AGELAB
Life Tomorrow

Technical Report 2013-17A

The Effects of a Production Level "Voice-Command" Interface on Driver Behavior: Reported Workload, Physiology, Visual Attention, and Driving Performance

Limited Release Version 2013-17 Issued: November 4, 2013
Minor Revisions for Public Version 2013-17A: November 18, 2012

Bryan Reimer, Bruce Mehler, Jonathan Dobres &
Joseph F. Coughlin

Abstract - This report details the rational, methods, and of an on-road study assessing perceived workload, physiological arousal, visual attention, and basic driving performance metrics while drivers engaged in a number of tasks with a production version, in-vehicle voice-command system. The same metrics were also evaluated while participants carried out an implementation of the manual radio tuning reference task (Driver Focus-Telematics Working Group, 2006) and three levels of an audio-presentation / verbal response delayed digit recall task (n-back) that is known to produce graded levels of cognitive demand. Extensive training on all tasks was provided prior to assessment under highway driving conditions. Results for an analysis sample of 60 drivers equally distributed across both genders and two age groups (20-29 and 60-69) are presented. Depending on the task assessed and measure evaluated, both positive features and concerns associated with the use of the voice interface were identified. Physiological arousal during the voice tasks was comparable or lower than that observed during the more difficult level of manual radio tuning task as measured by skin conductance and heart rate, respectively. Perhaps most notable was the identification of a high level of visual demand / engagement during selected tasks such as the use of the voice-command interface for entering addresses into the navigation system. It also appeared that different age / gender groupings tended to interact with the voice system in different ways.

These findings highlight that implementations of voice interfaces can be highly multi-modal and are not necessarily free of visual-manual demands on attentional resources. If one were to apply the current National Highway Transportation Safety Administration (NHTSA) visual-manual distraction guidelines to the tasks assessed, a number of "voice" interactions would not meet the total off-road glance time criteria of the guidelines. While these data were not collected in full alignment with NHTSA's simulation-based guidelines, the overall structure and metrics are similar, and so this work raises a number of important questions. It is clear that visual demand needs to be considered in the design of multi-modal voice interfaces. This highlights the question of how an acceptable level of visual demand should

Massachusetts Institute of Technology > AgeLab > 77 Massachusetts Ave, E40-291, Cambridge, MA 02139
Phone: 617.452.2177 > fax: 617.258.7570 > reimer@mit.edu > agelab.mit.edu

be defined in the context of multi-step and extended task time interactions that characterize activities involving voice-command interfaces. Finally, the results illustrate the necessity for additional research assessing the generalizability of these findings to other production level and hand-held "voice" interactions, and in developing methods of quantitatively assessing the net attentional costs and benefits of providing drivers with information across different modalities. Voice interactions can play an important role in the vehicle environment. Optimizing the selection of activities in which the driver utilizes voice interaction and the appropriate design of displays will help to maximize driver attentional focus towards information necessary for vehicle operation, while allowing, where appropriate, interactions with interfaces for comfort, convenience and communication functions.

Primary Study Contact:

Bryan Reimer
Research Scientist
Phone (617) 452-2177
reimer@mit.edu

Table of Contents

Note: Extensive appendices covering the areas listed below were developed as part of this report and should be considered as part of a single document for citation purposes even if they appear as separate files due to size considerations when provided in electronic form:

Appendix A: Glance-To-Device Analysis

Appendix B: Error-Free Task Analysis

Appendix C: Trial Comparison Analysis

Appendix D: Orienting Response Coding

Appendix E: Baseline Period Analysis

Appendix F: Glance Coding Guidelines & Procedures

Appendix G: Task Performance Coding Guidelines

Appendix H: Detailed Experiment Protocol & Task Scripts

Appendix J: Questionnaires

Appendix K: Questionnaire Data

Appendix L: Misc. Information

The Effects of a Production Level "Voice-Command" Interface on Driver Behavior: Reported Workload, Physiology, Visual Attention, and Driving Performance

Introduction

Background

Drivers continue to demand increased connectivity and more advanced entertainment options while underway. While automotive manufactures strive to provide drivers with convenient, safe, easy to access information to meet this growing demand, there remains no well-established method for optimally achieving this goal. Over the past several years, there has been a shift in automotive driver-vehicle interfaces (DVIs) from purely visual-manual interactions to voice-based or voice-assisted interaction. However, few DVI functions are presently controlled entirely through voice commands. At minimum, most, if not all, current voice-based in-vehicle systems require manipulation of a "push-to-talk" button. This characteristic is clearly evident in a recent advertisement for the Chevrolet Sonic with Apple Siri Eyes-Free iPhone integration (McCann Erickson, 2013), in which the narrator refers to "the button to end all buttons". While the advertisement is designed to highlight the conversational aspects of the HMI and the potential for reducing manual interactions, it is readily apparent that the driver does in fact need to manipulate a "push-to-talk" button to go "hands-free". Upon close inspection of the video, it can also be observed that the driver engages in a visual confirmatory glance off the roadway to orient to the button location. This clearly illustrates that even one of the latest, most modern voice interface systems continues to involve and evoke some elements of traditional visual-manual interaction. While an "eyes-free" system may be found to successfully minimize certain visual-manipulative demands, other in-vehicle voice-based interactions appear to place significant demands on these resources. For instance, in some voice enabled navigation systems, an address is entered into the interface verbally, but a list of candidate addresses is shown on the phone application or an in-vehicle display. A button press or additional voice interaction is then required to make a selection from this list. The demands associated with most, if not all, current embedded vehicle, handheld and other portable device voice systems are therefore likely to be multimodal.

Research has been directed for some time at developing an understanding of and assessing the safety, usability and demand related aspects of voice-interaction in the vehicle. Areas of interest have included hands-free phone conversations (for an extensive review see Horrey & Wickens,

2006), navigational guidance (Dalton, Agarwal, Freankel, Baichoo, & Masry, 2013; Jensen, Skov, & Thiruravichandran, 2010), comparisons of interface operational modalities (Carter & Graham, 2000), and the quality of the voice recognition (Kun, Paek, & Medenica, 2007). Particular emphasis in the literature appears comparing voice systems with handheld visual-manual based tasks (texting and phone dialing), voice-based smartphone applications, "Wizard-of-Oz" simulations , and a range of aftermarket and research-based voice systems in which there is a wide variation in the amount of detail provided on functional characteristics of the systems under evaluation (Barón & Green, 2006; Forlines, Schmidt-Nielsen, Raj, Wittenburg, & Wolf, 2005; Garay-Vega et al., 2010; Gärtner, König, & Wittig, 2001; Gellatly & Dinges, 1998; Geutner, Steffens, & Manstetten, 2002; Graham & Carter, 2001; Greenberg et al., 2003; Grothkopp, Krautter, Grothkopp, Steffens, & Geutner, 2001; Harbluk & Lalande, 2005; Hu, Winterboer, Nass, Moore, & Illowsky, 2007; Itoh, Miki, Yoshitsugu, Kubo, & Mashimo, 2004; Jamson, Westerman, Hockey, & Carsten, 2004; Lee, Caven, Haake, & Brown, 2000; Lee, Caven, Haake, & Brown, 2001; Maciej & Vollrath, 2009; Mazzae, Ranney, Watson, & Wightman, 2004; McCallum, Campbell, Richman, Brown, & Wiese, 2004; Neurauter, et al., 2009; Ranney, Harbluk, & Noy, 2005; Ranney, Mazzae, Baldwin, & Salaani, 2007; Strayer et al., 2013; Tsimhoni, Smith, & Green, 2004; Yager, 2013; Zhang et al., 2012) (A "Wizard-of-Oz" simulation is a research design where a participant interacts with a system that they are led to believe is autonomous but which is in reality operated partially or fully by an unseen human being.). Less research is available that covers the evaluation of the demands placed on the driver from engaging with production level, embedded in-vehicle voice interfaces (Carter & Graham, 2000; Chiang, et al., 2005; Harbluk, Burns, Lochner, & Trbovich, 2007; Owens, McLaughlin, & Sudweeks, 2010; Shutko, et al., 2009; Shutko & Tijerina, 2011).

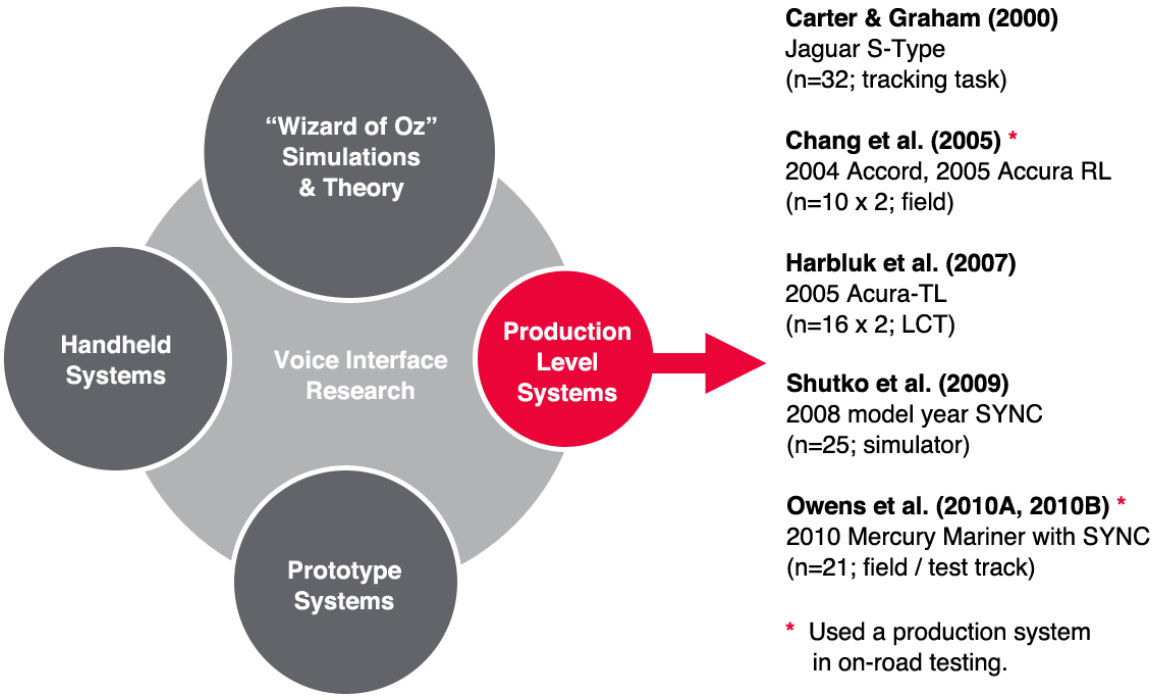


Figure 5: Illustration of a taxonomy of voice interface research and key studies that have addressed production level automotive voice systems. (Note: figure numbering starts at 5.)

In the literature on the safety of voice interfaces, various experimental approaches (task selection, study conditions, sample characteristics, interface selection, etc.) have produced some conflicting findings. For instance, (Neurauter, et al., 2009; Shutko, et al., 2009) suggest that the time a driver's eyes are off the road is significantly less with a voice enabled texting system as compared to a visual-manual system. In contrast, (Yager, 2013) reported that eye gaze to the forward roadway decreased significantly compared to a baseline driving condition regardless of whether a manual or voice-based (iPhone running Siri; Samsung phone with Vlingo) application was used. There appears to be more consistency in other efforts centered around embedded in-vehicle systems (Chiang, et al., 2005; Itoh, et al., 2004; Shutko, et al., 2009), where research shows significantly less visual demand for voice-activated tasks vs. traditional methods of visual-manual control. Interesting, many experiments have approached the evaluation of voice enabled systems in comparison to visual-manual alternatives without including a complementary comparative analysis to single task (baseline) driving (Jamson, et al., 2004; Neurauter, et al., 2009; Zhang, et al., 2012). Other research has provided comparisons of the distraction potential of voice interface demands relative to "just driving" (Maciej &

Vollrath, 2009; Owens, et al., 2010; Ranney, et al., 2005; Shutko, et al., 2009; Yager, 2013). A careful review of the data from the latter studies raises some significant questions.

While a complete review of the literature to date on the safety of voice interfaces is beyond the scope of this report, a select set of papers warrants particular consideration. Barón and Green (2006) completed a partial review of the human factors experiments published between 1998 and 2005 on the use of in-vehicle voice (speech) systems. Across the 15 publications they considered, sample sizes ranged from 4 to 48 (mode = 24), involved a mix of "true" speech systems and "Wizard of Oz" simulations, and were completed in low and moderate fidelity simulators, test tracks, and field driving conditions. The authors report that comparisons between studies are difficult, but "generally, driving performance is better (fewer lane departures, steadier speed), workload is less, and there is less time spent looking away from the road when using speech as opposed to manual interfaces." Furthermore, they report that optimal interface choice (voice vs. visual-manual) depends on the implementation of the interface for a task, the driving situation, and the individual.

In a follow-up article, Lo and Green (2013) provide an overview of examples of automotive speech systems, a review of a select set of more recent studies on the effect of speech interfaces on driving, reference to key design standards, and methods for evaluation. The overall focus on evaluation of speech interfaces appears to most strongly acknowledge the effort of Dybkjaer, Bernsen and Minker (2004) who proposed four broad aspects (and variables) of a system to be considered in evaluation: recognition, language understanding, perception of the speech synthesizer, and measures of the systems performance (technology and operator). Lo and Green (2013) conclude that "in terms of performance while driving, there is no standard or common method for evaluating speech interfaces, with evidence from bench-top, simulator, and on-road experiments being used." The authors note that performance on the speech task will depend upon the driving demand, a variable that is often not quantified.

Research with Simulated Interfaces

In a test track study of 36 participants across three age groups, Ranney, Mazzae, Baldwin and Salaani (2007) investigated a set of voice driven navigation tasks using a "Wizard-of-Oz" 511 system. Deteriorations in all aspects of driving performance were found when drivers engaged with the simulated voice interface. While the authors set out to assess visual behavior along with vehicle control, driver decision making, and target detection, they did not report on the eye position measures collected because the "confidence for the eye gaze data was less than 40%". Maciej and Vollrath (2009) looked at the performance of 30 participants across visual-

manual interactions and conceptually paired voice interface interactions (MP3 player and a prototype laptop-based voice-controlled music selector) in a driving simulator utilizing the lane change test. In comparison to baseline single task driving, the visual-manual and speech interactions generally impaired driving performance, with stronger impairments evident for the visual-manual interactions. The authors reported that in the manual control conditions, 30-40% of the mean task time was spent looking away from the lane change screen. Gaze away from the screen was reduced by the speech system to between almost none (music selection) and just under 20% (point of interest entry). The former finding provides support for the position that a voice interface can result in improvements in glace behavior away from the roadway, but the latter finding suggests that the interaction may not be cost-free relative to single task driving.

In an experiment on the Virginia Tech Transportation Institute's Smart Road, Neurauter, Hankey, Schalk and Wallace (2009) investigated drivers' behavior while performing three voice texting, three handheld visual-manual texting, and three voice-based destination entry tasks (full address, point of interest "nearest Starbucks", and point of interest category "gas station"). Handheld texting was completed on a touch-screen "smartphone" and the voice system was described as a "hands-free module" in the center stack. Data from 24 drivers collected across two age groups (18-30; 45-55) showed divergent performance on the texting task by age group and interface type. Among younger adults, 96% of the handheld texts and 85% of the voice texts were reported as successful. In comparison, the middle age group was successful in 82% of the handheld texts and 94% of the voice texts. Performance on the destination entry task, reported across the three types of activities, showed 92% and 89% accuracy for the younger and middle age groups, respectively. Lane deviations were more likely to occur during the visual-manual texting task (mean of .35 per participant) as compared to .03 and .05 for the voice-based texting and destination entry tasks respectively. While a significant interaction effect between task type and age was not reported, data presented suggests that the middle age drivers had significant difficulty maintaining lane positioning during the visual-manual texting task. Changes in driving speed were not evident across tasks. An interaction between age and task type on speed variance suggests that the middle age group had more difficulty controlling vehicle speed compared to the younger age group during the visual-manual texting tasks. However, speed variance of the middle-age adults dropped below that of the younger adults during the voice tasks. Overall performance generally showed improved vehicle control with the voice-based interactions versus the visual-manual texting. Eye glances, coded by manual, frame-by-frame video reduction, show that the duration of glances to the center stack (location of the hands-free module) for voice-based interactions were shorter than glances to the handheld device.

Consistent with this, eyes off-road time was significantly lower for the voice-based tasks (26%) than the handheld tasks (68%). While the interaction between task type and age on eyes off-road time was not significant, age did appear as a significant modifier. Younger adult drivers averaged 38% of the task time looking away from the road, while middle age adults spent 42%. Detailed differences in eyes off-road by task type were not reported. Subjective ratings of mental demand, frustration and situational awareness suggested significant benefits for the voice interface. However, the authors noted that the voice interface negatively impacted reported mental demand, frustration and situational awareness when compared to baseline driving. The authors summarized the effects of voice-based interactions as "equalizing performance" between the two age groups as compared to the "clear disconnect" the older sample displayed with manual texting.

