondingly 12 simulated absent target searches on 10 screens each were ran for calculating the predicted values. In summary, for both Grid and List menus, 10 screens with 6, 9, and 12 items per screen, totaling in 60, 90, and 120 item search tasks were analyzed and compared between the empirical data and the model predictions with N=12.

The relevant measures related to the number and duration of the in-car glances (i.e., glances to the touch display): the total number of glances, total glance duration, mean and

maximum glance duration, and number and percentage of glances over 2.0 seconds. The first three measures are related to the visual demand of in-car tasks, whereas the latter three are intended to measure safety-critical lapses of control. Glance durations over 2 seconds have been associated with increased risk of safety-critical events in real traffic (Liang et al., 2012). We wanted to also see if our model can predict the values of our in-car tasks for the NHTSA (2013) criteria on the total glance times (max 12.0 s for the 85th percentile), mean glance duration (max 2.0 s for the 85th percentile), and the percentage of over-2.0-second glances (max 15% for the 85th percentile).

Even if not in the focus of our current modeling approach, we also evaluated the relationship of the predicted lane deviation to the predicted continuous glance metrics to validate the relationship between the long glance durations and deviation in vehicle's lane position. Our proposed strategy for visual search posited that drivers could mark attended items to inhibit return to these items, and that these markers remain even after looking away briefly to the driving task. This assumption makes the implicit prediction that items would not be revisited during visual search. In order to test this assumption, the participants' eye movements on the display were recorded with a sampling rate of 500 Hz. For each condition and for each participant, a screen with interrupted search was randomly selected for closer calculation on revisits per item (12 x 6 = 72 screens in total).

For statistical analyses, repeated-measures ANOVAs with an alpha level of .05 were used. For pairwise comparisons, a Bonferroni correction with SPSS-adjusted significance level of .05 was applied. In this paper, SPSS's Bonferroni adjusted p-values are quoted, and for each ANOVA, assumptions of sphericity were confirmed. If the assumption of sphericity was violated, degrees of freedom were adjusted with the Greenhouse-Geisser correction. Partial eta-square and mean differences were calculated as measures of effect size. In order to evaluate the relationship of the predicted lane deviation to the predicted glance metrics, we

analyzed how total, mean and max glance duration per screen could predict average lane deviation per screen of the model simulations with linear regression models (12 x 6 x 10 screens, N=720).

For testing the goodness-of-fit between model and data, two measures, RMSSD (Root Mean Squared Scaled Deviation) and r^2, are used. RMSSD was used to evaluate deviation from exact location whereas r^2 is a measure of fit to relative trend. High values for r^2 are important for reliably pointing out the better user interface alternatives, whereas RMSSD should be small if predictions of passing acceptance criteria are to be made.

3.2 Results and Goodness-of-Fit

As shown in Figure 4 for a total of 10 screens, study participants showed a significant increasing effect of the number of menu items in the number of in-car glances, $F(2,22) = 7.880$, $p=.003$, partial $\eta^2 = .417$. The mean difference between 6 item and 12 item tasks was 10.25 glances, $p=.033$, and between 9 and 12 item tasks 6.88 glances, $p=.012$. As expected, the model predicted similar trends. As the number of items per screen increased from 6 to 9 to 12, so did the total number of items to be inspected increase from 60 to 90 to 120 for the total of 10 screens. Both the human and model results reflect the fact that more items simply take more time to encode and process. There was also a significant interaction between the menu type and the number of menu items, $F(2,22) = 3.910$, $p=.035$, partial $\eta^2 = .262$. For the 6-item tasks, the number of glances was slightly lower for Grid than for List, whereas for the 9- and 12-item tasks, the number of glances was lower for List than for Grid. The model did not show this interaction effect but predicted an overall lower number of glances for List compared to Grid. No main effect of menu was found.

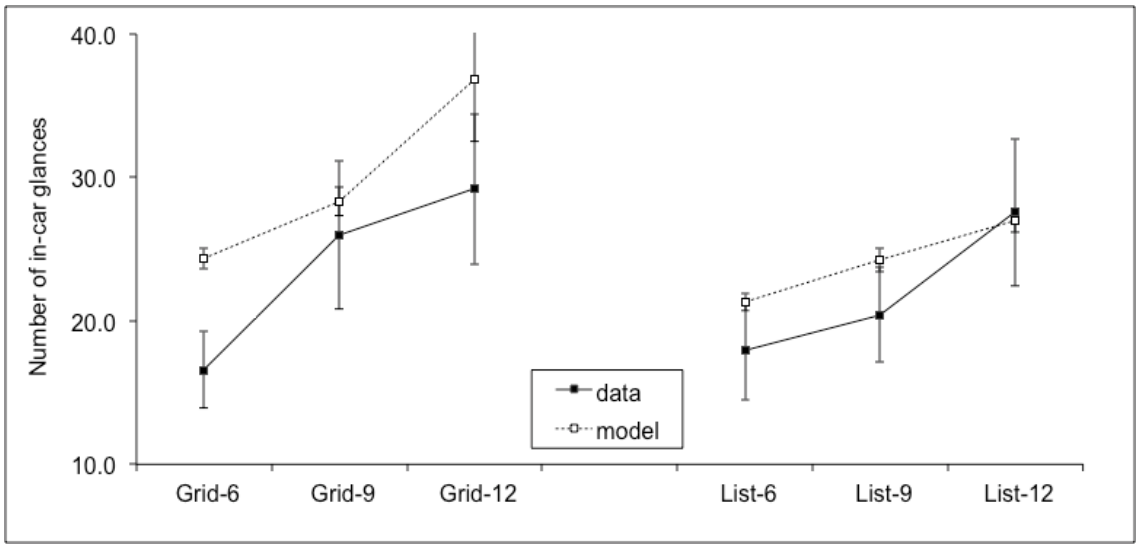

Figure 4. Number of in-car glances, scroll buttons right (10 screens, N=12), r^2=0.672, RMSSD=3.261. Bars represent 95% confidence intervals.

The total in-car glance durations in Figure 5 tell a similar story. There was a significant effect of the number of menu items on total glance duration, $F(2,22) = 12.765$, $p<.001$, partial $\eta^2 = .537$. The mean difference from 6 item tasks to 12 item tasks was 20.74 seconds, $p=.006$, and from 9 item to 12 item tasks it was 13.87 seconds, $p=.001$. The model predicted the relative trend and the deviation from the exact location of the data was small, although the model suggested somewhat lower total glance durations for the List menu structure than for the Grid. In the experiment, there were no significant differences between menus or significant interaction effects. Regarding the NHTSA (2013) criteria, none of the tasks would pass the criterion of total glance time being at most 12 seconds, as predicted.