A more recent study of 43 participants in the 16 to 60+ age range (majority in the 18-24 and 40-59 age groupings), compared voice-based and visual-manual based texting across two smartphones, the iPhone with Siri and a Samsung Galaxy Stellar running the Vlingo Android App (Yager, 2013). The Siri voice interface allowed for voice-initiated texting, content dictation and auditory message reading. The Vlingo Android App was less interactive and only allowed for the dictation of text as a single verbal input string, i.e. visual-manual initiation of texting and reading the content of texts. Participants completed five "short" text messaging tasks (uncorrected for content errors) using both voice interfaces and one of the two visual-manual interfaces (self-selected). Tasks consisted of sending a message (x1), reading and replying to a message (x3) and reading only (x1). Responses to a light detection task did not differ across tasks (sending, reading and replying, or reading only) or between the two voice-based systems and manual typing. Regardless of the texting interface, a significant decrease in response time (nearly two times slower) appears while texting in comparison to single task (baseline) driving. No statistical difference appeared in missed light detection events, mean driving speed, or standard deviation of lane position between any of the four conditions (baseline, voice Siri, voice Vlingo, and manual). While no differences in task completion times appeared between the interfaces in the text sending task, all three of the read and reply tasks independently show that voice-based texting (regardless of device) took longer to complete than visual-manual based texting. In the last read and reply task, the completion time of the Siri interaction also appears shorter than the Vlingo interaction. The same pattern appeared for the reading task. The percentage of time drivers spent looking at the forward roadway recorded by a faceLAB eye tracker decreased across all three conditions as compared to baseline. No difference in the percentage of time drivers glanced at the road appeared between the two voice systems and

visual-manual texting. However, it is important to note it is unclear to what extent a driver's self-selected positioning of the phone impaired the eye tracking system's ability to see the drivers face, i.e. many individuals tend to hold phones up in the vicinity of the top of the steering wheel to enter information while driving, while others tend to hold their phones near their lap. When phone placement impairs the eye tracker's ability to observe the driver's face, no eye positioning data is available. The authors do not acknowledge any adjustments of the data for this situation.

In a report covering a series of studies consisting of one each in a single task laboratory set-up (N= 38), driving simulation (N= 32), and an on-road experiment (N= 32), Strayer et al. (2013) compared interaction with a speech-to-text program to participants experiences listening to a radio station, listening to a book on tape, talking with a passenger, talking on a hand-held phone, talking on a hands-free phone, and a highly demanding surrogate task (OPSPAN) combining word memorization interspaced with math-verification problems. The speech-to-text program was presented in a "Wizard-of-Oz" paradigm in which an experimenter typed the participants' verbalizations into the program to eliminate any issues with speech recognition technology. Conceptually, the task as was presented as simulating a voice-based e-mail and text messaging system. Across a range of cognitive workload measures, the speech-to-text task generally ranked in the range of or higher than the phone tasks and markedly less than the OPSPAN task. This was most evident in the subjective ratings provided in the mental demand, temporal demand, and effort scales of the NASA-TLX. In terms of objective measures, in the driving simulator, the speech-to-text task was associated with relatively longer brake reaction times and longer following distance to a lead vehicle, the latter behavior perhaps representing a compensatory response to aid in managing the workload. In the on-road study, glances to hazard locations while engaged in the speech-to-text task were also nominally lower than other tasks with the exception of the OPSPAN. However, detection reaction task (DRT) reaction time scores during the on-road study were in the same range or even nominally lower for the speech-to-text task relative to the phone and passenger conversations. The authors make the intuitively compelling argument that as cognitive workload increases, available resources to respond to attention-demanding concurrent activities may decrease, and their data on increased brake reaction time and reduction in frequency of glances to hazard locations lend some support to this argument. At the same time, the degree to which the simulated speech-to-text task employed in this research effectively models a production voice interface design and other voice-command interactions beyond the Wizard-of-Oz e-mail composition scenario is unknown. It should be noted that the task was sustained over a continuous 10 minute period so

that event-related brain potentials (ERP) and DRT data could be collected. Other factors to be considered include characteristics of the system such as possible response delays (due in part to manual typed input by the experimenter), the user's ability to easily decode the output of the speech synthesizer given the modest level of voice quality, and the extent to which demand of continuous time pressure impact usability and workload. While it is conceivable that a voice-enabled e-mail composition may provide access to lengthy interactions, it seems unlikely that voice-command enabled interaction to change a radio station or enter an address into a navigation system would typically involve such continuous interaction and that many drivers would selectively choose to operate a system under such extended high temporal demand. The extent to which the level of diversion of attention observed under these conditions translates into a substantive safety risk for actual in-vehicle voice systems more generally remains to be established. The work does make a good case for why a better understanding of the nature of driver interaction with production level voice-interfaces in the automobile (embedded and portable) is needed.

Research with Production Level Systems

Published research on production level embedded in-vehicle voice interfaces is limited to a select set of vehicle systems. In the first reported research with a production system, Carter and Graham (2000) describe speech communication as "both hands-free and eyes free, allowing drivers to maintain visual attention on the road and their hands on the steering wheel". The paper reports on an investigation of the performance of a Jaguar S-Type speech recognizer in a laboratory setting. The sample of 32 participants was equally split by gender across a younger (21-35) and older (55-70) age group. Participants completed a PC-based tracking task designed to "mimic the important visual-manual tasks involved in driving while completing a peripheral choice reaction task and ten in-vehicle entertainment tasks across each of four interface configurations. Interfaces included standard button controls, steering wheel located button controls, speech with implicit feedback, and speech with explicit feedback (message center prompts). Tasks included operating the Tape Player, Radio, CD, Climate Control and Phone. In comparison to non-task periods, a decrease in tracking performance (lateral control) was observed for all interface conditions. Interaction with steering wheel-mounted buttons resulted in the most impaired performance on the tracking task, followed by the manual controls, and then both voice conditions, indicating an apparent relative advantage for the voice interface. Older drivers' tracking performance was more significantly impaired than the younger group during concurrent driving and operation of the in-vehicle entertainment functions. Results for collision data followed the pattern seen in the lateral control task. Reaction times to peripheral

targets where longest in the standard manual control and steering wheel-mounted button conditions. The two voice control conditions again showed an advantage over the manual interfaces in terms of reaction times. However, reaction times during all four conditions were longer than during the baseline task condition, indicating an attentional cost for engaging in the voice interaction. Task completion times for the two voice conditions were longer for the more traditional manual interface. While quantitative results on the NASA-RTLX were not reported, the authors note that the voice conditions were given the lowest workload rating for operating the interface while engaged in the "driving" task. In other subjective ratings, the authors reported that participants indicated that they would "prefer to have the speech control system in their cars than the manual controls". It was acknowledged that many participants experienced problems with the wheel controls utilized in this experiment and that improved designs may reduce effects.

In another laboratory PC-level simulation, Harbluk et al. (2007) evaluated a 2005 Acura-TL navigation system utilizing the Lane Change Task to investigate differences in driving performance in navigation tasks (point of interest and full address) performed using visual-manual and voice-based interfaces. In a between-subject design, 16 participants (24 to 58 years) were instructed to perform visual-manual entry tasks and 16 participants (21 to 48 years) voice-based tasks. Two difficultly levels of the point of interest and full address entry tasks were selected to provide a low and high level of complexity for both types of navigation task. Differences in complexity were varied by the amount of information the driver was require to input into the system. Mean deviation in the lane change path exceeded baseline for all task conditions (manual and voice). In the visual-manual condition, the mean deviation in lane change path was higher for the more complex tasks (point of interest and full address) as compared to the low complexity tasks. Complexity did not impact mean deviation of lane change path in the voice conditions. Similar results appeared for the initiation of the lane changes, except that the results of the lower complexity visual-manual full address entry followed that of the two higher demanding visual-manual conditions. In terms of task duration, the higher complexity tasks took longer than the lower complexity tasks to complete in both the visual-manual and voice conditions. Perhaps most notable in the results were the findings on the mean time to complete a secondary task. While low complexity address entry took an average of 6.6 seconds using the manual interface, it took 38.6 seconds in the voice interface mode; for the high complexity address entry, mean time for the manual interface was 45.2 seconds vs. 102.9 for the voice interface. The authors commented, "Although speech-based interfaces are increasingly popular choices for in-vehicle technologies, there are likely to be

advantages and disadvantages associated with their use in different applications (e.g., Tsimhoni et al., 2004)."

In the first reported field study of voice interfaces, Chang et al. (2005) reports on two experiments investigating driver behavior during destination entry considering a visual-manual interface and an interface with voice control. The first experiment, conducted in a 2004 Accord equipped with an embedded OEM navigation system, evaluated the behavior of 10 drivers (29-57 years of age; half engineers; 3 female). A point of interest based categorical address entry task and phone number based entry task were explored in both urban city streets and in freeway driving. Outcome measures included the number of entries, entry errors, glance-based metrics, and task time. Overall, fewer and shorter glances were observed to the navigation device with voice entry. However, even with the voice interface, glances to the navigation system during the point of interest task remain elevated compared to single task driving, reaching 15 to 17% (6.9-7.2 sec) of the task time in the freeway and city conditions respectively. The phone number entry method resulted in less visual demand, with 5 to 7% (1.8-2.3 sec) of the total task time spent looking at the interface. Interestingly, during the voice phone number entry tasks, more attention appeared to be allocated to "other" locations, resulting in relatively consistent percentages of eyes on-road time (76-79%) across all task types and operating conditions. Some decreases in lane keeping performance with the manual control interface were reported.

The second experiment reported in the paper (Chiang, et al., 2005) was conducted in a 2005 Acura RL equipped with the Acura navigation system. In contrast to the Accord navigation system evaluated in Experiment 1, the Acura navigation system supported the entry of full addresses, which required the stepwise entering of information, e.g. street name, followed by the address number, and city name. Verbal entry of the street name resulted in a display of street name choices on the navigation screen, with the user making a selection by saying a number. Speaking a city name produced a corresponding display of city choices to be selected from the list. The entry of destination information was again completed by ten participants, some of whom took part in Experiment 1 (34-50 years of age; 4 engineers; 5 female); the same city street and freeway driving locations were used. No statistical results are reported, however, the voice interface appeared to offer nominal advantages in mean single glance durations, mean total fixation time, and total glance time percentage on the roadway. The percentage of glance time to the display with the more complex, full address entry task increased over experiment 1, to 20 to 21% (10.4-11.9 sec) for the freeway and city conditions respectively. As in Harbluk et al. (2007), total task time for the full address entry method using the voice entry interface was markedly longer than for the manual entry method. For the visual-manual interface, mean entry

times were 34 seconds for both the city street and freeway conditions; for the voice interface, the corresponding times were 68 and 60 seconds. As was the case in Carter and Graham (2000), subjective ratings appeared to favor the voice interface in terms of how difficult each interface was to use, most notably for driving on the freeway.

In the first of two studies that evaluated voice interface activities with versions of the Ford SYNC® voice interface, Shutko et al. (2009) presented a simulation study of 25 manufacturer employees (18-65 years of age: 5 women) who reported being regular users of the SYNC hands-free system in their personal vehicle for accessing a personal music player and with a Bluetooth enabled phone. Data from two additional cases was collected, but dropped from the analysis, one due to voice recognition difficulties and the other due to "great difficulty using SYNC as configured for this experiment". Interaction with the voice system was compared to controlling a handheld music player with a display (e.g. Apple iPod) and a cell phone. The simulation buck was developed from a Lincoln Town car, retrofitted with an embedded 2008 model year SYNC system (Ford GAP AM/FM/CD audio system with integrated display and Ford Explorer steering wheel). A Motorola RAZR Bluetooth phone and an Apple iPod Nano were connected to the system to enable dialing and music search features through the SYNC system. In addition to voice-enabled tasks, participants utilized their own personal phones and music players for a set of handheld tasks. Participants were asked to complete seven paired tasks using both the embedded vehicle system and their handheld devices. The tasks included: select and play a specific song, select and play the first found song by an artist, select and dial a pre-selected contact, dial a familiar 10 digit number, receive an incoming call and reply with a short message, retrieve and review a text message (handheld phone – read; SYNC – listen through hands-free), and reply to a text message (handheld phone – keyed in response; SYNC canned reply). While the conceptual tasks were paired, i.e. the same general activity was performed using both interfaces, there were substantive differences in some of the actions. For instance, handheld song tasks involved "searching for a random song pre-selected by the experimenter from the participant's personal music device," while the paired SYNC hands-free device provided "direct request of a song". Similarly, texting with the handheld device required "keying in a response" while the paired SYNC hands-free device provided selection of a canned reply from a pre-defined list. Many of these activates appear to have been designed to highlight the efficiency of the SYNC system. Of the 350 trials (25 (subjects) x 2 (conditions) x 7 (tasks)), 7% (24) of the paired trials were not included in the analyses because of participant difficulties in completing one portion of the task (handheld or voice). Across all seven tasks, a significant difference in total task time, total eyes off-road time, and standard deviation of lane position

appears between tasks completed with the handheld interface and SYNC hands-free system. In all tasks except the 10-digit dialing task, the SYNC hands-free system enabled faster completion. Total eyes off-road time was lower for all of the SYNC hands-free tasks except for receiving an incoming call. Finally, a lower standard deviation of lane position appears for all SYNC hands-free tasks. In terms of task involvement impacting the volatility of driving speed (Max - Min) there were significant advantages reported for the two song selection tasks, the contact dialing task, and the text message reply task. Finally, the selection of an artist task, 10 digit dialing task, and text message review task with the SYNC hands-free interface resulted in faster reaction times to a pedestrian response scenario than with the paired handheld operations. The authors conclude, "As a hands-on-wheel, eyes-on-road system, SYNC provides a hands-free voice interface to safely control cell phones and personal music players while driving." While the aforementioned findings clearly favor the voice system over the hand-held device interactions, visual inspection of the individual participants' eyes off-road time (Figure 2 in (Shutko, et al., 2009)) suggests that following NHTSA's new visual-manual guidelines, i.e. at least 85% of the respondents (22 of the 25 participants) must complete a task with less than 12 seconds of off-road time (National Highway Traffic Safety Administration, 2013), that the SYNC hands-free 10-digit dialing and text messaging sending task assessed in this study would not meet this new governmental guideline. It does appear that it meets the original Alliance of Automotive Manufacturers (Driver Focus-Telematics Working Group, 2006) guideline (criterion 2.1 A) of under 20 seconds of total glance time to the device threshold for the tasks tested.

Owens, McLaughlin and Sudweeks (2010) reported on the behavior of 21 drivers, approximately half older (19-34 vs. 39-51 years), who were current owners of vehicles with the Ford SYNC® system in a field driving study. The experiment took place on a divided secondary road with a speed limit of 65MPH and was conducted in a 2010 Mercury Mariner equipped with the Ford SYNC® voice interface. The impact of different input modalities (voice control vs. handheld device) was reported for contact dialing, brief conversations, and playing a music track in comparison to baseline driving. In comparison to handheld operation, contact dialing and track playing tasks utilizing the voice interface interfered less with vehicle control, were completed faster, and received lower self-reported mental demand scores on the NASA TLX. Steering variance and maximum steering wheel velocity were larger for the handheld tasks compared to baseline driving. No significant differences in steering variance or maximum steering wheel velocity were reported between baseline driving and the three voice (SYNC) interface activities or handheld conversing. The voice interface also allowed for a more optimal orientation of the eyes towards the road than the handheld device. In comparison to all

conditions (baseline driving, handheld phone conversation and all voice interactions), the handheld contact dialing and handheld music track playing tasks resulted in more frequent, longer total and mean duration, and longer maximum duration glances to the interior of the vehicle. The voice music-playing task also had a longer mean glance duration then baseline. While the frequency of glances during the conversation task were reported in comparison to baseline, the authors note that there was no adjustment provided for the length of the conversational task, limiting the interpretation of these findings. Observation of the data reported in Figures 4 and 5 (Owens, et al., 2010) shows nominal, largely non-significant increases over baseline in visual demand associated with all of the voice interface conditions. Age appears to negatively impact lateral control during the visual-manual activities, but the impact on glance behavior during the same periods is inconsistent. While the authors center their presentation of findings on the advantages of voice control over visual-manual handheld interaction, they clearly acknowledge limitations associated with the limited experimental control employed in this study. In particular, they acknowledge that the selection of frequent users of the technology may under-represent differences in interactions with the interface types that might be observed across a more comprehensive sample.

Another report compared sending and receiving text messages using a hand held device and the SYNC in-vehicle voice system. This component of the study was carried out on a closed test track following the portion of the study focused on comparing voice with handheld control (Owens, et al., 2010). The samples considered in these reports are identical, except for one case that was dropped in the text messaging condition due to an equipment failure. Compared to the hand-held device, using the in-vehicle voice system to send "canned" text messages took less time, was rated as involving less mental demand, involved a lower number of glances, resulted in a lower total glance duration, and showed less performance degradation relative to baseline driving. It is clearly noted that using the voice system to send messages did result in more task-related interior glance time and higher reported mental demand than simply attending to the primary driving task. The authors note in the limitations section that future work might benefit from employing a larger and more varied sample than the one studied. They also state that evaluation on an open roadway might produce different results, as participants might feel that there is less risk on the test track.

Broad Research Objectives

An underlying aspect of the studies discussed above is the focus on a driver's visual orientation, driving performance, and self-reported demand while engaged in various types of voice dialog,

particularly in comparison to handheld operations. Given the importance of eyes-off-road time to safety (Klauer, Dingus, Neale, Sudweeks, & Ramsey, 2006), these are clearly key factors to consider in assessing the demands associated with voice systems. However, they may not provide a comprehensive and fully objective rating of the non-visual (cognitive) demands of such systems. The amount of "spare capacity" available to recognize and respond to surprise events when engaged with this type of interface is largely unexplored, as noted by Owens, McLaughlin and Sedweeks (2011) in their discussion of their study's limitations. In essence, although voice-based systems are intended to help keep drivers' eyes on the road, relatively little is known about the "holistic" visual, manipulative, and cognitive demand placed on the driver by interaction with production level voice-command systems under actual on-road driving conditions.

Research suggests that physiological indices of workload reflect an individual's investment of cognitive resources corresponding to task demand (Brookhuis & de Waard, 1993; Lenneman & Backs, 2009; Mehler, et al., 2012; Mehler, Reimer, Coughlin, & Dusek, 2009; Reimer & Mehler, 2011; Wilson, 2002). Physiological measures have been shown to be sensitive to subtle increases in demand before overt breakdowns in driving performance are observed (Mehler, et al., 2009). In contrast to earlier work where demands exceeded a driver's capability or willingness to engage in a secondary activity (Engström, Johansson, & Östlund, 2005), heart rate and skin conductance have been shown to scale relatively linearly with an increase in cognitive demand from an auditory presentation – verbal response working memory task (n-back) (Mehler, et al., 2012). In essence, the three levels of the n-back task create a three-stage ruler, e.g. low, moderate and high, against which the relative demand of other cognitive activities can be objectively and non-invasively scaled.

The degree to which demand placed on the driver through artificial secondary tasks such as the n-back relate to the demands of voice-based interactions or other non-visual activities is an open question. However, we hypothesized that the demands of voice-based interactions would fall between the lowest (0-back) and highest (2-back) levels of this secondary task. This hypothesis is framed by the 0-back task, a simple mirroring activity consisting of verbally repeating back auditorily presented single digit numbers, and the 2-back task, a more demanding activity that taxes working memory and which likely approaches the limits of most drivers' spare capacity (Ranney et al., 2011).

This report focuses on the presentation of detailed results from a field study conducted to measure drivers' perceived workload, physiological arousal, basic driving performance metrics,

and visual behavior while engaging in a number of tasks with a voice-based in-vehicle HMI, an implementation of the manual radio tuning reference task (Driver Focus-Telematics Working Group, 2006), and the three levels of the n-back task. The data were collected during a field experiment in which participants were given detailed training on the operation of the vehicle systems under study prior to the assessment of behaviors under highway driving conditions. As highlighted earlier, relatively limited data are currently available on how a broad-based sample of drivers interacts with a production level voice-based interface under actual on-road driving conditions. Therefore our initial objectives included the development of:

- Data highlighting physiological, visual attention, and driving performance measures as objective methods of measuring changes in cognitive workload associated with various voice-based tasks
- Analyses that relate the level of cognitive load involved with "everyday" voice enabled applications to scaled levels of cognitive workload established by the multi-level secondary task (n-back auditory delayed digit recall task)
- Findings that compare the relative levels of distraction associated with voice interaction and a traditional visual manipulative interface to complete the same functional task
- Findings that evaluate the level of difficulty involved in learning how to use a voice interaction interface
- Insights into where voice enabled tasks fall on a scale of acceptable cognitive workload
- As well as, consideration of the impact of age and gender on the points above

Formal Research Questions

One of the primary, if not the primary objective of introducing voice-command interfaces into automobiles is to reduce visual-manual distraction by allowing the driver to keep their eyes on the road and reduce or eliminate the need to remove a hand from the steering wheel while interacting with non-driving critical user interfaces. With this in mind, one of the planned research questions addressed in this study is an evaluation of the extent to which the voice-command interface in the test vehicle reduces the amount of visual-manual demand on the driver compared to doing the functionally equivalent task using the manual interface. For this purpose, direct comparison of two manual radio tuning tasks and the same tasks using the voice interface was developed. Commonly used methods for carrying out such a comparison

include a consideration of the level of visual engagement associated with a task and an examination of any impact a task has on various driving performance metrics. In addition, we consider self-reported workload and physiological measures of arousal that have previously proven sensitive to varying levels of cognitive demand in the driving environment as complementary approaches to assessing overall task demand associated with each interface type.

> **Question 1:** Does using the voice-command interface in the vehicle-under-test to control the radio result in reductions in driver distraction or workload, as assessed by various metrics, compared to carrying out the same tasks using the manual interface?

A second formal research question had to do with the extent to which the voice-command interface allows entry of full destination addresses into a navigation system at an acceptable level of demand upon the driver's attentional resources. A significant body of research has raised questions about the safety risk around manual entry of full address destination entry into a navigation system while underway, and many systems either lock-out manual destination entry while underway or present warning messages advising that this action should not be attempted while driving. To what extent does a representative production level voice-command interface resolve the issues associated with full address entry while underway? Since manual entry of a full address might be considered unsafe and was also locked-out while underway in the in-vehicle system under test, an evaluation point or points are required. Manual radio tuning has been adopted by both by the Alliance of Automotive Manufacturers (The Alliance) Driver Focus-Telematics Working Group (2006) and the National Highway Traffic Safety Administration (NHTSA) in the issuance of Visual-Manual Driver Distraction Guidelines (2013) as a reference for what is generally considered as a socially acceptable level of demand on a driver's attention. Therefore, we decided to compare the relative demand associated with using the voice-command system to enter a full destination address into the navigation system to a complex manual radio tuning task. To the extent that the impact on the driver is indistinguishable from or lower than that seen with complex manual radio tuning, this could be seen as a demonstration of advantages of this interface concept.

> **Question 2:** Does using the voice-command interface in the vehicle-under-test to enter a full destination address into the navigation system result in equivalent (or lower) levels of driver distraction or workload, as assessed by various metrics, compared to a complex manual radio tuning reference task?

As listed under the broad research interests for this study, we are interested in how various demographic factors are associated with how drivers interact with the various tasks and interface options presented in the study. Specific tests considering age group and gender were therefore included as formal questions.

> **Question 3:** Is there a significant effect of age on distraction or workload, as assessed by various metrics, on the behavior considered in this study?

> **Question 4:** Is there a significant effect of gender on distraction or workload, as assessed by various metrics, on the behavior considered in this study?

In the development of this work we aimed to assess these hypotheses with a large enough sample size to support a well powered design. While the selection of the sample size was based upon experience with data gathered in other areas of research on physiological reactivity to cognitive demands, it clearly does not have the necessary power to assess all potential outcome measures with the same degree of statistical power. In essence, the sample was not expected to provide sufficient sensitivity to fully assess all risk factors. One of our main expectations in developing this work was that the extended auditory-vocal interactions utilized by some voice command interface systems might place relatively high cognitive demands on the driver. Our expectation was that some voice interaction tasks might approach or exceed the 2-back task level of demand.

> **Question 5:** Using physiological measures previously shown to increase in magnitude in response to defined levels of cognitive demand, how do the various voice-interface based tasks scale in terms of physiological measures of workload compared to multiple levels of an auditory-vocal n-back task?

Consideration of Visual Metrics & Distraction Guidelines

As already described, both The Alliance (2006) guidelines for assessing driver interactions with advanced in-vehicle information and communications systems and the NHTSA (2013) visual-manual distraction guidelines specifically focus on the use of visual metrics during a complex radio tuning task as a reference point in evaluating an acceptable level of demand / distraction during driving. Therefore, it is clearly appropriate to apply various visual behavior measures to assess the visual demand of a visual-manual interface task such as manually tuning the radio. At the same time, it would seem to be a logical extension to apply these same measures to a voice-command interface that is specifically intended to allow the driver to carry out the same

task (radio tuning) while largely keeping their "eyes on the road". In other words, how else would one document if an interface option is doing a better job of keeping a driver's eyes on the road unless one measures their eye behavior when interacting with both interfaces?

It became readily apparent in working with the voice-command interface in the vehicle under test that much of the interaction with the interface was not purely auditory-vocal in nature. Beyond the initial requirement of engaging a press-to-talk button to initiate a voice-command interaction, many tasks such as full destination address entry into the navigation system involve the presentation of information on a display screen that the operator is expected to look at and then respond to with either a verbal or a manual touch screen response. This would seem to clearly put the interface in the realm being a multi-modal interface with visual demand. Does it not then seem reasonable to assess the visual demand components of a multi-modal interface just as one would with a purely visual-manual interface?

During the course of this project, NHTSA released its visual-manipulative driver distraction guidelines for in-vehicle electronic devices (NHTSA, 2013). As presented in some detail in that document, NHTSA largely accepted in principle the general approach developed by The Alliance for assessing visual demand and establishing acceptance criteria. However, NHTSA made several changes to The Alliance procedures and criteria. In particular, The Alliance approach to additional emphasis has been placed on the assessment of implications of the eyes-off-the-road metrics in comparison to more traditional approaches of categorizing glances to tasks, the characterization of "error free" performance in the evaluation of tasks, and specific demographic characteristics of study participants. To the extent possible, we have provided data analyzed along the methods proposed in the new NHTSA visual-manual guidelines, while clearly recognizing that the current study only provides an approximation of the specified assessment approach.

Methods

Participants

Recruitment drew from the greater Boston area using online and newspaper advertisements and consisted of two age groups, 20-29 and 60-69 years. These groups were selected to sample a cross-section of possible age related technology experience, mental models of how automobile interfaces "should" work, and age related cognitive and perceptual differences that might affect interaction with in-vehicle interface systems. Participants were required to read and sign an institutional review board approved informed consent form, to present a valid driver's license and attest to having had their license for more than three years, to driving on average three or more times per week, and be in self-reported reasonably good health for their age. A research assistant verified that participants clearly understood and spoke English. Individuals were excluded if they had been involved in a police reported accident in the past year, had a major medical illness resulting in hospitalization in the past 6 months, had a diagnosis of Parkinson's, Alzheimer's, dementia, mild cognitive impairment, or other neurological problem, were being treated for a psychological or psychiatric disorder, had a history of heart failure, angioplasty, coronary artery bypass grafting, a pacemaker, stroke, transient ischemic attack, or diabetes. Medication exclusions consisted of the use in the past twelve months of anti-convulsants, immunosuppressive, cytotoxic, anti-depressant, anti-psychotic, anti-anxiety drugs, or medications to treat a major medical condition such as cancer. Also considered was the use in the past two days of any medications causing drowsiness. A minimum score of 23 on the Montreal Cognitive Assessment (MoCA) was required as a cognitive status screen (Nasreddine et al., 2005) (See section on *Sample Statistics & Screening Results* under *Results* for a consideration of why this cutpoint was selected.) Potential participants were informed that the expected duration of the study was four to four and a half hours, including approximately two hours of on-road driving. Compensation was $90.

General Inclusion Criteria

- Age: 20-29 or 60-69
- A driver's license for more than 3 years
- Drive 3 or more times a week (on mean)
- Comfortable speaking and reading English

General Exclusion Criteria (based on self-report):

- A driver in a police reported accident in the past year

- Failure to positively endorse the statement "Would you be comfortable driving a full size sedan" as part of the study.
- Failure to positively endorse the statement "Are you in reasonably good health for your age?" or if self-rating of health on in-lab screening questionnaire as "poor".
- Any major illness resulting in hospitalization in the past 6 months
- Diagnosis of Parkinson's, Alzheimer's disease, dementia, mild cognitive impairment (MCI), or any other neurological problems?
- Current treatment for a psychological or psychiatric disorder
- Report of having ever had heart failure, angioplasty or coronary artery bypass grafting (CABG), a pacemaker, stroke or transient ischemic attack, diagnosis of diabetes
- Use in the past 12 months of anti-convulsant, immunosuppressive, cytotoxic, anti-depressant, anti-psychotic, anti-anxiety medications
- Medication to treat a major medical illness (such as cancer) in the past 12 months
- Use of medication that made them drowsy in the past 2 days

Other Exclusion Criteria:

- A Montreal Cognitive Assessment (MoCA) score less than 23 (see review of this selection in Results section)

Apparatus

An MIT owned 2010 Lincoln MKS with factory installed voice-command systems (Ford SYNC™ for voice control of the phone and media connected by USB and the "next-generation navigation system" with Sirius Travel Link) was selected as a convenient example of a widely available production level voice interface when this project was initiated in 2011.. The interface is engaged using a "push-to-talk" button on the right side of the steering wheel (see Figure 6). When the voice control interface is active, a display screen in the center stack typically supplies supporting information on system status and often provides information on prompts that the driver may use in dialog with the system (see Figure 7). A voice recognition training option is available in the system to optimize system capacity to recognize commands from an individual driver. This system training feature was utilized when a participant was introduced to the system to maximize the capacity of the system to correctly recognize commands from each participant.

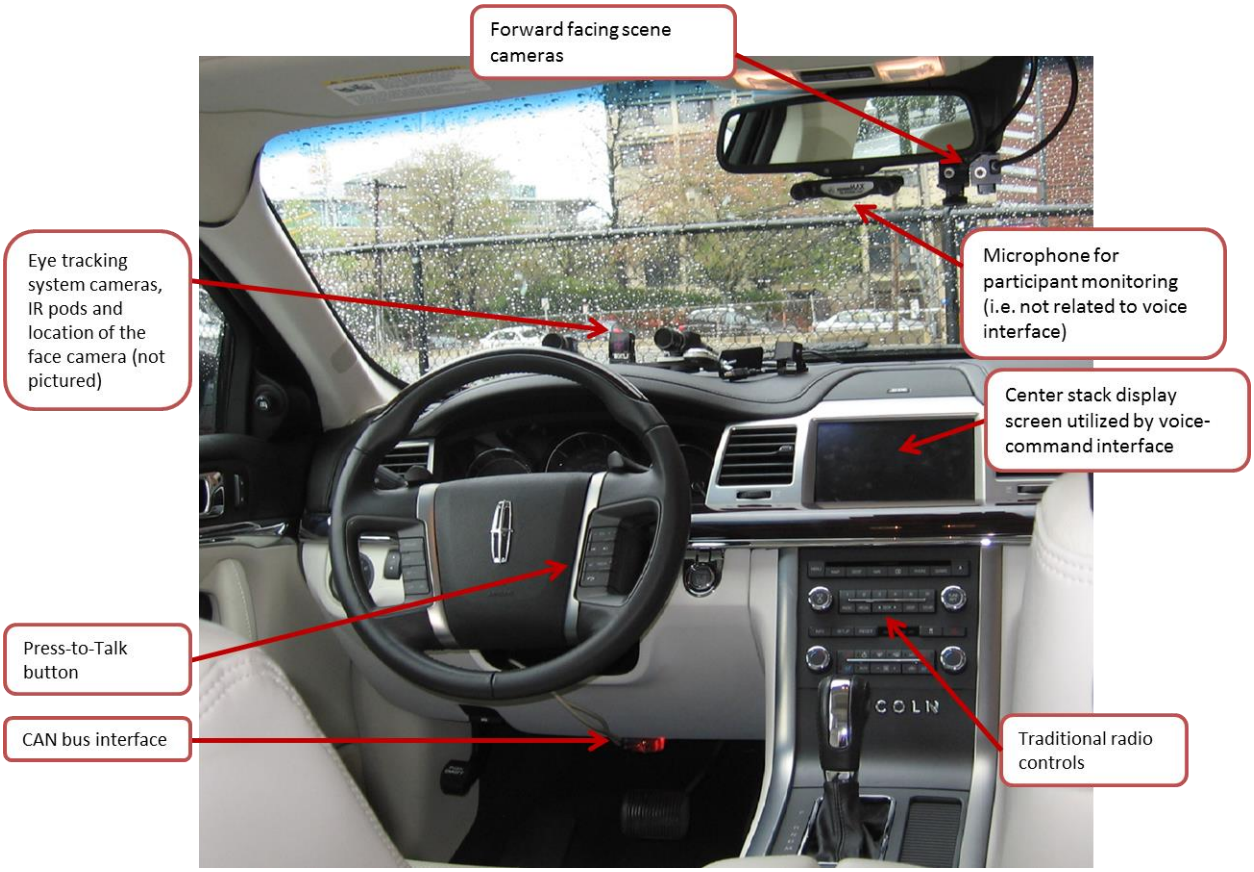


Figure 6: Interior of 2010 model year Lincoln MKS test vehicle. Note the Push-to-Talk button on the right side of the steering wheel that is used to initiate interaction with the voice-command system and the center stack display screen (see image below).

Figure 7: The screen above appears on the display screen at the top of the center console when the Push-to-Talk button is pressed.

As graphically visualized in Figure 8, the vehicle was instrumented with a customized data acquisition system for time synchronized recording of vehicle information from the controller area network (CAN) bus, a MEDAC System/3 physiology monitoring unit, FaceLAB® 5.0 eye tracking, cameras for capturing driver behavior and vehicle surroundings (see Table 1), a microphone, an Iteris AutoVue® Lane Departure Warning System for assessing lane position, and GPS tracking. CAN bus and lane position data were captured at 10Hz, GPS data at 1Hz, physiological data at 250Hz to support EKG feature extraction for accurate heart beat interval detection, and eye tracking data was recorded at up to 60Hz.