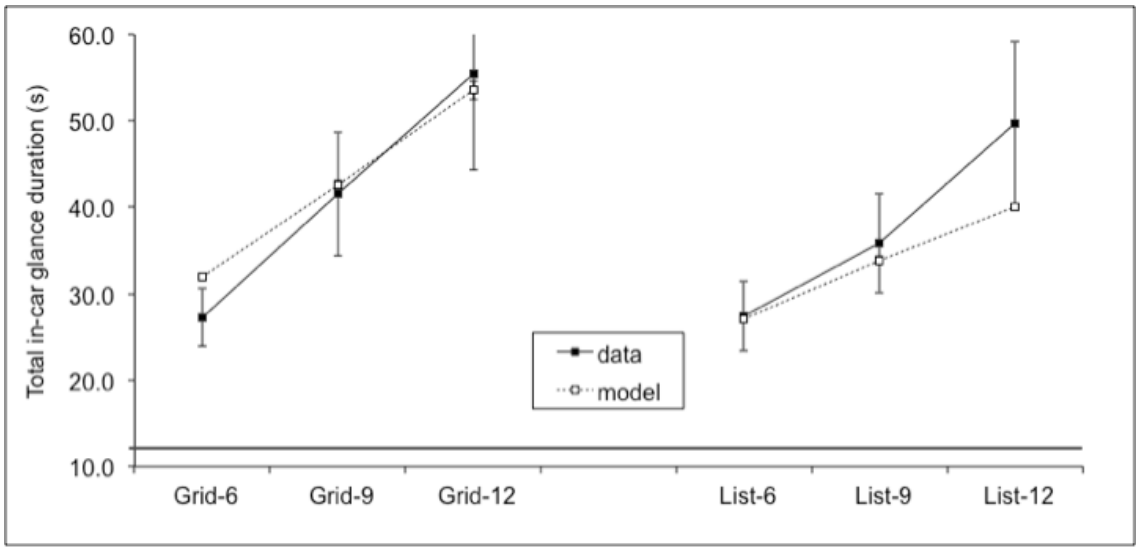

Figure 5. Total in-car glance duration (s), scroll buttons right (10 screens, N=12), r^2=0.843, RMSSD=1.615. Bars represent 95% confidence intervals. NHTSA (2013) verification threshold illustrated at 12.0 seconds.

The mean in-car glance durations (Figure 6) were close to the predictions, although the relative trends were somewhat off. The number of menu items had a significant effect on the mean glance durations, $F(2,22) = 7.108$, p=.004, partial η^2 = .393. The mean difference from 6 item to 12 item tasks was small (160 ms) but statistically significant, p=.005. No main effect of menu was found, but there was a significant interaction between menu and items, $F(2,22) = 5.187$, p=.014, partial η^2 = .320. For 6 item tasks, the mean glance durations were slightly shorter for List than for Grid, whereas for 9 and 12 item tasks, the mean durations were slightly longer for List. This could suggest the participants did take advantage of the closer distances between the titles in List-9 and List-12 and encoded more titles per glance than in Grid. In general, considering what the visual demands of the search task were compared to what the in-car tasks were like in the 1990s, the mean glance times were still near 1.6 seconds, in line with the upper limit of Wierwille's visual sampling model (1993). As predicted, all the tasks would pass the NHTSA (2013) criterion of mean in-car glance durations being at most 2.0 seconds for 85% of the participants.

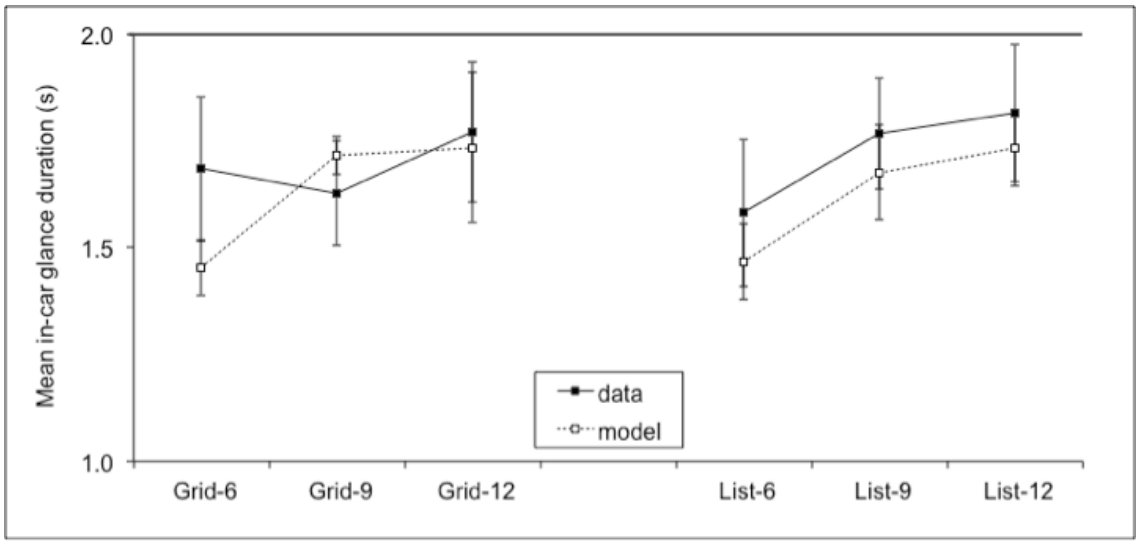

Figure 6. Mean in-car glance duration (s), scroll buttons right (10 screens, N=12), r^2=0.396, RMSSD=1.705. Bars represent 95% confidence intervals. NHTSA (2013) verification threshold illustrated at 2.0 seconds.

Whereas the measures above speak for the average visual demands of the in-car tasks, the maximum in-car glance durations and the number of very long in-car glances—representing lapses of control in visual sampling—provide safety-critical information and can also be more challenging to predict. Although the observed maximum in-car glance durations shown in Figure 7 were shorter than predicted, there was some similarity in the relative trend. As predicted, the number of menu items had a significant effect on maximum glance durations, $F(1.333,14.658) = 11.411$, p=.002, partial η^2 = .509. The mean difference between 6 item and 12 item tasks was .54 seconds, p<.001. Against predictions, no main effect of menu was found.

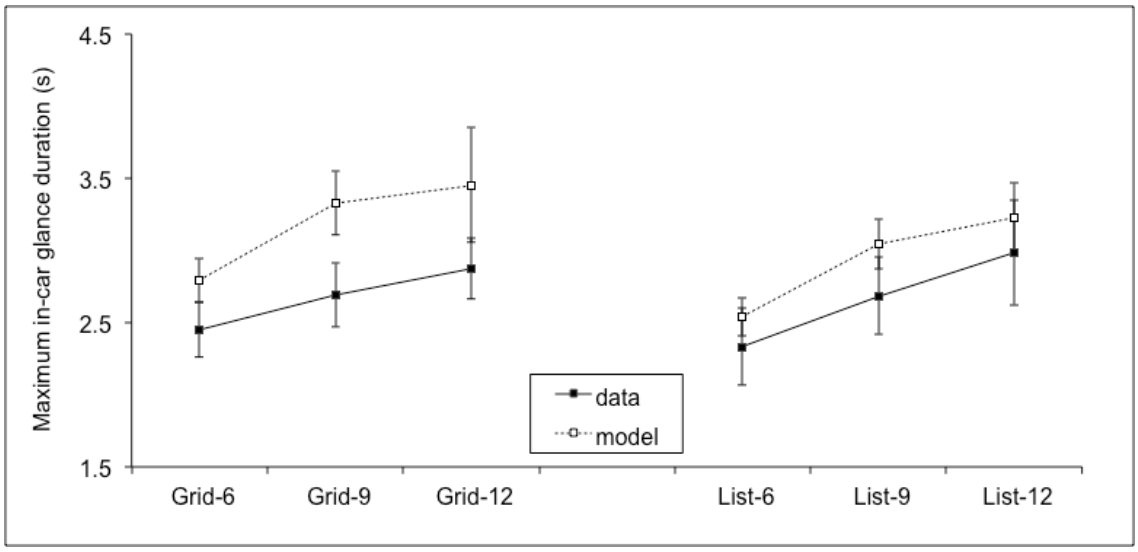

Figure 7. Maximum in-car glance duration (s), scroll buttons right (10 screens, N=12), r^2=0.766, RMSSD=4.180. Bars represent 95% confidence intervals.