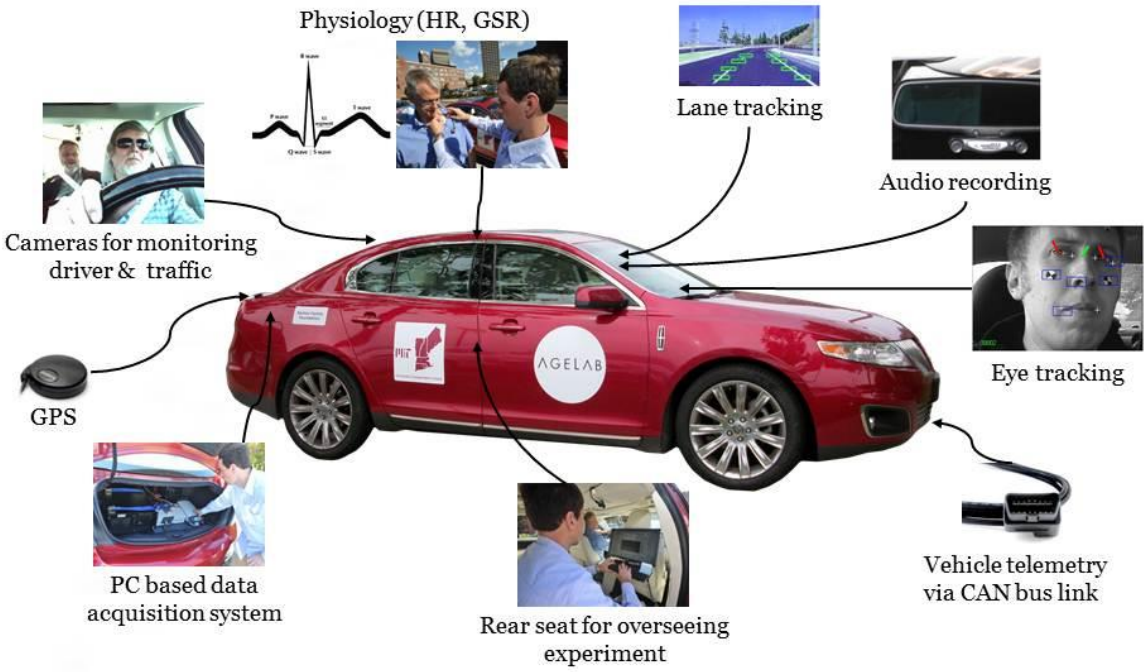


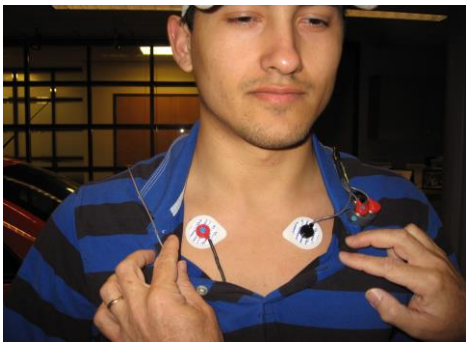
Figure 8: Experimental vehicle with key components noted. The identifying graphics shown on the side of the vehicle are removed during experimental sessions to avoid drawing attention to the vehicle and driver which might potentially impact normal traffic flow and interaction.

Table 1: Camera configurations and description.

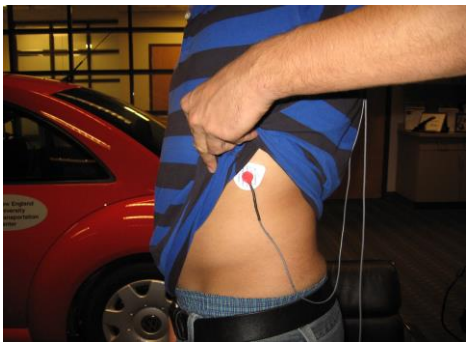
Camera Name / description	Frame Rate (fps)	Color	Image Size	Camera type / Lens
Forward View	30	N	640x480	Guppy Pro F125C/ Fujinon DF6HA-1B (6mm)
Forward View Wide Angle	15	Y	1024x240	Guppy Pro F125C/ Kowa LM4NCL (3.5mm)
Driver Face	15	Y	640x480	Guppy F033C/ Kowa LM6NCM (6mm)
Driver Bust	15	Y	640x480	Guppy F033C/ Kowa LM6NCM (6mm)
Over the Shoulder Dash	15	Y	640x480	Guppy Pro F125C/ Kowa LM4NCL (3.5mm)
Rear	15	Y	640x240	Guppy Pro F125C/ Kowa LM4NCL (3.5mm)

EKG recordings employed a modified lead II configuration; the negative lead was placed just under the right clavicle (collar bone), the ground lead just under the left clavical, and the positive lead on the left side over the lower rib. The skin was cleaned with isopropyl alcohol and standard pre-gelled silver/silver chloride disposable electrodes (Vermed A10005, 7% chloride wet gel) were applied. Skin conductance was measured utilizing a constant current configuration and non-polarizing, low impedance gold plated electrodes that allow electrodermal recording without the use of conductive gel. Sensors were placed on the underside of the outer segments of the middle fingers of the non-dominant hand and secured with medical grade paper tape. The thin surface design of the electrodermal sensors minimized interference with a natural grip of the steering wheel associated with the use of more traditional cup style electrodes. All wires were taped to participants for safety and positioned to allow for free movement (see Figure 9 for illustration of sensor placement).

EKG Sensors (3 contacts) – Employing a modified lead II configuration, active leads were placed just under the right collar bone and over the bottom rib on the left side of the body create a vector across the heart. The sensor just under the left collar bone is the ground / reference. The skin is cleaned with alcohol and wiped dry before placing sensor.

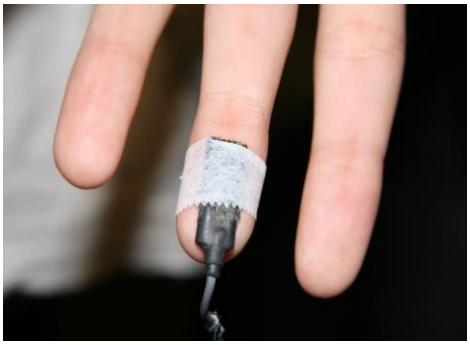


EKG with blue dot on right side; orient cable up over right shoulder and gather together with left lead on the left shoulder as shown below:

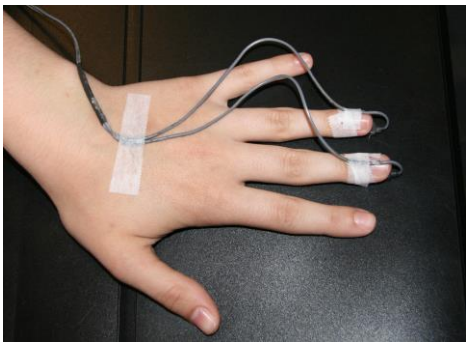


EKG placement over lower rib on the left side. Exact placement is not highly critical for this lead and it can be placed lower and somewhat farther back.

EDA Sensors for Skin Conductance (SCL) - Gold contacts were placed on the underside of the tip of the two middle fingers of the left hand. The inner edge of the gold contact is placed far enough forward so that the outer segment of the finger can bend normally around the steering wheel.



Lead wires are folded up and back over the top side of the fingers and held in place with medical paper tape.



To determine where paper tape should be placed on the back of the hand, the participant is asked to make a fist and draw their arm up toward their right shoulder as shown; the tape is then attached. Bending the fingers and elbow in this way corresponds to maximum pull that will occur on the lead wire.

Lead wire is taped at the 3 points shown and on top of shoulder.

Figure 9: EKG & electrodermal sensor attachment employed.

FaceLAB® calibration was performed, following the manufacture suggested procedures, as follows. Participants were instructed to sit in the driver's seat of the vehicle and look straight ahead. Two cameras mounted on the dashboard captured an image of the participant's face (this was reduced internally by the eye tracking system's internal algorithms to a representation of the participant's eye and facial features). From these data, FaceLAB® generated a model of the participant's face and eyes and tracks the changes in the positioning of these features in relation to a virtual "world model" environment that approximated the layout of the vehicle cabin (see "Automated Eye-Tracking", below). To verify that the eye tracker was properly calibrated, participants were asked to make a series of overt glances to objects of interest in the vehicle (i.e. the speedometer, rearview mirror, center stack touch screen, and finally, straight toward the front windshield). If FaceLAB®'s estimates of these glance targets were not accurate, the research assistant adjusted the positioning of world model objects was until the system produced an "observed" accurate estimate of the participants gaze positioning. The system was re-calibrated daily through the manufacture specified producers, "picture of a checker board calibration at various angles" to ensure that the system internal representation of the camera positions remained accurate.

Subjective workload ratings were obtained using a single global rating per task on a scale consisting of 21 equally spaced dots oriented horizontally along a 10cm line with the numbers 0 through 10 equally spaced below the dots and end points labeled "Low" and "High" on the left and right respectively (see Appendix I for details on the instrument). All of the scales were presented on a single sheet of 11x17 inch (legal size) paper so that participants were able to rate each task relative to tasks that they had already rated. Participants were told that workload may involve mental effort, the amount of attention required, physical effort, time pressure, distraction or frustration associated with trying to do the task while continuing to drive safely, and that workload is best assessed by the person doing the task. They were instructed to circle a point along each scale that best corresponds to how much workload they felt was involved in trying to do each task.

An experimenter was seated in the rear of the vehicle and was responsible for providing driving directions, ensuring safe vehicle operation, that participants understood and followed instructions, recording telemetry was working properly and that the experiment proceeded according to a predefined script. The data acquisition system supported playing recorded audio and the experimenter used a set of F-key presses at predefined points to trigger steps in the experiment. This ensured that primary instructions and tasks were presented in a consistent manner.

Secondary Tasks

There were six in-vehicle task areas: manual control of the radio, voice command control of the radio, navigation system destination entry, song selection (from an MP3 storage device), stored phone number dialing, and three difficulty / demand levels of an auditory presentation / verbal response calibration task (n-back). Illustrations of selected tasks can be reached through the following You Tube links:

- Manual Radio Tuning - http://www.youtube.com/watch?v=Kmd6oI2FWBc&feature=youtu.be
- Voice Radio Tuning - http://www.youtube.com/watch?v=0oiyV-S6KYs&feature=youtu.be
- Voice Navigation Entry - http://www.youtube.com/watch?v=X6gzg9k6T1U&feature=youtu.be
- Voice Song Selection - http://www.youtube.com/watch?v=OO5Qhkqt1OQ&feature=youtu.be
- Voice Contact Dialing - http://www.youtube.com/watch?v=zb862JV3u9U&feature=youtu.be
- N-Back Calibration Tasks - http://www.youtube.com/watch?v=08tbf7ak-wU&feature=youtu.be

Radio Tasks

Basic radio interaction was modeled on guidelines established by The Alliance (2006) and protocols developed as part of the Crash Avoidance Metrics Partnership (CAMP) Driver Workload Metrics project (Angell et al., 2006). An "easy" task consisted of changing a station by the single step of pressing a specified preset button in the radio-manual control version. The corresponding voice-command system interaction involved 3 steps (1 voice button press and 2 verbal inputs / confirmations, i.e. "preset-1", "yes"). Following the approach taken by CAMP, the "hard" version of the manual radio task consisted of turning the radio on, selecting a radio band, and then tuning to a specified station by rotating a manual tuning knob. It should be noted that the radio turning reference task adopted by The Alliance and recently endorsed by NHTSA, is slightly easier in that it assumes that the radio is already on. (The implications of this difference are considered in the Discussion.) A listing of the tasks and the steps required to complete each task appear below. Full procedural details including task specific training, introductory prompts, and any intermediate configurations steps are provided in greater detail in Appendix H. The voice system offered an advanced option for dropping confirmatory responses which would have reduced the number of steps in some of the interactions. Since this was not the system default mode, it was not used in the present assessment. In the 2010 Lincoln MKS studied here, the "harder" radio-manual task required 4 steps (pressing the volume control to turn the radio on, pressing a 'RADIO' button to access the band selection, pressing a touch screen band button (i.e. 'FM2'), and rotating the tuning knob to the specified frequency number). The corresponding voice-command interaction also involved 4 steps (1 voice button press and 3 verbal inputs / confirmations, i.e. "Radio", "100.7", "yes").

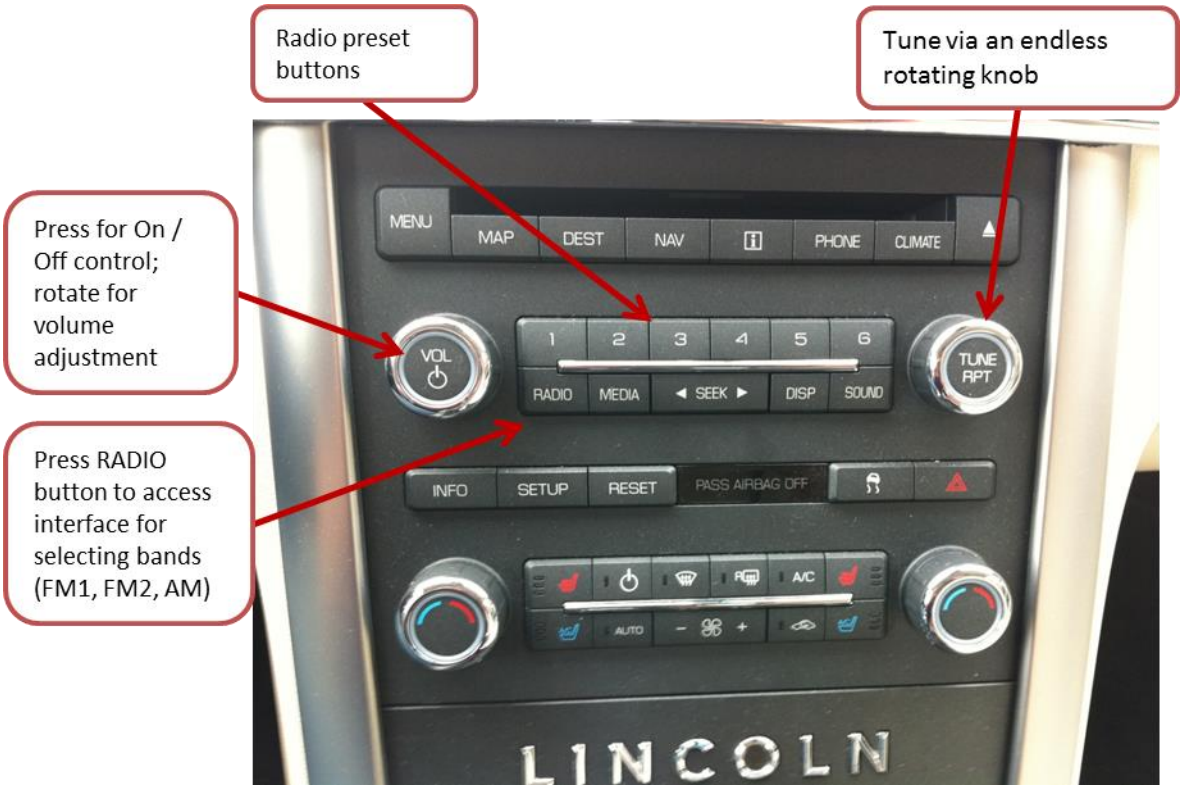


Figure 10: "Traditional" visual-manual radio interface in 2010 model year Lincoln MKS test vehicle. Switching between FM1, FM2, and AM radio bands was carried out using touch screen buttons on the main display screen located directly above the center console shown here.

Manual Radio Tasks

EASY 1: Your task is to change the radio to **preset-1**

EASY 2: Your task is to change the radio to **preset-5**

HARD 1: Your task is to **turn on** the radio, switch to **FM2**, and tune to **100.7**.

HARD 2: Your task is to **turn on** the radio, switch to **FM1**, and tune to **95.3**.

Voice Radio Tasks

EASY 1: Your task is to change the radio to **preset-1**

EASY 2: Your task is to change the radio to **preset-5**

HARD 1: Your task is to turn on the radio on using the **push-to-talk** button and request **FM 100.7**.

HARD 2: Your task is to turn on the radio on using the **push-to-talk** button and request **FM 95.3**.

Task Execution Notation:

[voice] a button press; i.e., press voice command button on steering wheel

(tune) rotate manual tuning knob

"yes" say a voice command

Radio Easy task: Radio is on, change radio station to Preset-X. For the manual task, the Preset buttons are classic style, numbered hard physical buttons as shown in Figure x.

Manual - 1 step **[Preset-1]**

Voice - 3 steps **[voice] → "Preset-1" → "yes"**

Radio Hard task: Radio is off, turn on, and tune to a specified frequency (station) when the radio is not already in the desired frequency band (i.e. AM/FM1/FM2). For the manual task, the [RADIO] button is a physical (hard) button located directly below the preset-1 button. The band selection is done by placing a finger on one of the touchscreen (soft) buttons that appear on the display when the [RADIO] button is pressed. Band options are: AM, FM1, FM2, Sat1, Sat2, Sat3. It is important to note that the use of soft buttons for band selection was chosen over an option that would have allowed users to toggle between bands by multiple presses of the radio button. This selection was made for two reasons. First, the soft buttons were highly salient one the [RADIO] button was pressed and likely to be the default option chosen by many participants even if instructed otherwise. Second, since no subscription to the satellite radio system was purchased, when toggling past those bands a message was presented on the touch screen noting the lack of subscription. This message needed to be closed by a press of the touch screen to move past each of the three Sat station bands.

Manual - 4 steps **[Vol] → [RADIO] → [FM2] → (tune)**

Voice - 4 steps **[voice] → "Radio" → "100.7" → "yes"**

Navigation System

Voice-command interaction with the navigation system consisted of two subtasks, entry of a street address and cancelation of the route request. Assuming there were no overt errors in interaction with the system, address entry required between 12 to 16 discrete steps. The number of steps appeared to vary depending on the confidence level of the system on the recognition of a voice entry and the extent to which a given street or town entry had variant options that the system required the user to select from.

Voice Navigation Tasks

HARD 1: Your task is to enter the destination address: **177 Massachusetts Avenue, Cambridge, Massachusetts**.

EASY 1: Your task is to cancel the route using the command 'Navigation Cancel Route'.

HARD 2: Your task is to enter the destination address: **293 Beacon Street, Boston, Massachusetts.**

EASY 2: Your task is to cancel the route using the command 'Navigation Cancel Route'.

Task Execution:

Address Entry - 12 to 16 steps – variable depending on speech recognition and whether a listing of selections to choose from was presented by the system:

[voice] → "Destination Street Address" → "yes" → "Cambridge" → "yes" → "Massachusetts Avenue" → "yes" → "One Seven Seven" → "yes"

[voice] → "Set as Destination" → "yes"

Cancel Route - 3 steps:

[voice] → "Navigation Cancel Route" →"yes"

Song Selection

For the song task, a USB drive containing MP3 files was pre-connected to the system. The primary task required 3 steps (1 button press, saying "USB" and then saying "Play Artist xxx'). Following this, participants were given a selection to request that did not exist on the device. This task ("song fail") was presented to observe how drivers interacted with the system when it was unable to comply with a request. Following this "failure" condition, participants were informed that, "The last task deliberately requested a song that did not exist on the storage device to simulate a condition where the voice system did not appear to recognize your request. This is the only time that this will be done intentionally during the study. Please continue driving."

Song Selection Tasks

EASY 1: Your task is to play music by The Rolling Stones.

EASY 2: Your task is to play music by Johnny Cash.

FAIL: Your task is to play Let It Be by The Beatles.

(Note: No music by The Beatles was present on the music device.)

Task Execution:

Song Easy task – 3 steps

[voice] → "USB" → "Play Artist The Rolling Stones"

Song Fail task – no completion possible

[voice] → "USB" → "---"

Voice Initiation of a Phone Call

The final system task involved placing a phone call to a stored number using the voice interface. An introductory prompt reminded participants of the steps involved in placing a call to minimize the extent to which actual memory of the task from the training period was a significant variable in the assessment. Placing a call to a stored number required 3 steps (1 button, saying "phone" and then "call contact x"). Terminating the call required "pressing" the End button on the center console touch screen. Complete protocol details are provided in Appendix H.

Voice Phone Tasks

Task 1: Please place a phone call now to **contact-3**.

Task 2: Please place a phone call now to **contact-4**.

Task Execution:

Placing Call – 5 steps: **[voice] → "phone" → "yes" → "call contact 3" → "yes"**

Terminating Call – 1 step: **[End]** (touch screen)

N-Back Surrogate Task

In 2006, the MIT AgeLab began a project to assess the feasibility of biometric-based state detection under driving conditions that considered a wide range of possible physiological measures. To carry out this evaluation, a method of reliably inducing multiple levels of arousal or demand was required. A number of different methods were considered and eventually a variation of a cognitive task widely used in neuropsychological and medical research was adapted for use in the project. The resulting n-back task variation developed in the AgeLab requires participants to hold single digit numbers in memory and to repeat them back verbally either immediately (0-back), after another number has been presented (1-back), or after two additional numbers have been presented (2-back). As shown in the example below (Table 2), the numbers are presented as a random ordering of the digits 0-9 with a typical spacing of 2.25 seconds between numbers. Single 10 item stimulus sets were employed in this study, resulting in task periods of approximately 30 seconds in duration.

Table 2. Example of an N-back task set.

Stimulus	6 9 1 7 0 8 4 3 5 2
0-back Response	6 9 1 7 0 8 4 3 5 2
1-back Response	. 6 9 1 7 0 8 4 3 5
2-back Response	. . 6 9 1 7 0 8 4 3

As can be seen from the table above, for the 0-back task the participant simply has to repeat each number as it is presented. In the 1-back task, the participant is to hold a number in memory, wait for the next number to be presented and then enter it into memory, and then verbalize the previous number while continuing to hold the most recent number in memory. The 2-back extends upon the 1-back by requiring that the participant hold the most recent two numbers in memory. The vocal demands of this task are relatively consistent, with the 1-back requiring one less vocalization than the 0-back and the 2-back requiring one less vocalization than the 1-back. Consequently, the task largely represents a manipulation of the level of demand on working memory.

Extensive research has been undertaken on the use of a delayed digit recall task (n-back) as a method for inducing graded levels of cognitive demand during simulation and actual on-road driving (Mehler, et al., 2012; Mehler, et al., 2009; Reimer, 2009; Reimer & Mehler, 2011; Reimer, Mehler, Wang, & Coughlin, 2012; Son et al., 2011).

Procedure

Detailed task protocols with step-by-step instructions used during the experiment (laboratory, training and driving evaluation) appear in Appendix H. The protocol checklists provide instructions for the research assistant, language guidance for all key interactions with the participant, the specific language that was pre-recorded for scripted interactions (i.e., auditory prompted instructions), conceptual steps that were expected for error free interactions with each of the tasks and areas for recording specific (yes / no / scores), and general open ended notes on the progress of the experiment.

Outline of Intake and Initial Training Phase of Study

In-Lab Start Phase

- When Participant Arrives - Consent Forms / Payment Form / Emergency Contact Form
- Review of Eligibility (Interview & MoCA)
- Pre-Experimental Questionnaire
- N-Back Training
- Workload Scale Rating Explanation

Bathroom Break

- Physiological Sensor Attachment

Move to Vehicle

- Set Participant Up in Vehicle / Eye Tracking Calibration / N-Back Practice
- Voice system calibration to participant
- Training in MIT parking lot on first set of in-vehicle tasks

Following informed consent, a review of eligibility criteria, cognitive screening on the MoCA, and completion of a pre-experimental questionnaire, participants were trained to minimal competency criteria on the n-back task as in Mehler, Reimer and Coughlin (2012) and then given an explanation of how to complete the workload rating scale. A bathroom break was offered, physiological sensors attached, and participants were then escorted to and given an orientation to the research vehicle. The participant was instructed to adjust seat and mirrors and asked to back up the vehicle a few feet before picture were taken for calibration of the eye tracking system. Additional 10-item sets of each of the levels of the n-back task were practiced in the stationary vehicle as the RA configured the eye tracker (see procedures above). An introduction to the voice command system was provided that included going through the individual voice calibration option.

As detailed previously, there were six in-vehicle task areas: manual control of the radio, voice command control of the radio, navigation system destination entry, song selection (from an MP3 storage device), stored phone number dialing, and three difficulty / demand levels of an auditory presentation / verbal response calibration task (n-back). Each task type was presented twice. For purposes of the study, the radio tasks, voice-based navigation entry, and song selection were approached as the primary system tasks for evaluation and their order of presentation was counterbalanced across the sample taking into account age and gender. The phone task included some exploratory components beyond the basic task of using the voice system to place a call to saved contact and, as a consequence, was always presented last in the task sequence. The surrogate auditory presentation / verbal response n-back task was intended specifically as a reference task for calibration / comparison scaling against the primary system tasks. With this in mind, the two presentations of the n-back task were interspaced between the primary system tasks. A general outline of task workflow and counterbalancing is shown in Figure 11.

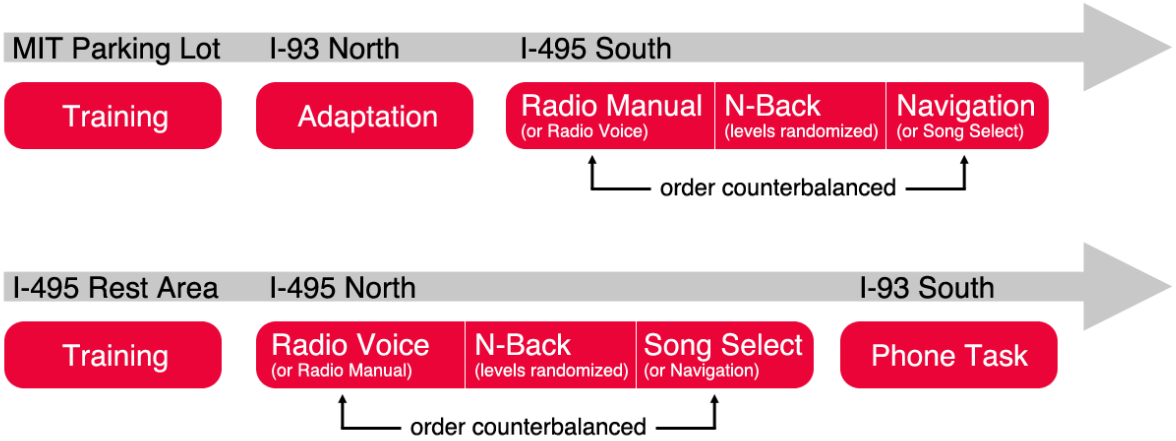


Figure 11: A flowchart illustrating the ordering of training and task periods. Task ordering was counterbalanced across participants as shown, resulting in eight possible task configurations. N-back tasks were always performed as the middle task in a block.

On-Road Assessment

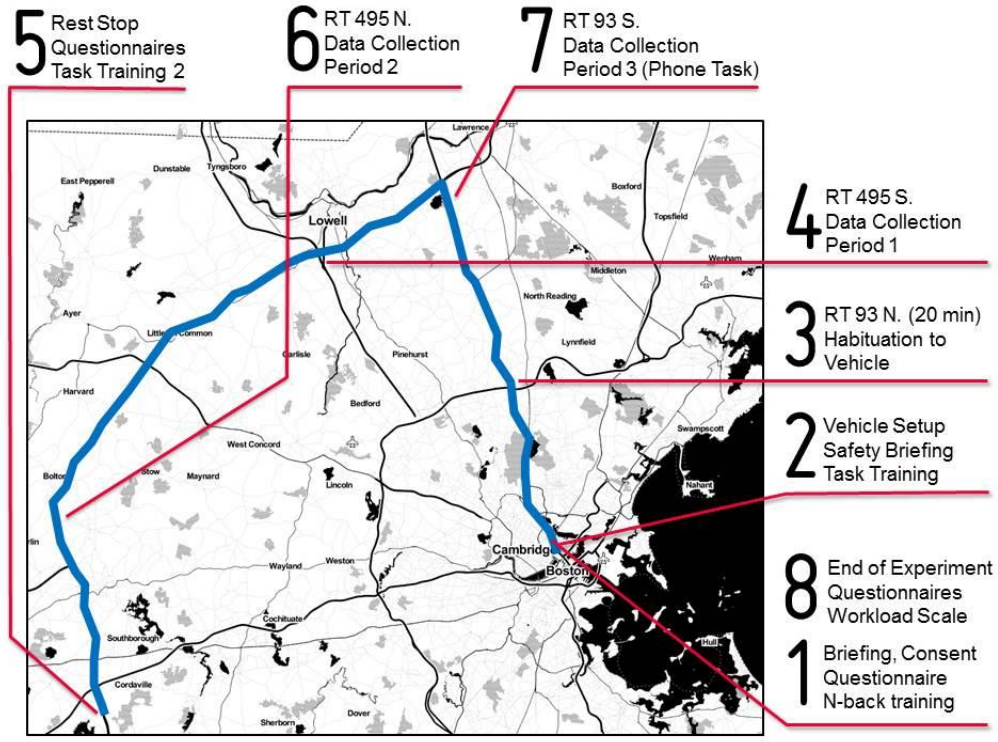


Figure 12: Experimental route with key protocol periods.

The driving portion of the study was conducted on roadways in the greater Boston area and divided into four segments (see Figure 12). The first segment (Figure 12-3) consisted of a period of approximately 10 minutes of urban driving to reach interstate highway I-93 and continued north on I-93 for an additional 20 minutes or so to the I-495 intersection. This allowed a total adaptation period of approximately 30 minutes of driving prior to the assessment portion of the study. The second segment (Figure 12-4) consisted of driving south on I-495 to the exit 19 rest area and averaging approximately 40 minutes. The third (Figure 12-6) was from the rest area back north on I-495 to I-93 and the fourth (Figure 12-7) was the return on I-93 south. The radio-manual, radio-voice, navigation-voice, and song selection-voice tasks were presented in a counter-balanced order during segments two and three with the exception that the radio-manual and radio-voice tasks were never presented in the same segment. The 3 levels of the n-back were presented twice, once each in the middle of segments two and three; ordering of the levels was randomized. The phone task was always presented during segment four. Detailed training was provided in the MIT parking lot (Figure 12-2) on the tasks to be completed during the first half of the drive. Training and practice on the remaining tasks were provided during the rest-stop between segments two and three (Figure 12-5). Self-report workload ratings were obtained at the rest stop and following the completion of the drive for the tasks completed during the first and second halves of the drive, respectively.

Measurements

Measures of driving speed, steering wheel position, and acceleration data were recorded directly from the vehicle CAN bus.

Steering Wheel Metrics

Steering wheel reversals were classified as proposed in the final report of the European Union AIDE project (deliverable D2.2.5, section 7.12) (Östlund et al., 2005) . This metric captures the number of steering wheel inputs exceeding an angular reversal gap of either 3° for major or 0.1° for minor reversal events. The rate of steering wheel reversals per minute was obtained by dividing the raw reversal rate by the task trial duration.

Additionally, the standard deviation of steering wheel angles, reported in angular degrees, was calculated based upon raw steering wheel angle information.

Lane Departures

The frequency of lane departure events was measured for the sample based on the Iteris AutoVue 3G lane departure warning system.

Mean and Standard Deviation of Velocity

Two additional driving performance metrics included in this report are the mean and standard deviation of forward vehicle velocity, both measured in m/s. Input data for these metrics was obtained from the vehicle CAN bus.

Acceleration Events

CAN bus data of longitudinal and lateral acceleration was used to calculate independent acceleration events, as proposed in Reimer et al. (2012). This measure examines unidirectional acceleration, computed from individual lateral and longitudinal measures using the Pythagorean Theorem. Classification of independent acceleration events is parameterized with thresholds for both temporal separation and acceleration magnitude. For this report, an acceleration threshold of 0.1g (0.98m/s^2) and a temporal separation of 2 seconds between independent events were applied.

The count of acceleration events was normalized by each participant's trial duration, yielding the acceleration event rate, expressed in units per minute.

Physiological Metrics

Heart beats were detected through identification of R-wave peaks in the EKG signal. Processed records were reviewed by trained RAs to identify and resolve any detection issues. High frequency noise in the skin conductance level (SCL) signal was removed through a wavelet transform (see Reimer & Mehler, 2011). Gross low frequency movement artifact was identified by manual inspection and removed.

Automated Eye-tracking

FaceLAB® is built on the concept of a "world model", a collection of virtual objects (planes and spheres) that approximates the layout of the instrumented vehicle cabin (see **Figure X**). This model allows FaceLAB® to automatically estimate and label the target of the participant's gaze at any point in time. When a participant's gaze vector intersects with one of the world model regions, the system records this region as the *gaze object* for that frame. After consulting with the FaceLAB® support team, the world model was configured with four planes to represent the

driver's broad surroundings: Front (dark blue plane in **Figure X**), Left, Right, and Bottom. In addition, we defined three objects that were relevant to the present investigation: rearview mirror, instrument cluster (located on the center of the steering wheel where the prompts for voice instructions were placed), and the entertainment cluster (or "center stack", colored in dark green in **Figure 13**) at the far right side of the console, which represented the Ford SYNC touch screen. As noted in the procedures above these regions were moved in the world space during system setup to "best" align with each participants actual glances to a region.

However, subsequent analysis determined that when a gaze vector intersected with more than one world model object, i.e. where the objects overlap with each other relative to the gaze angle, only one of the gaze objects is recorded (the object that was created first, according to FaceLAB®'s internal index). In the present study, any gaze that intersected with the instrument cluster object also intersected with the Front object, and therefore, was always recorded as a gaze to the Front. It is also likely that many glances to the entertainment cluster were incorrectly recorded as Front glances. This issue in combination with data loss due to characteristics of the field driving environment (high degree participant movements, lighting changes etc.) limited the correspondence between automated eye-tracking analysis methods and manual validation of the drivers gaze patterns using recordings from the face camera. As a result, we resorted to coding participant glance behavior manually from in-vehicle video recordings (see next section).

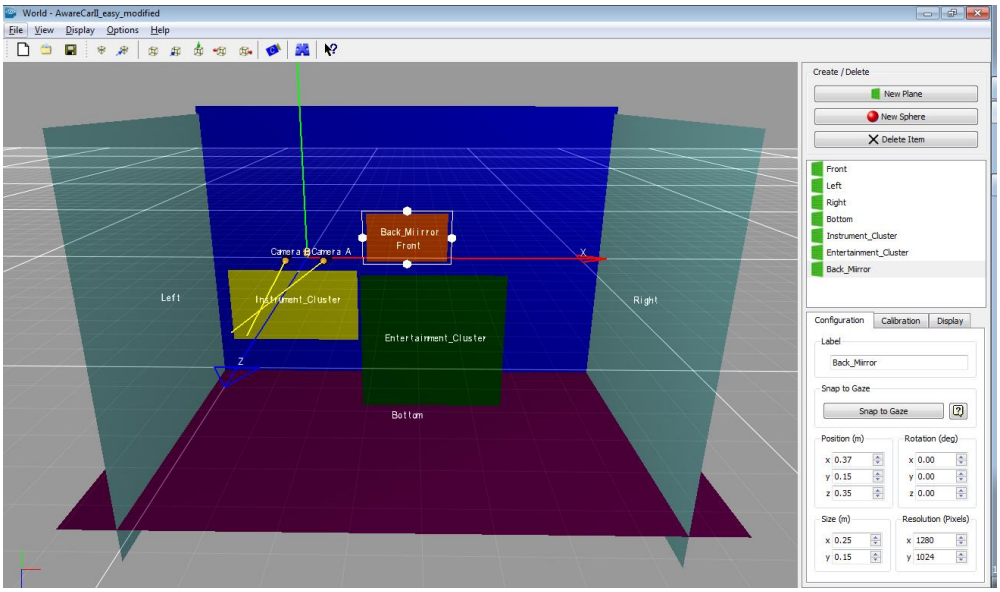


Figure 13: An example of the FaceLAB® world model used during data collection.