For the safety-critical measure of number of glances over 2.0 seconds, the observed effect of the number of menu items was somewhat stronger than predicted for the List-conditions, in particular (r^2=0.725, RMSSD=1.425), $F(2,22) = 9.604$, p=.001, partial $\eta^2 = .466$. This effect was particularly strong for the List condition. The mean difference between 6 and 12 item tasks was 3.79 glances, p=.008, and between 9 and 12 item tasks 3.29 glances, p<.001. SEMs were large for this measure, which could explain partly the absence of the expected effect of menu structure. However, there was a significant interaction between menu and items on the percentage of over-2.0-second glances (Figure 8), $F(2,22) = 4.138$, p=.030, partial $\eta^2 = .273$. For List-6 the percentage was significantly lower than for Grid-6 whereas for the 9 and 12 item tasks the Grids had somewhat lower percentages. Regarding the NHTSA (2013) acceptance criterion, the model predicted pass for List-6 (percentage of over-2.0-second glances max 15% for 85% of participants) whereas the data suggests all tasks would fail.

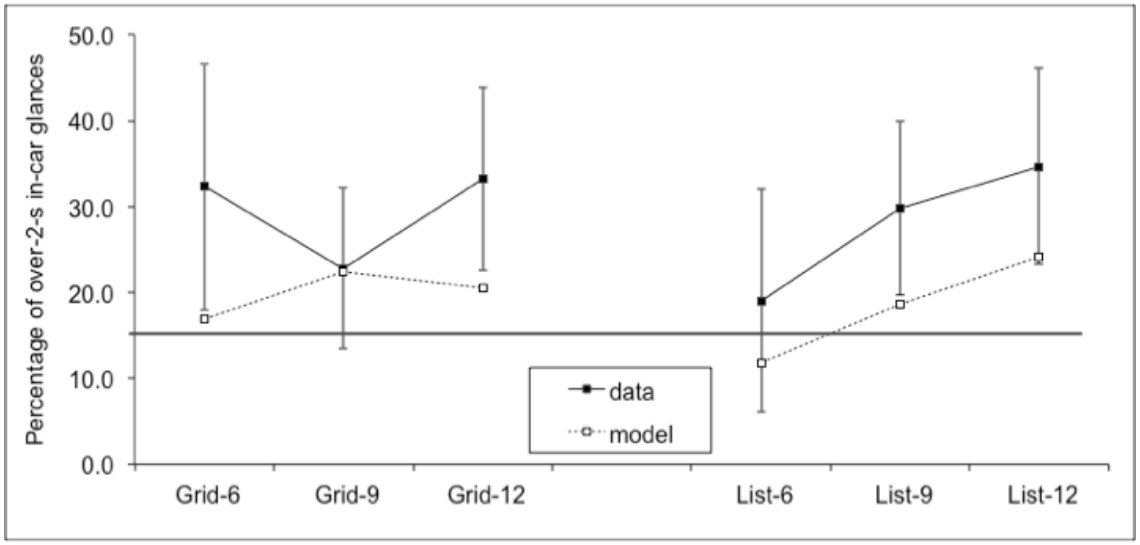

Figure 8. Percentage of over-2.0-second in-car glances, scroll buttons right (10 screens, N=12), r^2=0.316, RMSSD=2.005. Bars represent 95% confidence intervals. NHTSA (2013) verification threshold illustrated at 15.0 percent.

A closer analysis on gaze revisits per item (for 72 screens in total) revealed that for the List conditions, there were no revisits at all, and for the Grid conditions there were a few revisits for 4 participants on 5 screens (G6: 1, G9: 3, G12: 1). In total, only 7% of the searches included revisits, thus supporting (albeit not decisively) the assumptions underlying the model and the notion that, in general, a driver's visual search is efficient in avoiding revisits on items. In addition, the assumption that drivers begin search at any item and move their eyes to the nearest unattended item gained support by visually inspecting the gaze paths on these screens: systematic top-down or similar strategies seemed to be rare, and saccades tended to stay minimal in length and tended to avoid revisiting attended items.

The relationship of the predicted lane deviation of the vehicle to the predicted glance metrics was analyzed with linear regression models with a single screen as a sample (12 x 6 x 10 screens, N=720). Only the maximum in-car glance duration was able to predict average lane deviation (F(268.338, p<.001, β = .522, r^2 = .272, t(718) = 16.381, p<.001, 95% CIs

[.027, .035]). Total or mean glance duration did not correlate with lane deviation in a way that these could be used to predict lane deviation in the simulation. The finding gives support for the assumed relationship between individual glance durations and the driving stability in our model.

3.3 Discussion

In Experiment 1, the model was able to predict the observed increases in the number of in-car glances as well as in the total in-car glance durations, as the number of items to be inspected increased from 60 to 90 to 120 (with 6 to 9 to 12 items per screen across 10 screens). The model, as the embodiment of the search-while-driving strategy proposed earlier, did well at predicting the qualitative glance-behavior trends in the empirical data, although the quantitative predictions were sometimes less accurate. The model was able to predict the lower number of glances for the List-9 and List-12 tasks compared to corresponding Grid tasks. The model suggests the advantage is due to the shrinking distance between text labels in the List condition, as opposed to the distances in Grid condition, which remain farther apart.

In general, the model was able to predict that as the task durations increase (as seen in the total number and duration of in-car glances), the mean and maximum glance durations, as well as the percentage of over-2-second glances, tend to increase as well. The model was also able to predict correctly an advantage for the List structure over Grid on the individual glance lengths as the number of items is at 6. However, the model generally overpredicted the maximum in-car glance durations. The predicted relative trend for the increase of maximum glance durations with more items was generally present in the data, but the expected small difference between the menu structures was not observed. In addition, the observed increase of glances over 2.0 seconds, as the number of items increase, was much larger than predicted for Lists in particular.

The detailed gaze data revealing that only 7% of screens included a few revisits gave support for the assumption of stable markers of visually attended items. The analyzed data still does not reveal if the marker span is limited in duration—that is, if the survival of the markers is dependent on the duration of the interruption. However, the findings give support for the perseverance of the inhibition of return mechanism for facilitating visual search on an in-car display when interrupted by the visual demands of the driving task.

In the model simulations, only the maximum in-car glance duration predicted vehicle's lane deviation whereas total or mean glance duration did not. This finding suggests that in particular the long glances can lead to instability in driving and validates our model behavior in that also in our simulation a long off road glance affects the stability of the vehicle. However, the relationship in real traffic between lane deviation and crash risk is unknown, whereas there is a wealth of evidence of the more direct association between crash risk and long off road glances (Liang et al., 2012). It seems that in general, drivers do not try to optimize their lane position while multitasking and some lateral deviation is allowed even if the driving is prioritized (Janssen and Brumby, 2010). Keeping the own lane seems to be a sufficient and more critical goal than keeping an optimal lane position at all times, which is why we focus on glance metrics here.

4 EXPERIMENT 2: SCROLL BUTTON POSITION

When comparing the findings of Experiment 1 to those of Kujala and Saariluoma (2011), on the effects of menu structure on maximum in-car glance durations and overlong glances, the effects in the present study are much weaker. One possible explanation is that, in Experiment 1, the lapses of control that induce long glances seemed to arise from the driver's choice to complete a subtask before switching task (Bailey and Iqbal, 2008), where a subtask boundary would be reflected in the pressing of the down button to indicate completion of one screen of items. This mechanism is visible in the manually scored videos: the longest glances were

often terminated when the driver pressed the scroll button after finishing searching a screen. The scroll buttons were located in Experiment 1 on the right side of the display (Figure 3). This position could have undermined the advantage of List menu structure compared to Grid because of the large distance of the text labels in List to the touch screen button due to encoding demands no matter to what title the search ends at the screen. Besides increasing the number of individual long in-car glances, this scroll button position could have also decreased the total number of the glances if the participants decided to prolong a glance instead of investing a dedicated in-car glance for locating and pushing the button. It seems these types of events were more frequent in the empirical data than our model was able to predict.