Glance Coding

In-vehicle video for all participants and task periods of interest were coded to summarize participant glance behavior. Eye glance behaviors were coded manually with the assistance of software specifically designed for this purpose. Two independent coders manually assessed video of each task. A third party resolved any discrepant glance codes. Detailed procedures for glance coding and mediation are given in Appendix F.

Glance Metrics

Specific glance metrics are discussed in the introduction to the Glance Analysis portion of the Results section under the heading, "Glance Measures & Off-Road Glance Metrics"..

Orienting Response

A single coder manually assessed videos for indications of an orienting response, a behavior in which the participant appears to engage directly with the in-vehicle display as if were the location of the voice-command interface. For example, the participant might begin speaking toward the display's location, lean towards it, change his posture, turn his body, or otherwise behave in a manner that suggests he has begun to prioritize interaction with the in-vehicle display. The detailed coding guide for orienting response behavior is given in Appendix D.

Task Completion

Based on audio recordings of the driving sessions (and protocol notes in the case of the Radio Manual tasks), two coders independently assessed the participant's ability to successfully complete each secondary task. Performance on each task was rated as to whether participants required minimal or substantial assistance, needed to backtrack to correct an error, and whether they ultimately completed the task correctly. An independent mediator who did not perform the initial coding resolved any discrepancies in rating. See Appendix G for details.

Data Reduction & Analysis

Baseline reference values were computed for selected metrics as average values obtained across seven-two minute long single task driving periods. Each of these periods was drawn immediately prior to the seven different task periods. In the case of eye movement and driving performance data, these baseline periods are presented along with the metrics describing performance during tasks. In the case of physiological measures, task periods are presented as change scores from the baseline as well as in absolute heart rate and skin conductance values. (Appendix E breaks out the individual baseline periods to look at the consistency across periods

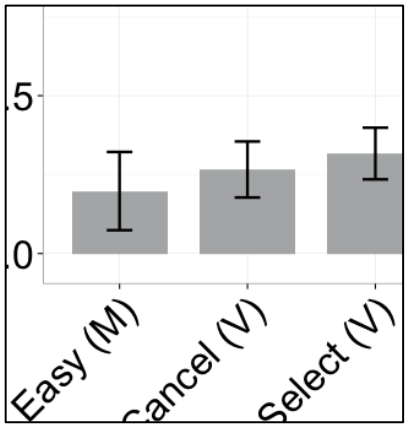
and the extent to which the use of individual vs. aggregated baselines impact the data.) Each task type was presented twice during the drive. For the Primary Analysis, the two presentation periods were averaged for analysis purposes (see Appendices A-C for alternative analyses).

Statistical Analysis

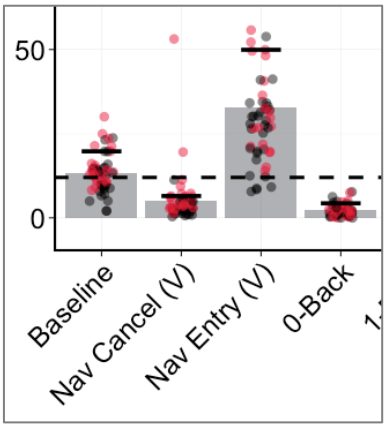
Statistical tests for main effects of task period, age, and gender are presented for each dependent variable. Given the large number of tasks that could be compared on an individual basis, the total number of statistical comparisons across the study was minimized to avoid alpha error inflation. In line with the formal research questions outlined in the Introduction, two task specific comparisons are presented for each variable. The manual Radio Hard task and the voice Radio Hard task are compared to assess whether the voice-command method of carrying out the task results in equivalent or lower impact on the variable. Second, the use of the voice-command system to enter addresses into the navigations system (Nav Entry) is compared on each variable against the manual Radio Hard task.

Statistical analyses were performed in R (R Core Team, 2013). Owing to non-normality of the sample data and /or the use of ratio data (percentages) for several dependent measures, in most cases non-parametric statistics such as the Wilcoxon signed rank test and the Friedman test were used (similar to the t-test and repeated-measures ANOVA, respectively). For selected analyses, repeated-measures ANOVAs are presented to maintain consistency with earlier reporting.

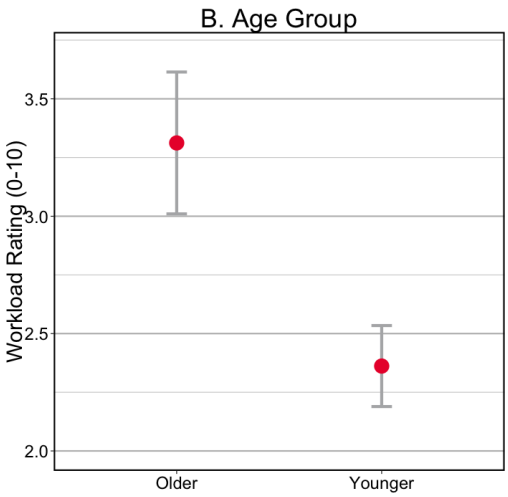
Data Visualization (plotting)



This report makes extensive use of three types of plots. For each measure of interest in the Primary Analysis section, figures are presented in which tasks are ordered by their mean values, with their variability indicated with error bars representing the mean adjusted standard error of the mean (SEM) (Loftus & Masson, 1994). Additionally, the n-back tasks are highlighted in these figures as a kind of "cognitive ruler" by which to compare tasks. This allows for gross comparisons to be made between task types at a glance, without presenting a large amount of per-subject data.



The second figure style, which we call a "participant performance plot", displays a large number of variables pertaining to a dependent measure. For each task (shown on the x-axis), the gray bar represents the sample's mean performance. Individual participants are plotted as points. Red points correspond to older participants, while black points correspond to younger participants. Dot positions are jittered horizontally to minimize visual overlap. The dashed horizontal line represents NHTSA criterion values (where applicable). The horizontal line segments aligned with each bar represent the 85th percentile of performance for that task. If the line segment is above the dashed criterion line (as in "Nav Entry", shown here), the sample has failed to meet NHTSA's criteria for that task. If the line segment is below the criterion line (as in "Nav Cancel"), the sample has met or exceeded the criteria. Lastly, the number of participants available for each task sample is displayed along the top of the plot (not pictured). These graphics present individual participant performance, summary statistics, and their relationship to government mandated pass/fail criteria in one succinct picture.



The third plot type, called a "statistical plot", displays statistical relationships for a given factor and measure. For example, the plot at the right shows the mean difference between workload ratings for older and younger participants. Red points represent the mean for each group, and the gray bars represent ±1 mean-adjusted standard error (SEM).

Results (Primary Analyses) with Commentary

This section presents the primary data analysis for the study. Glance metrics are considered in this section using the "eyes off-the-forward-roadway" criteria advocated by NHTSA. An alternative analysis that considers glances to the in-vehicle device, rather than all glances off-road, is presented in Appendix A. Other alternate analyses that break down the sample by participant performance ("error-free" cases) and task trial (first vs. second) are presented in Appendices B and C.

Sample Statistics & Screening Results

Figure 14: Graphic representation of the progression from participant recruitment through inclusion of cases in the final analysis dataset.

As illustrated in Figure 14, (102) participants were recruited for participation in the study to obtain the final target sample, equally distributed across age group and gender, of 60 cases. A number of individuals did not proceed to the driving portion of the study due to failure to pass

the cognitive screen test (MoCA), personal scheduling conflicts that precluding having enough time available to complete the study protocol, etc.

In the context of assessing how drivers interact with a production level voice interface, it seems appropriate to highlight that only two participants were excluded from the driving portion due to clear difficulties with the system in recognizing their voice commands in the parking lot. This was observed in spite of a wide range of speech patterns and was markedly less than what we initially anticipated. We have no measure of the extent to which taking participants through the voice calibration procedure influenced this; however, having taken all participants through the procedure, the overall recognition success of the system under static conditions was quite good.

Of those who proceeded on to driving portion, the primary reason for exclusion from the analysis sample was unavailability of good EKG recordings for purposes of obtaining heart rate data (13 cases). Availability of heart rate data was originally set as an inclusion requirement due to interest in using this metric as an objective workload measure. An alternate listing of reasons for exclusion of cases presented in order of frequency is provided in Appendix L. Descriptive statistics for the final analysis sample broken-down by age and gender grouping are presented in Table 3.

Table 3: Demographic statistics. Each cell represents the mean (SD) [range] for 15 participants.

	Female	**Male**	**Total**
Younger	24.73 (3.0) [20.0 - 29.0]	24.00 (2.7) [20.0 - 29.0]	30
Older	64.13 (3.0) [60.0 - 68.0]	66.20 (2.9) [60.0 - 69.0]	30
Total	30	30	60

Table 4: Highest level of education completed, by age group (30 participants in each group).

	Younger	**Older**
High School Graduate	0	2
Some College	10	5
College Graduate	12	5
Some Graduate Education	2	5
Completed Graduate Degree	6	13

Generally speaking, the sample was highly educated (see Table 4). All of the younger participants had enrolled in college, completed college, or completed a graduate degree. Along the same lines, the older sample was highly educated, with 18 out of 30 older participants completing at least some graduate work.

Additional demographic data broken down by age and gender is provided in Appendix K, which presents descriptive statistics (mean, standard deviation, min. and max. values) for all of the pre- and post-experimental questionnaire data collected. Selected items are presented here.

Participants were asked to rate how safely they drove on a scale from 1 to 10 (1 being very unsafe, 10 being very safe). All participants rated themselves fairly highly on this scale (see Table 5), with older drivers rating themselves more highly than younger drivers, and men rating themselves slightly more highly than women (none of these differences were statistically significant).

Table 5: Self-reported driver safety. Each cell represents the mean (SD) [range] for 15 participants.

	Female	**Male**
Younger	8.47 (1.2) [6.0 - 10.0]	8.73 (1.0) [7.0 - 10.0]
Older	8.93 (0.8) [8.0 - 10.0]	9.13 (0.9) [7.0 - 10.0]

Participants were also asked a range of questions concerning their familiarity and comfort level with new technologies (see Table 6). Younger participants rated themselves more highly on these measures (on a scale from 1 to 10), indicating a greater level of experience with technologies such as cell phones, automatic teller machines, digital cameras, computers, etc. (question #11), more willingness to try and adopt new technologies (question #12), and a greater ability to learn new technologies (question #16). None of these self-reported measures differed significantly between age groups or genders, aside from a borderline significant effect of age group for trust in technology ($p = .044$).

Table 6: Self-reported technology experience and engagement comfort (1-10 scales).

Age Group	Gender	Technological Experience (#11)	Willingness to Try New Technologies (#12)	Ability to Learn New Technologies (#16)
Younger	Female	8.07	7.2	8.33
	Male	9.13	8.13	8.67
Older	Female	7.87	6.93	8.2
	Male	7.87	6.93	8.8

Cognitive Screening

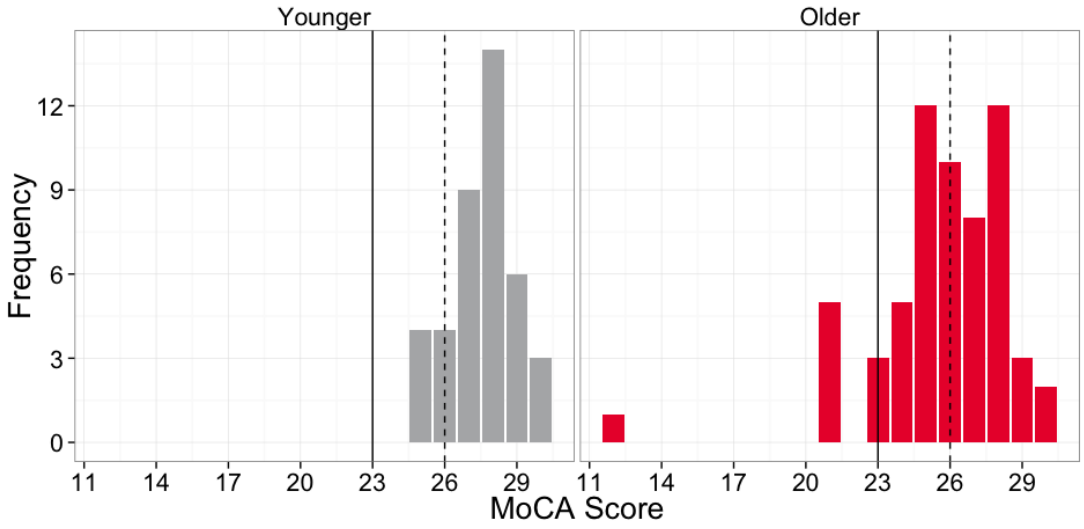


Figure 15: Distribution of MoCA scores by age group for 101 individuals screened for participation in the study. The relative impact of using cutpoints of 23 and 26 on the different age groups is illustrated using the solid and the dashed vertical lines, respectively.

Though not a focus of the current study, an observation concerning our experience with the use of the Montreal Cognitive Assessment (MoCA) (see Nasreddine, et al., 2005) as a brief screening tool may be informative for other groups involved in driving related research. We had for many years used the Mini-Mental State (MMSE) (see Folstein, Folstein, & McHugh, 1975) as an established method for identifying individuals with possible cognitive impairment that we would prefer not to include in driving studies for safety and other considerations. In our previous experience, use of the MMSE occasionally identified one or two individuals per participant group with clear issues of confusion. However, it was observed that a number of

research participants found the MMSE objectionable. The MoCA was suggested as a possible alternative screening device and a review of the available literature indicated a number of potentially attractive psychometric characteristics of the scale. At the same time, an open question existed for the MoCA regarding what would an appropriate cutpoint for purposes of driving research screening. We are not aware of research that has yet established a validated MoCA screening value specific for driving. In their 2005 paper, Nasreddine et al. proposed that a cutoff score of 26 (out of 30) was sensitive to identifying individuals with mild cognitive impairment who would fall within the normal range on the MMSE. In a study of older drivers completed prior to the initiation of the current study, we administered the MoCA and explored the use of a score of 26 as a cutpoint. This resulted in a significant number of potential participants being excluded from the study; the impression of the research staff was that the majority of these individuals did not otherwise show any outward signs that would make the staff members uncomfortable driving with them. A review of the distribution of those scores suggested that lowering the cutpoint so that individuals with scores less than 23 were excluded might be more reasonable.

A total of 101 individuals completed the MoCA. The distribution of the resulting scores is shown broken out by age group in Figure 15 on the previous page. The vertical dashed line indicates a cutpoint value of 26 and the solid line a cutpoint value of 23. Using 23 as a screening value resulted in allowing all of the younger participants to proceed to the initial MIT parking lot training portion of the protocol and excluded 6 older participants. If the cutpoint of 26 had been used, then a total of 30 out of the 101 individuals (29.7%) would have been excluded. One older participant (65 year old male) with a MoCA score of 24 was removed from the study during the parking lot training period due to experimenter concerns around the participant's capacity to engage in the tasks while driving. The other two participants who were withdrawn based on experimenter concerns during the parking lot training period both had MoCA scores of 26 (68 year old females).

Self-Reported Workload

After the completion of a block of tasks, participants rated how much workload they felt was involved in trying to do each task. A 0 (low) to 10 (high) scale that allowed for half point resolution was employed. See Appendix I for background on the scaling method used and instructions participants were given on how to conceptualize and rate workload. A reproduction of the workload rating sheet is also provided.

Figure 16 presents the self-reported workload ratings for each task in rank order. A repeated measures ANOVA indicates that there was an overall main effect of task type on perceived workload ($F(11, 506) = 32.8$, $p < .001$). Mean ratings ranged from just under one for the 0-back task to just over six and a half for the Song Fail task. (Means and standard deviation values for the sample as a whole and broken down by age are detailed in Table x.)