To address these questions raised in the first experiment, we conducted a second experiment to elucidate whether the earlier observed difference between the menu structures (Kujala and Saariluoma, 2011) would be due to the distance of the interaction elements (text labels and the scroll buttons) on the display. After Kujala and Saariluoma (2011) an advantage for List over Grid was expected in the in-car glance durations when the scroll buttons are positioned closer to the items in the List. This modification in turn would help to further validate the model's account with respect to small changes in the secondary task.

4.1 Research Method

The Experiment 2 repeated the design of Experiment 1 but with one critical difference: the display's two scroll buttons (up and down) were positioned on the left side of the screen (see Figure 3). This subtle change has differing performance implications for the two conditions: for List, the decreased distance to these elements should decrease total encoding and task time, whereas for Grid, the position of the scroll buttons should not make a significant difference. With this change in the experimental design, we wanted to validate that the model would still work even if the elements of the particular in-car task were slightly modified, and

a relative advantage for List could be expected in the empirical data. The model with its parameters used for the predictions of Experiment 2 was exactly the same as for Experiment 1.

The driver sample was different than in Experiment 1 but represented the same demographics. In the statistical analyses of the empirical data, the sample size was 11 because of missing data on G12-tasks for one participant due to erroneous selections of items before the fifth screens in two tasks. Detailed eye movements were not recorded or analyzed in Experiment 2 and the analyses focused on the glance metrics.

4.2 Results and Goodness-of-Fit

For this experiment, there was a significant effect of menu structure on the observed number of in-car glances as expected (Figure 9), $F(1,10) = 6.657$, p=.027, partial η^2 = .400. The mean difference between Grid and List was 4.54 glances, p=.027. Again, as expected, the number of menu items had also a significant increasing effect on the number of glances, $F(2,20) = 32.835$, p< 001, partial η^2 = .767. The mean difference between 6 and 12 item tasks was 8.95, p<.001; between 9 and 12 item tasks, 4.00, p=.007; and between 6 and 9 item tasks, 4.95, p=.002. The fit on the relative trend is fair but the deviation from the exact location is still large.

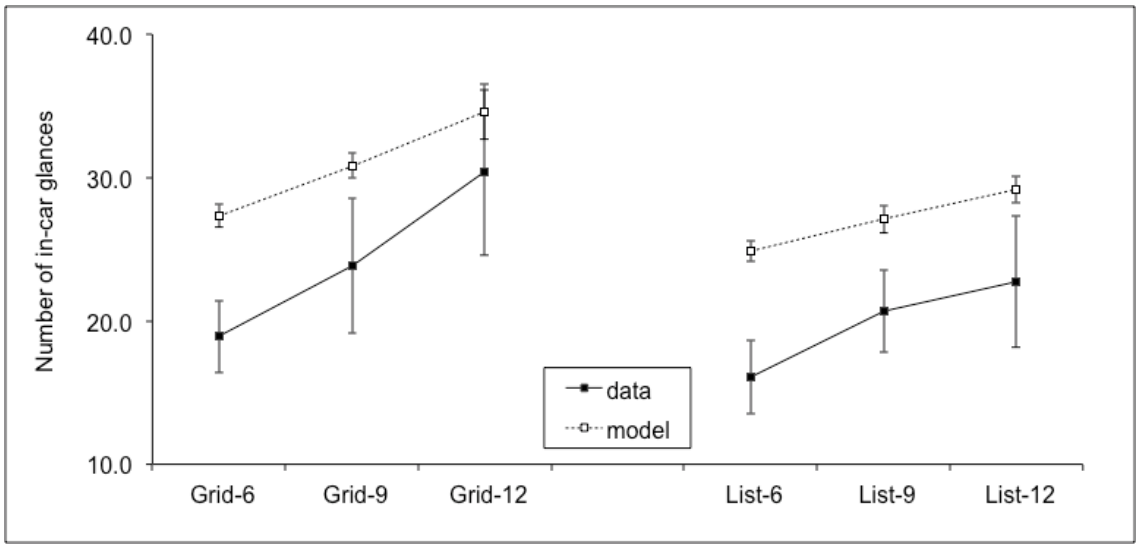

Figure 9. Number of in-car glances, scroll buttons left (10 screens, N=12), r^2=0.972, RMSSD=5.119. Bars represent 95% confidence intervals.

As expected, there was also a significant effect of menu structure on the total in-car glance durations (Figure 10), $F(1,10) = 6.961$, p=.025, partial $\eta^2 = .410$. The mean difference from Grid to List was 6.53 seconds, p=.025. For the effect of number of menu items on the total glance durations ($F(2,20) = 69.751$, p<.001, partial $\eta^2 = .875$), the effect size between 6 and 12 item tasks was 17.30 seconds, p<.001; between 9 and 12 item tasks, 8.42 seconds, p<.001; and between 6 and 9 item tasks, 8.88 seconds, p<.001. Again, none of the tasks would pass the NHTSA (2013) criterion on maximum total glance time of 12 seconds, as predicted.

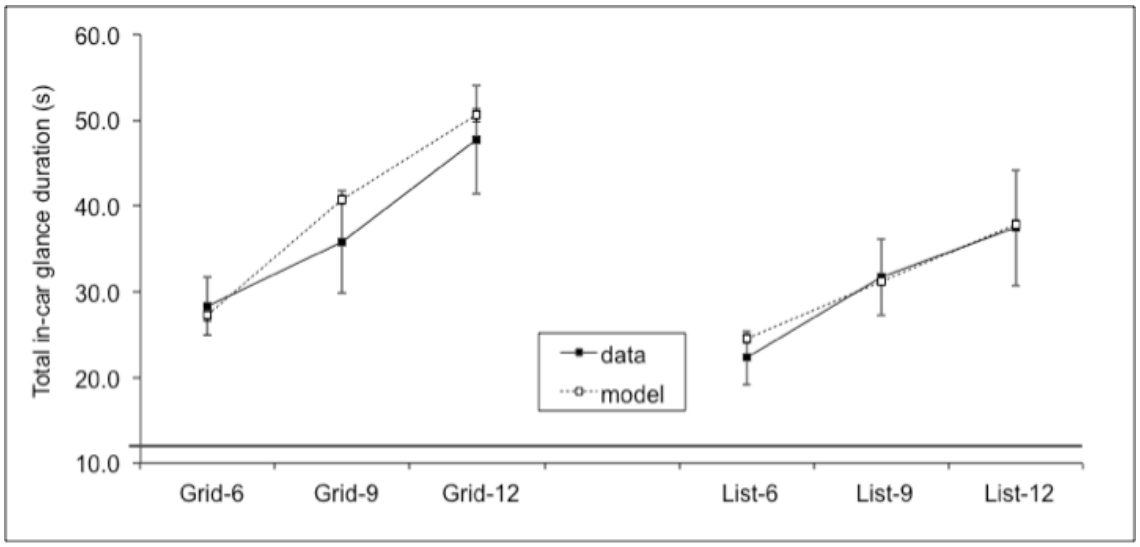

Figure 10. Total in-car glance duration (s), scroll buttons left (10 screens, N=12), r^2=0.950, RMSSD=1.108. Bars represent 95% confidence intervals. NHTSA (2013) verification threshold illustrated at 12.0 seconds.

Again, the mean in-car glance durations were close to the 1.6-second upper limit of Wierwille's (1993) model (Figure 11). The number of menu items had again a significant effect on mean glance durations, $F(2,20) = 10.317$, p=.001, partial $\eta^2 = .508$, with the small mean difference between 6 and 12 item tasks of .197 seconds, p=.003. The positioning of the scroll buttons on the left side of the menu seemed to bring down the mean glance durations as compared to Experiment 1 (see Figure 6). The model was able to predict the increase in mean glance duration by the number of items but the magnitudes were off, in particular for Grid. This was probably due to the overestimated number of glances. In addition, List did not have the predicted relative advantage over Grid in the data, there was no significant main effect of menu or a significant interaction effect. As predicted, all the tasks would pass easily the NHTSA (2013) limit of 2.0 seconds for mean glance duration.

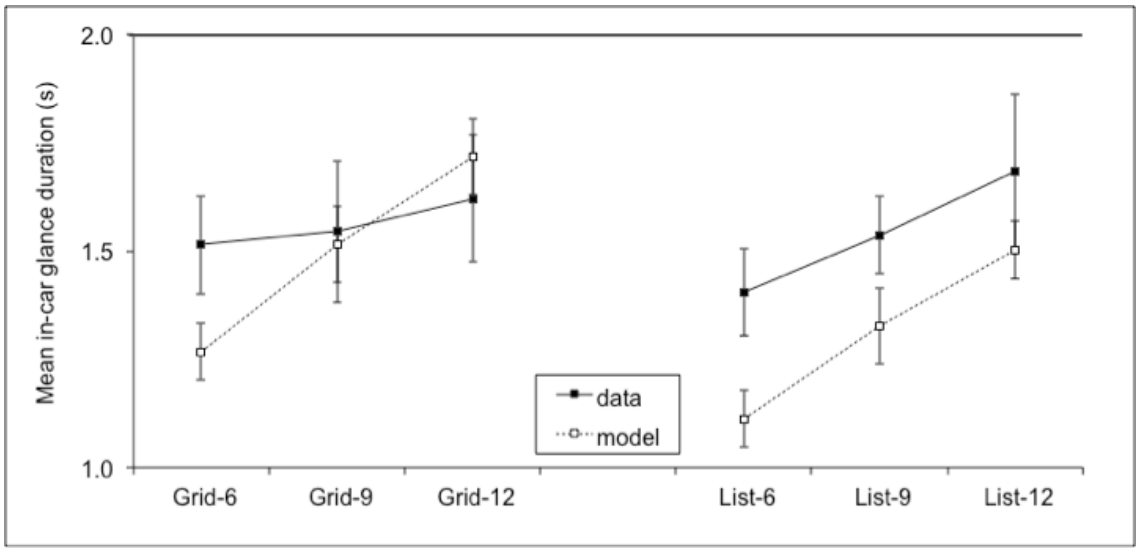

Figure 11. Mean in-car glance duration (s), scroll buttons left (10 screens, N=12), r^2=0.659, RMSSD=4.038. Bars represent 95% confidence intervals. NHTSA (2013) verification threshold illustrated at 2.0 seconds.

The left-side position of the scroll buttons seemed to slightly lower also the observed maximum in-car glance durations (compare Figures 7 and 12). Again, the predicted maximum durations were significantly higher and the predicted effect of menu structure did not become visible in the data. The effect of the number of menu items was this time smaller but still significant, $F(2,20) = 4.295$, p=.028, partial η^2 = .300. The significant mean difference between 6 and 12 item tasks was .32 seconds, p=.026.

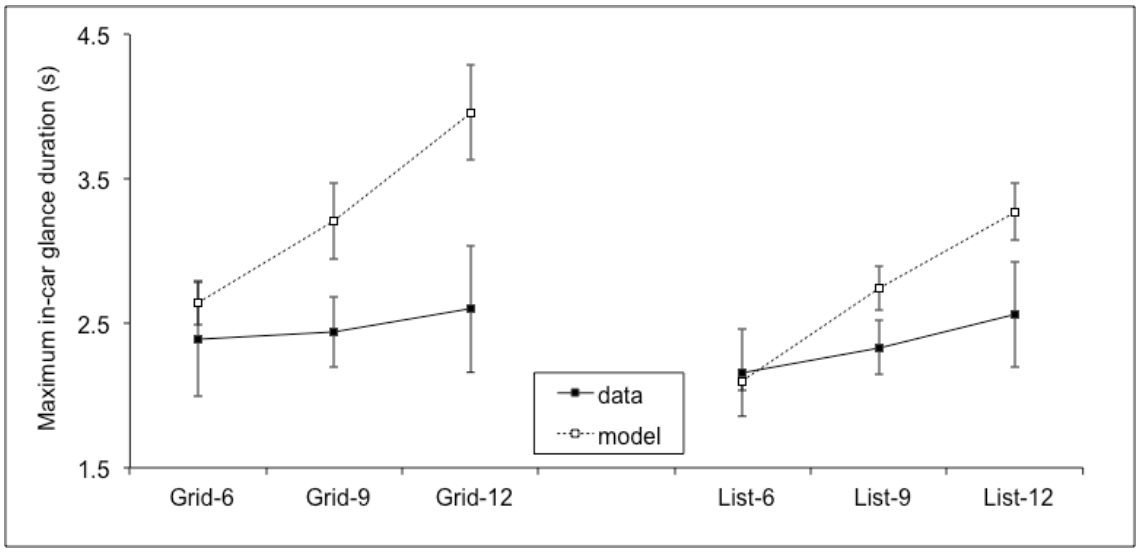

Figure 12. Maximum in-car glance duration (s), scroll buttons left (10 screens, N=12), r^2=0.888, RMSSD=4.799. Bars represent 95% confidence intervals.

A significant effect of the number of menu items was observed on the number of over-2.0-second in-car glances, $F(2,20) = 17.330$, $p<.001$, partial $\eta^2 = .634$. The significant mean difference between 6 and 12 item tasks was 3.73 glances, $p<.001$; and between 6 and 9 item tasks, 1.86 glances, $p=.026$. There was no significant main effect of menu but a significant interaction between menu and items, $F(2,20) = 6.997$, $p=.005$, partial $\eta^2 = .412$. The interaction seems to suggest that Grid is worse for 6 item displays, whereas List is worse for 12 item displays.

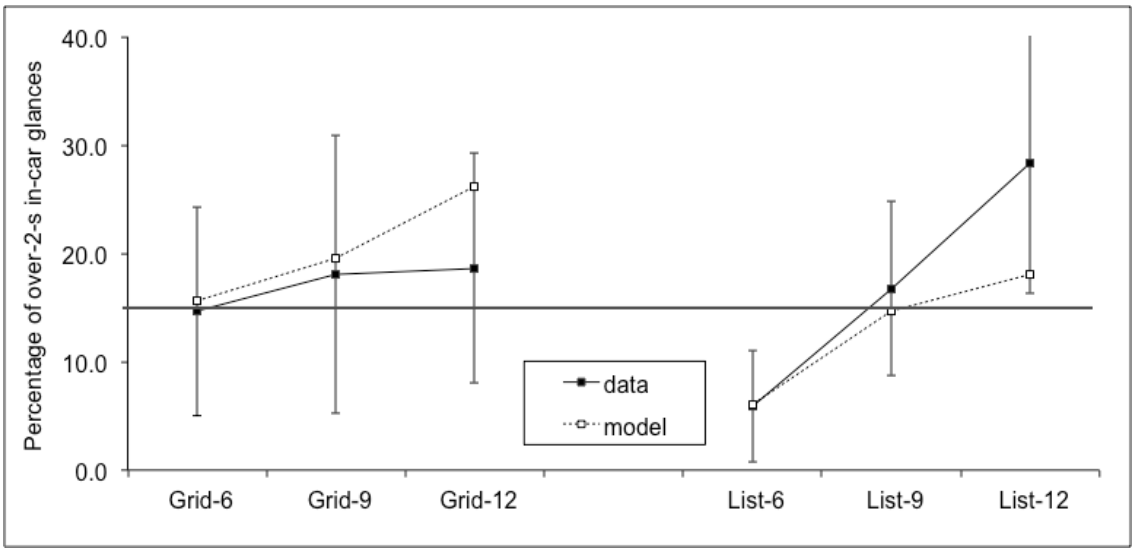

Figure 13. Percentage of over-2.0-second in-car glances, scroll buttons left (10 screens, N=12), r^2=0.421, RMSSD=1.037. Bars represent 95% confidence intervals. NHTSA (2013) verification threshold illustrated at 15.0 percent.