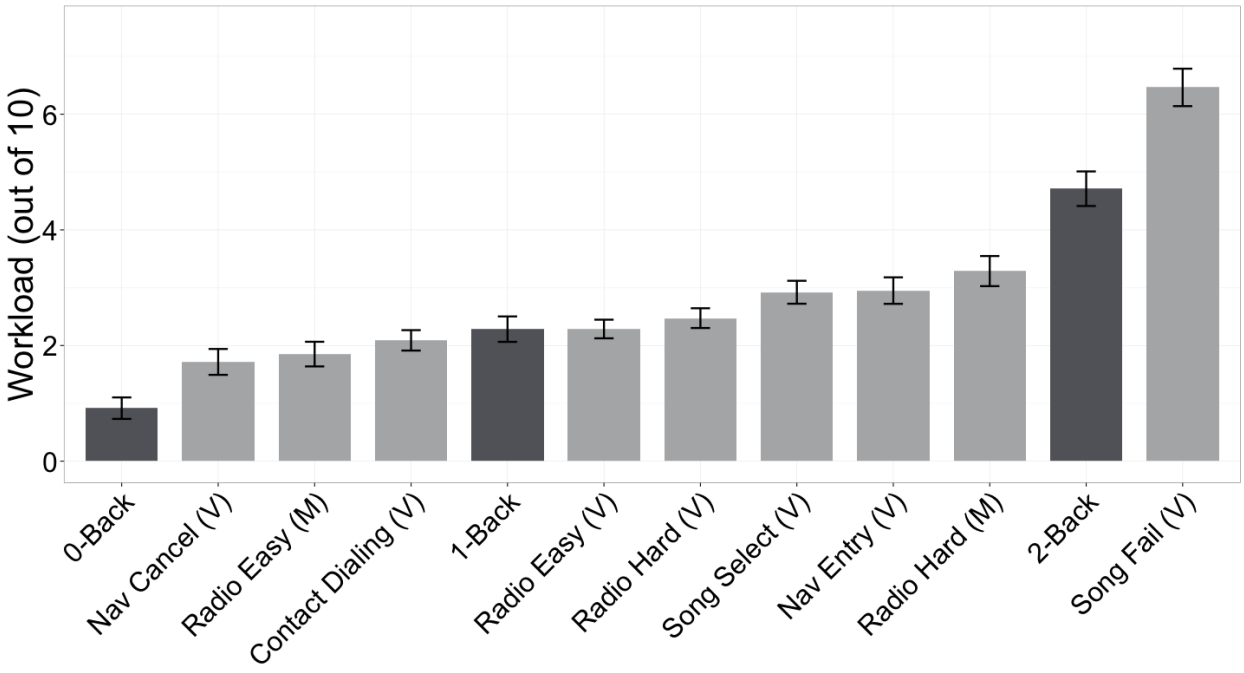


Figure 16: Tasks listed in ascending order for mean reported workload level. N-back reference tasks are denoted with darker bars. Error bars represent 1 SEM. Tasks marked (V) used the voice interface. Tasks marked (M) utilized traditional manual/tactile interactions. (Figure adapted from (Reimer, et al., 2013).)

Inspection of Figure 16 suggests that the three levels of the n-back calibration / reference task bracket the in-vehicle tasks in a conceptually consistent manner. The low demand 0-back task received the lowest workload rating, the medium demand 1-back task received an intermediate rating, and the high demand 2-back task received a workload rating that was only exceeded by

the Song Fail task that was functionally impossible for participants to complete. It is of both practical and conceptual interest to observe where the various tasks fall relative to the Manual Radio Hard turning task, which matches (with the exception of having to turn on the radio) what the original Alliance guidelines (2006) recommend and the more recent NHTSA (2013) guidelines specify as a reference task for the upper bound of a "generally accepted" level of secondary demand on a driver. Participants rated all of the other standard in-vehicle tasks (with the exception of the impossible Song Fail task) as involving less workload than this relatively complex visual-manual task. In line with the overall design goals for the voice interface, the voice-based method of completing the Radio Hard task received a lower subjective workload rating than the manual method. In contrast, for a simpler task, selecting a pre-set radio station, the single manual step of pressing a preset button was rated less demanding than using the voice interface which required a button press and two voice commands. It can be noted that the navigation Destination Entry, the most complex task in terms of number of steps, was rated nominally lower than the manual Radio Hard tuning task. (Statistical tests for the latter two comparisons are presented at the end of this section.)

Table 7: Means (and standard deviations) of self-report global workload ratings.

Task	Younger	Older	(All)
Nav Cancel	0.87 (0.9)	2.55 (2.7)	1.69 (2.2)
Nav Entry	2.53 (2.1)	3.38 (2.6)	2.95 (2.4)
0-Back	0.33 (0.5)	1.55 (2.6)	0.93 (1.9)
1-Back	1.83 (1.4)	2.79 (2.5)	2.31 (2.0)
2-Back	4.67 (2.5)	4.86 (2.9)	4.76 (2.7)
Phone	1.03 (0.9)	3.31 (2.4)	2.11 (2.1)
Radio Manual Easy	1.87 (1.8)	1.91 (1.9)	1.89 (1.8)
Radio Manual Hard	3.67 (2.1)	2.96 (2.6)	3.31 (2.3)
Radio Voice Easy	1.43 (1.5)	3.19 (2.3)	2.29 (2.1)
Radio Voice Hard	1.45 (1.5)	3.59 (2.4)	2.48 (2.2)
Song Fail	6.48 (2.4)	6.65 (2.8)	6.56 (2.6)
Song Select	2.16 (1.8)	3.81 (2.8)	2.96 (2.4)

Table 7 provides mean and standard deviation values for each task for the overall sample and broken down by age group.

As detailed at the end of this section, the older group gave all tasks, with the exception of the manual Radio Hard tuning task, higher workload ratings than the younger group.

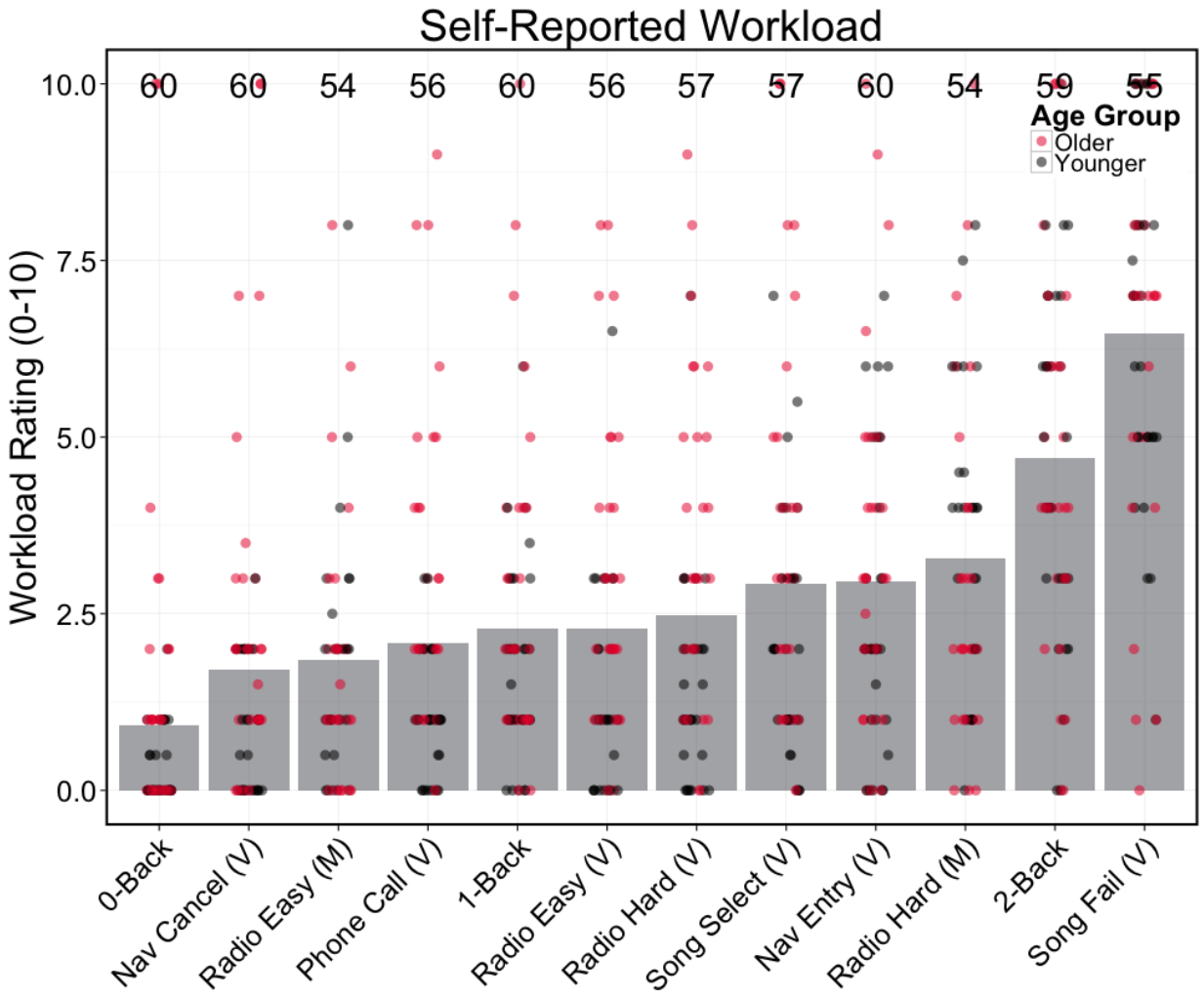


Figure 17. Response distributions for the self-reported workload across each of the tasks. See the section on Data Visualization for a description of the graphical format used throughout this report. Gray circles represent individual scores of Younger participants and red circles are scores of individual Older participants. The gray bars represent the full sample mean. Note the high variability in self-reported workload both within and between tasks.

Inspection of Figure 17, which shows the distribution of individual self-reported workload ratings, makes clear the generally wide variability in how individuals chose to rate the workload of various tasks. All of the tasks were given a 0 (low end of the scale) at least once. (The one participant to give every task a 0 rating was a 67 year old female.) However, most participants did rate tasks along a range of values. It can be noted that the mean values and the

variance in the ratings was higher for the older sample across the tasks. The higher workload rating across tasks by the older participants is statistically significant as detailed on the next page. As a methodological observation, 24% of the participants made use of the scaling option to rate a task as having a workload intermediate between two whole numbers (i.e., half-point resolution between 0 – 10, possible 21 point resolution in the scale) (see Appendix I).

A final set of graphs on the next page compare combined workload ratings by gender, by age, and highlight the workload ratings for a selected set of tasks (the manual Radio Hard tuning task, the voice Radio Hard tuning task, use of the voice-command system to enter a full street address, and the 1-back auditory-vocal surrogate task). Statistical tests are provided for a selected set of comparisons to address the questions of whether: there are overall main effects of gender, main effects of age, whether engaging with the voice-command interface to tune to a specific radio station is less demanding than using the traditional manual tuning method, and whether engaging with the voice interface to enter a full street address into the navigation system is more or less demanding than a version of the manual radio tuning reference task. While all possible task comparisons could be considered and reported, this has been avoided in the current technical report due to concerns around spurious Type I errors arising from excessive multiple-comparisons. This same presentation style is continued for the presentation of the majority of the variables considered in the primary analysis.

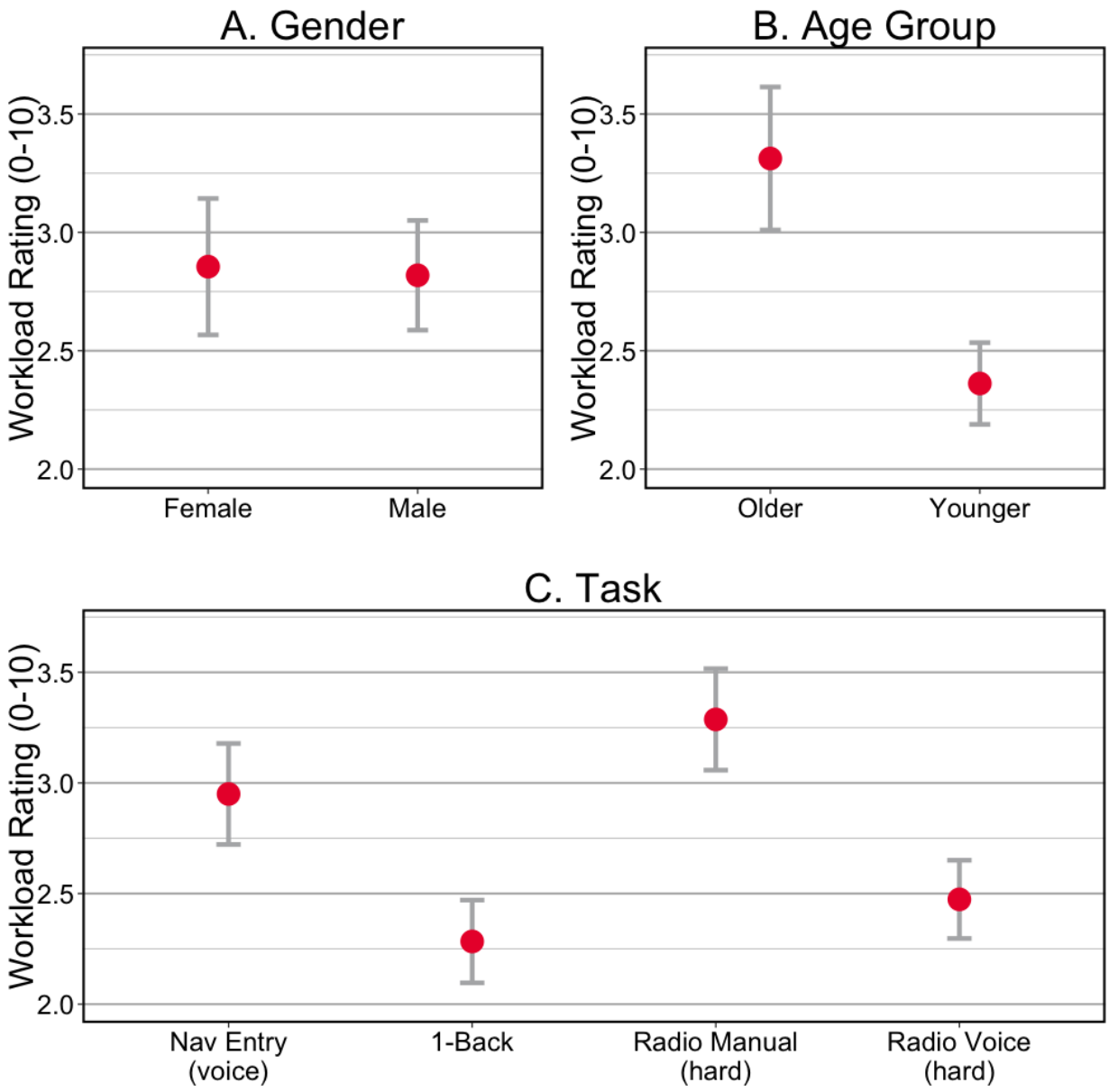


Figure 18: Selected statistical summary plots for self-reported workload.

There was no overall effect of gender on self-reported workload (p = .836, Wilcoxon test). There was, however, an effect of age group (p = .011, Wilcoxon test), with older participants reporting significantly higher workloads than younger. Interestingly, there was one notable exception in this pattern; younger participants rated the traditional manual tuning of the radio by manually rotating the tuning knob to locate a specific station as more demanding (3.67) than older participants (2.96). One might speculate that this represents a situation where years of

familiarity with an "archaic"method of accomplishing a task resulted in a higher acceptability. There was a significant difference between the voice and manual interfaces for the radio tuning task ($p = .036$), with the voice command option for the radio tuning task being rated as less demanding. The Nav Entry task was nominally lower, but did not differ statistically from the manual Radio Hard task ($p = .275$) on the global self-report workload measure.

.

Task Completion Time

Task completion time represents another way of evaluating the demand and potential distraction associated with a task. Figure 19 shows a plot of the total time in seconds from the start to the completion of each of tasks that could be evaluated along this dimension. The N-Back tasks and Song Fail task are excluded, as their total task times were fixed durations.

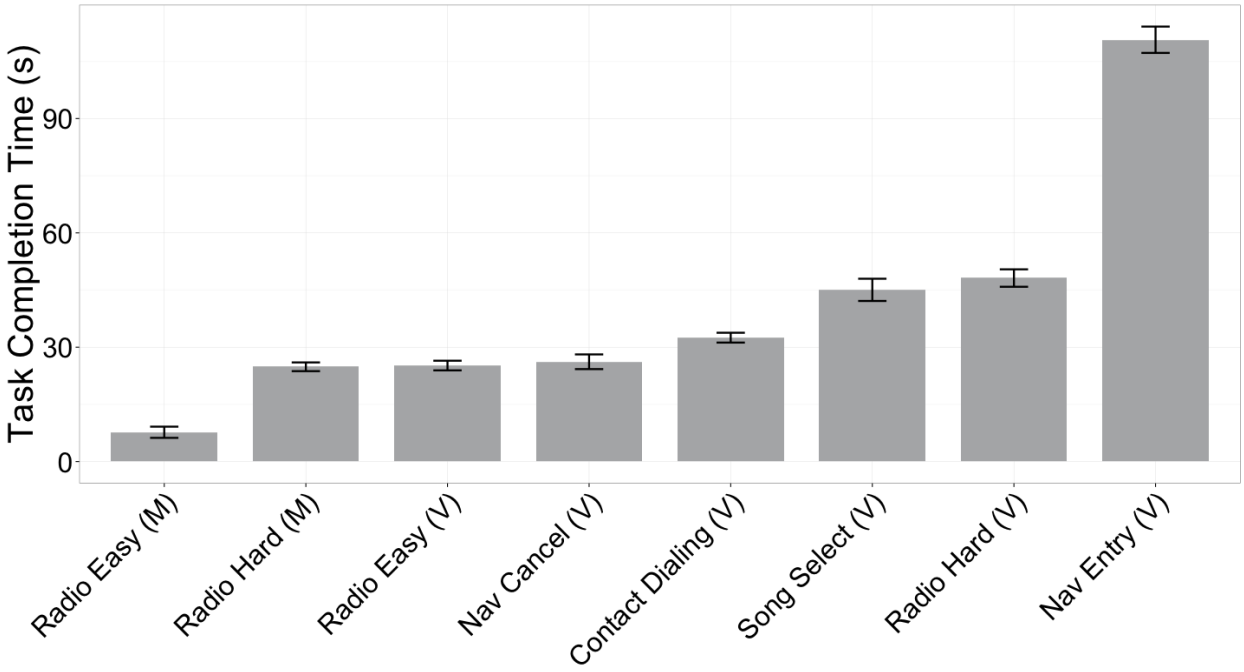


Figure 19: Tasks listed in ascending order for the amount of time needed to complete each task. Error bars represent 1 SEM. Tasks marked (V) used the voice interface. Tasks marked (M) utilized traditional manual/tactile interactions. (Note: the n-back tasks and song fail task are of fixed duration and therefore are not represented in the plot.)