The number of menu items had also a significant effect on the percentage of glances over 2 seconds, $F(2,20) = 6.473$, p=.007, partial $\eta^2 = .393$ (Figure 13). The mean difference from 6 to 9 item tasks was 7.88 percentage points, p=.037, whereas the mean difference between 6 and 12 item tasks was 13.89 percentage points, p=.015. There was no significant main effect of menu but a similar significant interaction between menu and items as with the number of over-2.0-second glances, $F(2,20) = 5.369$, p=.014, partial $\eta^2 = .349$. The direction of the relative trend on the number and percentages over 2.0 second glances was predicted fairly by the model although the magnitudes are again somewhat off and the model predicted lower percentages for List-12 and higher percentages for Grid-12 than what was observed. In this case, the List-6 task would pass the NHTSA (2013) criterion on the percentage of over-2.0-second glances (max 15% for 85% of the participants), as predicted.

4.3 Discussion

Experiment 2 repeated the findings of Experiment 1 but with better fit on the relative trends. Besides the more obvious effects of the number of items to encode, this time the List menu structure gained more advantage compared to Grid in the predicted as well as observed glance numbers and durations. This advantage derived from the placement of the scroll buttons to the left side of the screen and thus, shorter encoding times, as expected. However, unexpectedly also the Grid got advantage from this button placement. This could be due to the reduced distance between the buttons and the driving environment, as the focus typically shifted back to driving after a button press.

There are still critical differences between the predictions of the model and the observations, in particular regarding the safety-critical maximum in-car glance durations and the numbers of long, over-2-second glances. The fit on the relative trend for the maximum glance durations is fair but the predicted maximum glance durations are much higher than observed, and the observed differences between the menu structures did not become significant with this sample size. These observed differences were also much smaller than the predicted. What is important to note, the data tells that the Grid menu structure is worse than List on the long glances only for the 6-item displays, but on displays with 9 or 12 items, the difference becomes insignificant. The model predicts advantage also for the List-9 and List-12 layouts compared to Grid. It could be that the denser layout of the titles in List-9 and List-12 encourages encoding more items per glance than in Grid and thus, undermining the positive effect of shorter encoding times.

Also for the over-2-second long safety-critical glances, the observed interaction effect seems to suggest that Grid is worse for 6 item displays, whereas List is worse for 12 item displays. However, the large SEMs could suggest that 9 or 12 items per screen (i.e., 90/120 item tasks) are too much for these types of in-car search tasks, regardless of the menu

structure. What is important, both the data and the model suggest consistently that the List menu structure with 6 items, with scroll buttons near the items, may be advantageous with respect to drivers' visual sampling performance.

5 GENERAL DISCUSSION

In this paper, we investigated how drivers perform visual sampling while driving, focusing on various effects of grid versus list layouts and the number of menu items on an in-car touch display. The central question explored is how drivers share visual resources temporally between the search and driving tasks. Both the empirical study and the general visual sampling strategy proposed in this paper offer explanations for the empirical findings in several ways.

5.1 General Findings and Implications

The empirical data as well as the model seem to suggest that increases in in-car task duration may not only increase the total number and duration of the in-car glances, but may also increase the individual glance lengths—a critical result given the potential hazards of long in-car glances while driving. This effect has been observed in previous studies (e.g., Lee et al., 2012; Kujala and Saariluoma, 2011) but a theoretical explanation for the effect has been lacking. Our proposed strategy and model seem to offer a plausible explanation for this phenomenon, as described below.

The findings indicate that as the number of on-screen items increases, the task time increases proportionally, with higher total in-car glance duration and larger number of in-car glances. The higher total glance duration is a basic set size effect and this is what our model predicts. The number of glances increases in our model because of the time limit for a single in-car glance. Drivers seem to interleave search and driving efficiently in general, in the sense that mean in-car glance durations are kept well below the threshold of 2.0 seconds that

is linked to increased crash risk in real traffic (Liang et al., 2012). A plausible explanation is that they succeed in this by using their perception of time to determine when their searching has reached a temporal "limit" at which time they must switch back to driving.

Due to the upward adjustment of the time limit after each in-car glance when the driving stays stable, the longer the task (i.e., the more in-car glances), the higher and riskier the time limit is able to grow during an in-car search task (more upward adjustments can be made). Because of the inbuilt delay and noise in the human time perception mechanism (Taatgen et al., 2007), larger time limits translate to a greater chance of overlong glances. The dynamic adjustment of the time limit following the model of Wierwille (1993) together with the inbuilt delay and noise in the human time perception mechanism might also explain some of the variance in the observed in-car glance durations in studies of in-car multitasking in highly controlled driving scenarios (e.g., Horrey and Wickens, 2007).

The List structure has an advantage regarding total glance time and the total number of glances over Grid, in particular when the scroll buttons are close to the list of items and the number of items increase. Our model is able to predict this and the effect can be explained by the titles being closer to each other in List, reducing encoding time from item to the next in the assumed nearest uninspected item next search strategy. The empirical findings give support for one of these key assumptions of the search strategy; when performing search in unordered menus while driving, drivers seem to be able to inhibit return to already visited items. Only a few of the randomly selected screens included gaze revisits on visited items.

The data as well as our model suggest the List-6 alternative is the least distracting in terms of the safety-critical long in-car glances in any case. The List-6-task with 60 search items in total is possible to be conducted in such a small number of glances (in particular when the scroll buttons are placed left near the titles and the driving view) that the time limit won't grow as high as with the other menu structures. Our model suggests the advantage of

List-6 over Grid-6 with the same total number of items and with scroll buttons close to the items is caused by the decreased distance between the search items in the list-style menu compared to the grid-style menu, and thus, lower total task durations and total number of glances.

The model and the data seem to indicate the placement of the scroll buttons to the left-hand side of the display decreases the mean and maximum glance durations, as well as the number and percentages of over-2.0-second in-car glances compared to the right-hand side (for both List and Grid). Our model is able to predict this in particular for the long glances towards the List menus. The advantage of placing the scroll buttons to the left in the List condition arises from their proximity to the titles for reducing encoding time for screen change. In addition, for both menus, left-side buttons decrease the length of the shift of visual attention from the display to the driving scene after a screen change. It might be noted, however, that this effect may be mitigated or eliminated with practice if the touch screen scroll buttons were replaced with physical buttons, freeing visual resources (e.g., Burnett and Porter, 2001).

Regarding the NHTSA (2013) criteria for in-vehicle electronic devices, the List-6 with scroll buttons next to the list items was the only task in our studies that would have passed the maximum 15% rate of glances over 2 seconds. The model was able to correctly predict this result. None of the tasks would have passed the criterion for total glance time (maximum 12.0 s) and all tasks would have passed easily the criterion for mean glance duration (maximum 2.0 s). The outcomes on total glance times are not as demanding to predict as the first one because the total glance times can be estimated with basic tools that utilize knowledge of human visual processing, such as GOMS (Card et al., 1983) and SAE-J2365 (2002). The pass of all in-car tasks on the mean glance duration criterion is not either hard to predict because according to Wierwille's (1993) visual sampling model, drivers try to

keep the mean in-car glance durations below 1.6 seconds in all traffic scenarios. More recent studies in simulators as well as in the field (as well as our current experiments) have indicated that for some reason the mean in-car glance durations are still kept between 0.5 to 1.6 seconds regardless of the type of the in-car task. For example, the naturalistic driving study with a sample of 100 drivers by Klauer et al. (2006) indicated similar in-car glance duration distributions.

The empirical results illustrate that the visual design of in-car displays can have a significant impact on the potential for visual distraction. Task length and the spatial separation between interaction elements, especially those encoded sequentially, arise as two of the critical factors for the probability of in-car glances to exceed the safe glance limits in this context. The findings suggest that visual designers should try to minimize task duration as well as the durations of all visual encoding steps required for the in-car task. This means, for example, that the number of available menu items should be limited and that the distance between interaction elements encoded one after another in a task sequence should be minimized to a level where clutter is still avoided. Given a prolonged search task, an extended estimate of a safe in-car glance duration, inaccuracy in driver's time perception ability, and a longer individual encoding step near the end of a glance, milliseconds can truly matter in this context (Gray and Boehm-Davis, 2000). The idea of minimizing visual encoding steps relates to the idea of Janssen et al. (2012) of providing shortest possible "*action sets*" and thus, natural breakpoints to encourage task switching and reduce distraction by secondary tasks. In short, shorter visual encoding steps should give more room for breakpoints.

5.2 Limitations and Further Research

Our current understanding of drivers' visual sampling strategies in the model has several limitations. It overestimates somewhat the number of in-car glances per task in general. A

plausible explanation is that the human drivers preferred to extend a glance to press a scroll button as a natural breakpoint for a task switch even if the time limit was close or passed (Janssen et al., 2012), whereas our model tended to move attention back to driving immediately after exceeding the time limit. The underestimated number of overlong glances for List menus supports this explanation in particular when the scroll buttons were farther away from the items (Experiment 1). It is also possible that the human drivers were able to encode more titles per fixation than the model, giving similar fixation durations per title and thus similar total glance times, but with the model splitting the encoding steps into greater number of fixations and thus greater number of glances. Ojanpää et al. (2002) have shown that, for vertical lists in particular, the word identification span covered with a single fixation could be even 4-5 words, but this increases fixation (i.e., encoding) time compared to a single word.

The model predicted somewhat greater relative disadvantage for the Grid than what was observed. One plausible explanation is the local density effect (Halverson & Hornof, 2004), which suggests that people spend less time per word searching sparser layouts as opposed to denser layouts. It could also be that the denser layout of the titles in List-9 and List-12 encouraged encoding more items per glance than in Grid. This is a possible strategic choice that is not built in our current model. Besides the higher number of items to encode, there may be additional set-size effects related to the difficulty of discriminating and/or encoding the stimuli (see, e.g., Palmer et al., 2000), which are not currently represented in the model.

Another limitation relates to having in-car search tasks of varying complexity. Some tasks may result in cognitive capture (Blanco et al., 2006) or might have other properties than the task used here, such as reading news or Facebook posts, or tasks with multiple decision-making elements. Other search tasks might evoke the central menu performance phenomena

as well as directed search strategies (see Bailly et al., 2014), whereas we have only considered exhaustive visual search with target absent and in unordered menus. The items changed for each screen per task, and thus, no practice effects were incorporated in the model. The timing and control of eye movements has been found highly adaptive to varying tasks and demands (Sims et al., 2011), and the learning effects associated with different types of ordered and static menus should be modeled. Here, our main focus and contribution was on the model of drivers' gaze allocation and timing strategy between the primary task of driving and a secondary in-car task. We expect that this strategy should be generalizable across in-car tasks but certainly this generalizability needs to be better understood. The strategy and the cognitive model based on it can be easily extended to evaluate other types of in-car task layouts, but each in-car task requires a specific task model.

Yet another limitation is the simplicity of the driving scenario: Although we have used the standard and fairly simple NHTSA (2013) testing scenario which can help elucidate specific behaviors, it remains to be seen how these strategies would generalize to more complex and more realistic driving scenarios. However, increase in driving task demands, such as using a curved road, should affect the stability of the vehicle, which should reduce the maximum time limit available for an in-car glance in our model. This behavior would correspond to the visual sampling model of Wierwille (1993), suggesting that drivers reduce individual in-car glance durations according to the demands of the driving task.

Future studies of visual sampling while driving will also need to take into account drivers' individual differences and individual task and time-sharing strategies that can be seen in the eye-tracker's video data. Janssen and Brumby (2010) have shown that people sometimes strategically control the allocation of attention in multitasking to meet specific performance criteria. For our purposes, for example, one might imagine strategic tradeoffs between stable driving and finishing a screen by a slightly longer in-car glance than the

current time limit would allow, as discussed above. The stability parameter in the model could be adjusted for accommodating individual differences in these priorities as in Distract-R (Salvucci, 2009). The model seems to overestimate the maximum glance durations in particular for Grid and for the higher number layouts (9 and 12). This could suggest there is some additional chunking strategy in work, limiting the maximum time limit for a single in-car glance (Janssen and Brumby, 2010).

Besides the cautious strategy of resetting the glance time limit always back to the lower limit of 500 ms by Wierwille (1993) after each instance of instable driving, other plausible strategy would be to retrieve the latest successful time limit and reset there. As in Salvucci et al. (2006), this would simulate behavior of the driver trying to determine on the basis of accumulating experience the optimal (the longest safe) length of in-car glance duration depending on the visual demands of the particular driving scenario. More stable factors, such as driving experience (Wikman et al., 1998) and the tolerance of uncertainty (Furnham & Marks, 2013), may also provide an avenue for understanding how individual drivers determine their particular point in the space of tradeoffs.

Uncertainty, event expectancies, and internal task state estimates (Johnson et al., 2014; Wickens, et al., 2001; Senders et al., 1967), saliency, as well as the expected effort and value (reward) of gathering visual information from a particular source (Sullivan et al., 2012; Wickens, et al., 2001) can certainly play a role in driver multitasking behavior and may provide alternative venues or improve the current model for explaining the empirical findings. The current model is based on the theory of threaded cognition in multitasking by Salvucci and Taatgen (2008) and does not require modeling of uncertainty or internal task state estimates (for now at least) or an explicit central executive process in multitasking situations; instead, a straightforward “threading” mechanism suffices to interleave the resource processing between tasks.

Beyond developing a better understanding of how drivers perform in-car visual search, we would also like to incorporate this knowledge in computational tools that help to quantify and use this knowledge. Specifically, design tools such as Distract-R (Salvucci, 2009) are currently capable of predicting drivers' performance on defined tasks with a particular in-car user interface; further knowledge of visual in-car sampling can greatly augment the functionality of such systems. The predictions of these tools would ideally be used in conjunction with standard experimentation as well as broader guidelines (e.g., NHTSA, 2013). Our hope is that these systems ultimately help to pinpoint the distracting visual features of in-car display designs, and thus, in guiding the visual designer in creating better and safer user interfaces for multitasking behind the wheel.

Acknowledgements

This research was supported by OPTIMUM (Optimal Multitasking with Ubiquitous Media), a research project funded by TEKES (Finnish Funding Agency for Technology and Innovation), Nokia, Cassidian Finland, and University of Jyväskylä. The second author was supported in part by Office of Naval Research grant #N000140910096. We thank Eepsoft Oy for developing and sponsoring the driving simulator software, Juha Hämäläinen for programming the in-car search tasks, and Jussi Jokinen for providing valuable advice with regard to statistical analysis.

References

Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y., 2004. An integrated theory of the mind. *Psychological Review 111*, 1036-1060.

Bailey, B.P., Iqbal, S.T., 2008. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction*, 14, 1-28.

Bailly, G., Oulasvirta, A., Brumby, D.P., Howes, A., 2014. Model of visual search and selection time in linear menus. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'14),* 3865-3874.

Blanco, M., Biever, W.J., Gallagher, J.P., Dingus, T.A., 2006. The impact of secondary task cognitive processing demand on driving performance. *Accident Analysis & Prevention, 38*(5), 895-906.

Burnett, G.E., Porter, J.M., 2001. Ubiquitous computing within cars: designing controls for non-visual use. Int. J. Human-Computer Studies, 55, 521-531.

Card, S., Moran, T.P., Newell, A., 1983. *The Psychology of Human Computer Interaction*. Lawrence Erlbaum Associates.

Chisholm, S.L., Caird, J.F., Lockhart, J., Fern, L., Teteris, E., 2007. Driving performance while engaged in MP-3 player interaction: Effects of practice and task difficulty on PRT and eye movements. In *Proceedings of the 4th Intl Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design* (pp. 238-245). Iowa City, IA: University of Iowa.

Furnham, A., Marks, J., 2013. Tolerance of Ambiguity: A Review of the Recent Literature. *Psychology, 4*, 717-728.

Gray, W.D., Boehm-Davis, D.A., 2000. Milliseconds matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied*, 6, 322-335.

Halverson, T., Hornof, A.J., 2007. A minimal model for predicting visual search in human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'07)*, 431-434.

Halverson, T., Hornof, A.J., 2004. Local density guides visual search: Sparse groups are first and faster. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting.*

Horrey, W.J., Wickens, C.D., 2007. In-vehicle glance duration: Distributions, tails and a model of crash risk. *Transportation Research Record, 2008*, 22-28.

Janssen, C.P., Brumby, D.P., 2010. Strategic adaptation to performance objectives in a dual□task setting. *Cognitive Science, 34*(8), 1548-1560.

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.

Jeon, M., Gable, T.M., Davison, B.K., Nees, M.A., Wilson, J., Walker, B.N., 2015. Menu Navigation With In-Vehicle Technologies: Auditory Menu Cues Improve Dual Task Performance, Preference, and Workload. *International Journal of Human-Computer Interaction, 31*(1), 1-16.

Johnson, L., Sullivan, B., Hayhoe, M., Ballard, D., 2014. Predicting human visuomotor behaviour in a driving task. *Phil. Trans. R. Soc. B, 369*, 1-7.

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* (DOT HS Rep. 810 594). U.S. National Highway Traffic Safety Administration, Washington DC.

Klein, R., 2000. Inhibition of return. *Trends in Cognitive Sciences, 4*, 138-146.

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

Lee, J.D., Roberts, S.C., Hoffman, J.D., & Angell, L.S., 2012. Scrolling and driving: How an MP3 player and its aftermarket controller affect driving performance and visual behavior. *Human Factors, 54*, 250-263.

Liang, Y., Lee, J.D., Yekhshatyan, L., 2012. How dangerous is looking away from the road? Algorithms predict crash risk from glance patterns in naturalistic driving. *Human Factors, 4*, 1104-1116.

National Highway Traffic Safety Administration, 2013. Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices (NHTSA-2010-0053).

Ojanpää, H., Näsänen, R., Kojo, I., 2002. Eye movements in the visual search of word lists. *Vision Research, 42* (12), 1499-1512.

Palmer, J., Verghese, P, Pavel, M., 2000. The psychophysics of visual search. *Vision Research, 40*, 1227-1268.

Posner, M.I., Cohen, Y., 1984. Components of visual orienting. In H. Bouma & D. Bouwhuis (Eds.), *Attention and Performance Vol. X* (pp. 531-556). Erlbaum.

Pylyshyn, Z.W., 1989. The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition, 32*, 65-97.

Salvucci, D.D., 2009. Rapid prototyping and evaluation of in-vehicle interfaces. *ACM Transactions on Human-Computer Interaction*, *16*, 9:1-9:33.

Salvucci, D.D., 2006. Modeling driver behavior in a cognitive architecture. *Human Factors*, *48*, 362-380.

Salvucci, D.D., 2001. An integrated model of eye movements and visual encoding. *Cognitive Systems Research*, *1*, 201-220.

Salvucci, D.D., Markley, D., Zuber, M., Brumby, D.P., 2007. iPod distraction: Effects of portable music-player use on driver performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 243-250). New York, NY: ACM Press.

Salvucci, D.D., Taatgen, N.A., 2008. Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review*, *115*, 101-130.

Salvucci, D. D., Taatgen, N. A., Kushleyeva, Y., 2006. Learning when to switch tasks in a dynamic multitasking environment. In *Proceedings of the Seventh International Conference on Cognitive Modeling* (pp. 268-273). Trieste, Italy: Edizioni Goliardiche.

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, No. 195*, 15-33.

Sims, C.R., Jacobs, R.A., Knill, D.C., 2011. Adaptive allocation of vision under competing task demands. *The Journal of Neuroscience, 31*(3), 928-943.

Society of Automotive Engineers, 2002. SAE-J2365 Calculation of the Time to Complete In-Vehicle Navigation and Route Guidance Tasks. SAE, Warrendale, PA.

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

Sullivan, B.T., Johnson, L., Rothkopf, C.A., Ballard, D., Hayhoe, M., 2012. The role of uncertainty and reward on eye movements in a virtual driving task. *Journal of Vision, 12*(13), 1-17.

Taatgen, N.A., Rijn, H. v., Anderson, J.R., 2007. An Integrated Theory of Prospective Time Interval Estimation: The Role of Cognition, Attention and Learning. *Psychological Review, 114*, 577-598.

Wickens, C.D., Helleberg, J., Goh, J., Xu, X., Horrey, W.J., 2001. *Pilot Task Management: Testing an Attentional Expected Value Model of Visual Scanning* (ARL-01-14/NASA-01-7). Savoy, IL: University of Illinois, Aviation Research Lab.

Wierwille, W.W., 1993. An initial model of visual sampling of in-car displays and controls. In A.G. Gale, I. D. Brown, C. M. Haslegrave, H. W. Kruysse, and S. P. Taylor, (Eds.), *Vision in Vehicles IV* (271-279). Elsevier Science, Amsterdam, NL.

Wikman, A.S., Nieminen, T., Summala, H., 1998. Driving experience and time-sharing during in-car tasks on roads of different width. *Ergonomics, 41*, 358-372.

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 (pp. 99-119).* New York: Oxford.

Vitae

Tuomo Kujala is a Post-Doctoral Researcher at the Department of Computer Science and Information Technology in the University of Jyväskylä. He obtained his PhD degree in Cognitive Science in 2010 at the University of Jyväskylä, Finland.

Dario Salvucci is a Professor of Computer Science in the College of Computing and Informatics at Drexel University. He obtained his PhD degree in Computer Science in 1999 at Carnegie Mellon University.

Appendix 1. Example of search targets and the orders of tasks (for participant P02).

# of items	Search target	Target screen/position of the target on the screen
	GRID	
12	Bad Clues	2 / 8.
	Deafening Feelings	6 / 10.
	Hospitality To Make You Cry	7 / 8.
6	Silver Restrictions	4 / 2.
	Promises Are Just A Start	12 / 4.
	Ruin Is The Best	14 / 2.
9	Accusation Of The Century	3 / 2.
	Imitations Are For Girls	8 / 7.
	Beautiful Nature	9 / 8.
	LIST	
12	Modern Distractions	2 / 8.
	Summer Dreams	6 / 10.
	Frowns From Hell	7 / 8.
6	Feelings For A Dollar	4 / 2.
	Giving Em Scandals	12 / 4.
	Ideas Flashing Before Me	14 / 2.
9	Karma Moves	3 / 2.
	Giving Em Scandals	8 / 7.
	A Flash Of Wisdom	9 / 8.