A listing of the means and standard deviation values for each of the tasks depicted in Figure 19 is provided in Table 8 on the next page. As expected, as a single step task, pressing a single preset button in the manual Radio Easy task was completed relatively rapidly, with mean time from the prompt to start the task ("Begin") to completion of the task (experimenter pressing a timing key when the participant says "Done") being less than 8 seconds (M 7.7; SD 8.7) across the younger and older participants. This value compares favorably with the 10.5 second mean (SD 5.4) duration for the manual Radio Easy task collected under on-road conditions as part of the CAMP DWM work (Angell et al., 2006). The fact that the task completion time for the manual Radio Easy task in the present study was less than that obtained in the DWM study may be useful to keep in mind when considering the total completion times for other tasks.

It is appropriate to note that some of the total task time in the numbers from both studies consists of the time taken by the participant to say "done" and for that to be recorded by the experimenter. The Alliance (2006) guidelines explicitly recognize that timing of the task based on a completion indication on the part of the participant may often need to be used as an operational definition of the task end state. Similarly, the new NHTSA (2013) guidelines specify that the end of data collection "means the time at which a test participant informs the experimenter that they have completed a testable task either by speaking the word, 'done', or, by a non-verbal means (such as a button press) indicating the same thing" (p. 248). In the case of voice-command interactions in this study, the experimenter pressed a time recording button at the point the participant spoke whatever command terminated the task.

Table 8: Means (and standard deviations) for task completion times.

Task	Younger	Older	All
Nav Cancel	21.00 (2.5)	31.37 (26.0)	26.18 (19.0)
Nav Entry	100.44 (18.1)	120.90 (41.7)	110.67 (33.5)
Phone	26.72 (7.8)	38.49 (15.5)	32.51 (13.5)
Radio Manual Easy	4.94 (2.1)	10.43 (11.6)	7.68 (8.7)
Radio Manual Hard	20.20 (4.9)	29.53 (7.4)	24.86 (7.8)
Radio Voice Easy	20.74 (4.8)	29.64 (13.2)	25.19 (10.8)
Radio Voice Hard	40.28 (9.7)	56.01 (27.0)	48.14 (21.6)
Song Select	30.78 (12.0)	59.32 (31.1)	45.05 (27.4)

Table 9: CAMP DWM statistics for task completion times for the radio tasks (from Angell et al. (2006), Appendix Q, p. Q-20) for 101 participants.

Task	Mean	SD	Minimum	Maximum
Radio Manual Easy	10.45	(5.4)	3.91	40.54
Radio Manual Hard	15.39	(6.29)	6.57	48.9

In contrast with the results for the manual Radio Easy task, mean task completion time for the manual Radio Hard task was markedly longer in this vehicle than what was reported in the CAMP WDM results. Total task time in the WDM study was 15.4 seconds for the manual Radio Hard task while the mean value in our test-vehicle was close to 25 seconds. It does not seem likely that this difference is due to characteristics of our sample or procedure since our participants completed the manual Radio Easy task in less time than those in the WDM study.

The radio in our test vehicle used a traditional rotary tuning knob, which should be optimal for manual tuning of stations (Perez et al., 2013). On the other hand, when the driver wishes to change radio bands, physically pressing the "RADIO" button causes soft buttons to appear on the touch screen display for selecting AM, FM1, FM2, SAT1, SAT2, or SAT3. In a recent simulator study comparing phone dialing using traditional push-buttons vs. touch screen soft buttons, we found that interacting with the touch screen interface took longer (Reimer et al., 2012). It may be that orienting to and interacting with the relatively small touch screen buttons may account for a portion of the relatively long total task time seen here compared to having a discrete band selection push-button as in the older WDM vehicle interface. The movement to soft buttons to present the expanded list of "band" selections in modern entertainment systems seems like a reasonable approach to dealing with the added options. However, this illustration highlights the fact that even the "basic" radio interface in the vehicle has become substantially more complex than when manual radio tuning was originally proposed as a fundamental reference task for acceptable visual-manipulative demand in the vehicle.

While the voice-command method of carrying out the Radio Hard task was given a lower self-reported workload rating than the manual method of completing the task, it is also clear that the total time to complete the Radio Hard tuning task using the voice interface was significantly longer, at a mean duration of 48 seconds. Not surprisingly, given the number of steps involved in the task and the pacing aspects of voice input and confirmation, voice-based destination entry required participant engagement for the longest time of any of the tasks. The sample as a whole had a mean task completion time of 111 seconds and the older adults showed a mean task time of just over 2 minutes. Burns, Harbluk, Foley and Angell (2010) provide a very useful discussion relative to considering total task time as an important metric in considering designs intended to limit distraction.

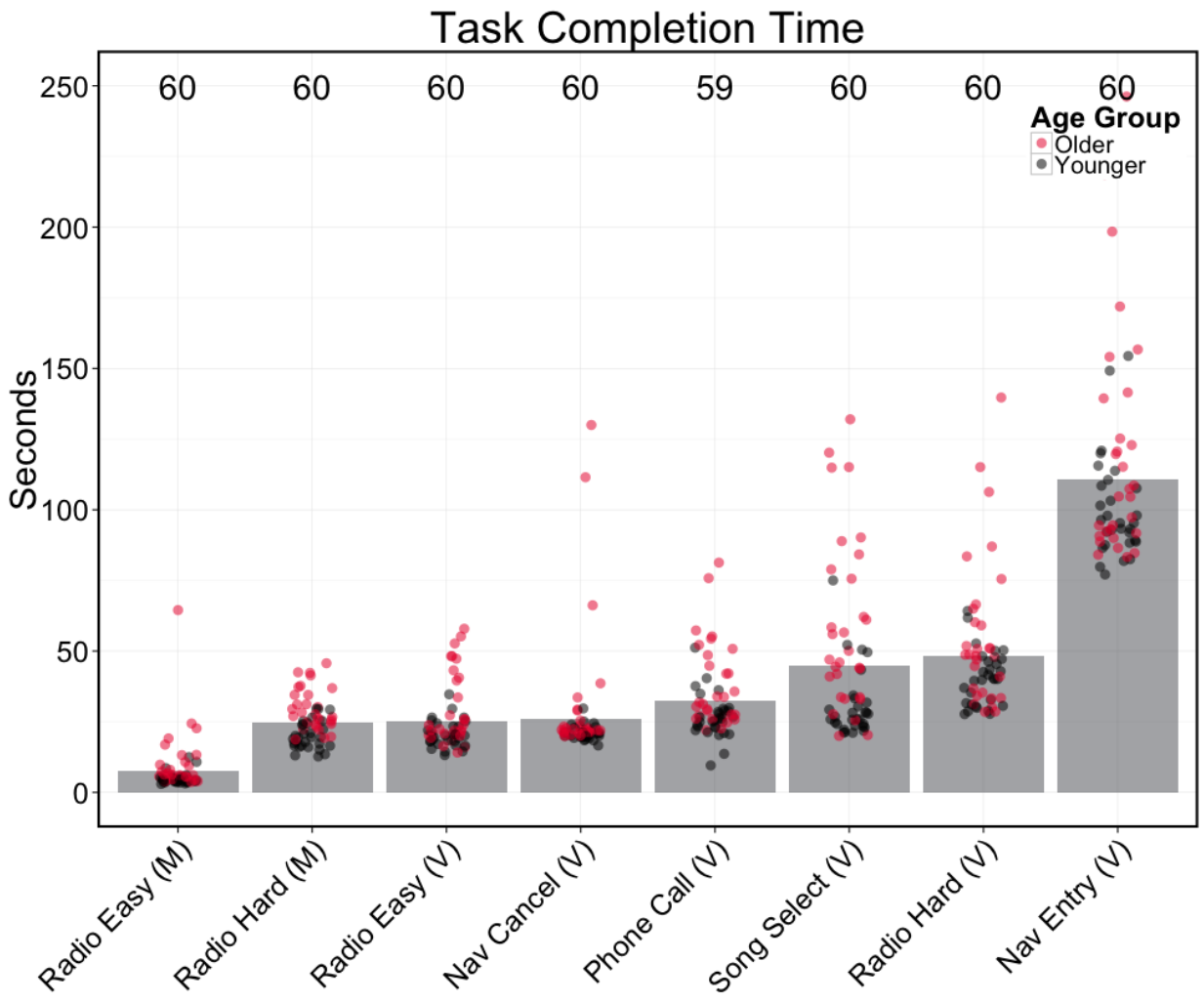


Figure 20. Distribution of the amount of time needed to complete each task across participants. (Note: One participant was unable to complete the Phone tasks due to equipment failure and is not shown.)

As can readily be deduced from the distribution of gray (younger, 20-29 years) and red (older, 60-69 years) circles, the range of task completion times was much wider for the older participants. While many older participants completed tasks well within the central distribution of times displayed by younger participants, the distribution tail for older individuals extended markedly in the direction of longer completion times. The overall effect of age on task completion time was statically significant ($p < .001$) as shown on the planned comparison tests on the next page.

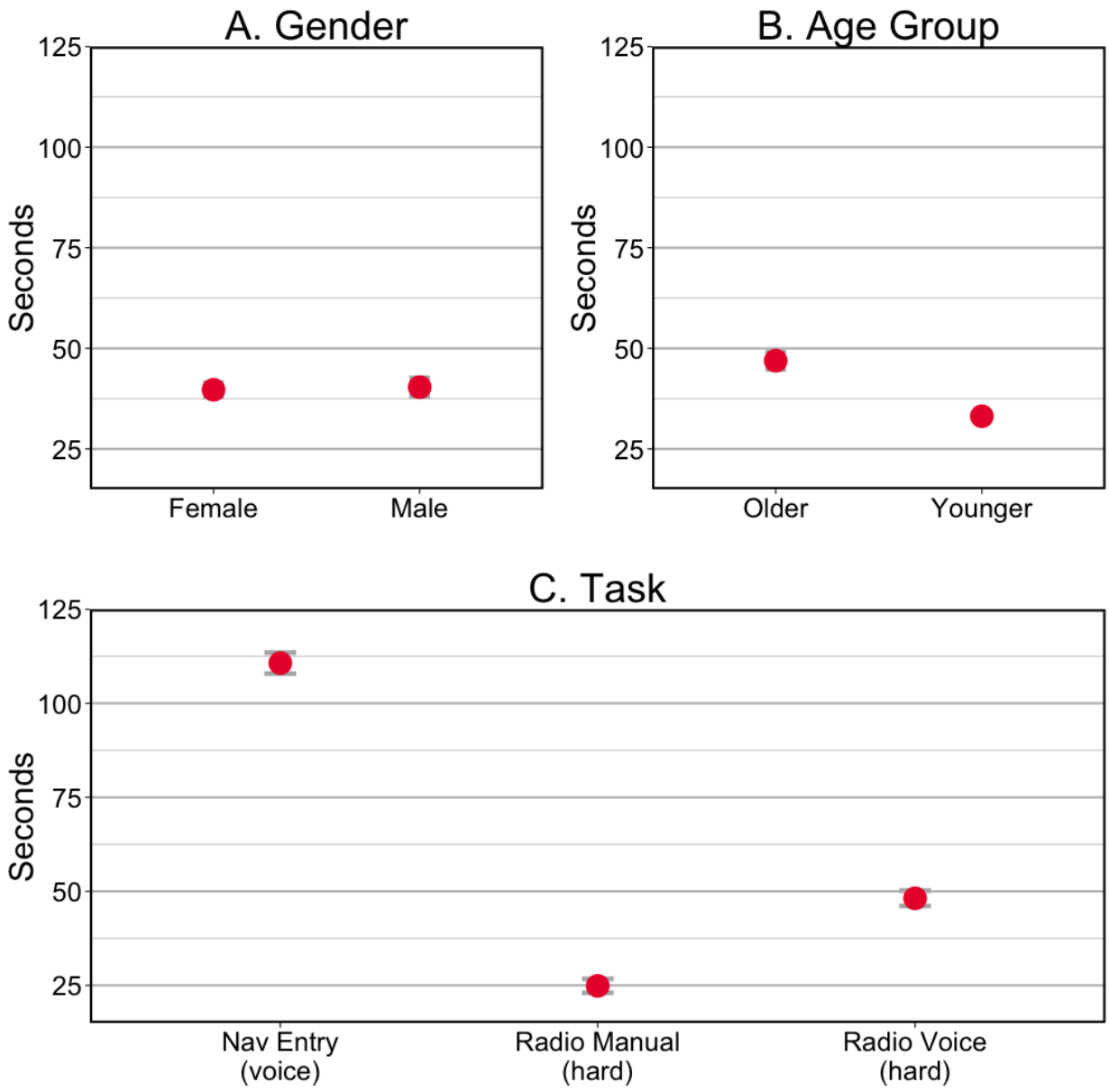


Figure 21: Statistical summary plots for task completion time.

The effect of gender on task completion time was not significant ($p = 0.935$), without even a nominal difference apparent between males and females. As already discussed, the effect of age group was significant ($p < .001$) with younger participants generally completing tasks in less time than older participants. The voice command method of completing the Radio Hard task took more time than the manual method of tuning the radio (48 vs. 25 seconds; $p < .001$), although this may be compensated for in part by the lower self-reported workload rating detailed in the previous section. As noted previously, at 111 seconds, the Nav Entry task took

significantly longer to complete than the reference manual Radio Hard tuning task ($p < .001$). The 1-Back task is not included in the plot because it employed a fixed task duration.

Physiological Measures

Heart Rate

Task periods are listed in order of mean absolute heart rate in Figure 22 below. This plot also includes a broad baseline that combines mean heart rate values across all of the baseline periods. A repeated measures ANOVA on the heart rate data shows that significantly different arousal levels were present across the tasks ($F(12, 684) = 13.01$, $p < .001$).

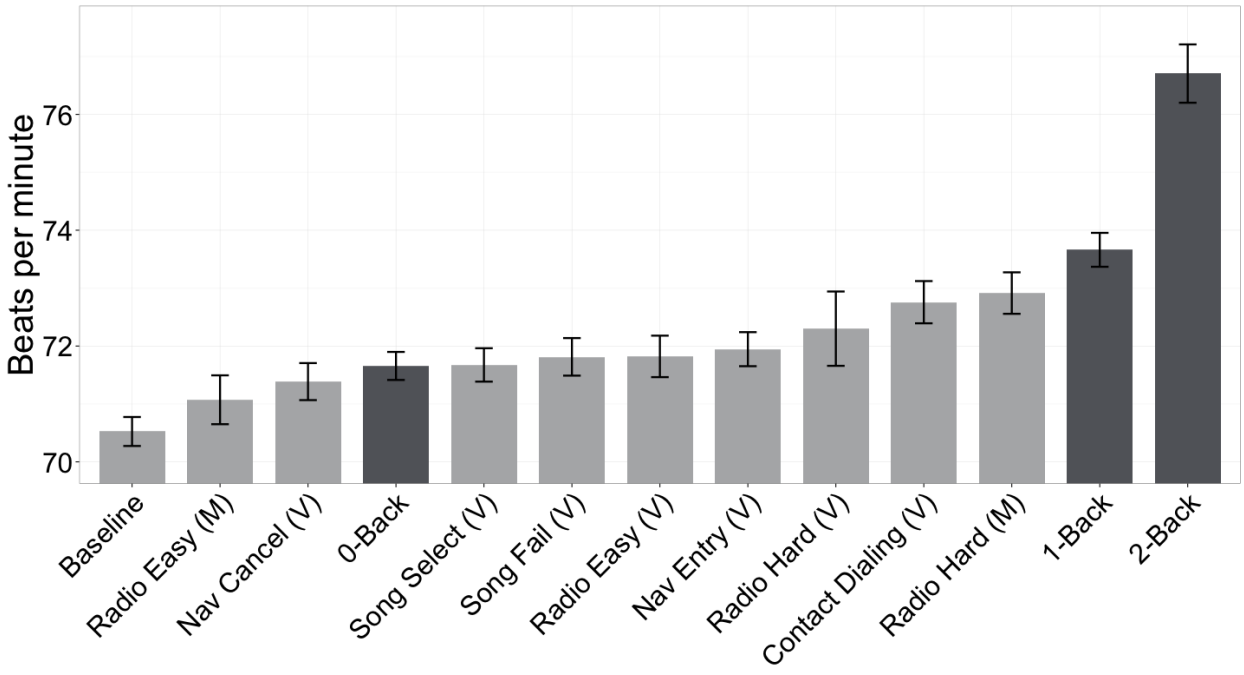


Figure 22: Tasks listed in ascending order for mean heart rate. The baseline shown represents the combined baseline periods recorded prior to each task. Error bars represent 1 SEM. (Figure adapted from (Reimer, et al., 2013).)

As expected, the lowest mean heart rate was present during the baseline (single task driving) reference period. Overall, heart rate showed a 1.6% increase over the baseline period during the 0-back task, a 4.5% increase during the 1-back, and an 8.9% increase during the 2-back. This pattern of response is consistent with previous findings for heart rate responses to randomly ordered n-back tasks in simulation (Mehler & Reimer, 2013; Son, et al., 2011) and a large on-road study (Mehler, et al., 2012) of 108 individuals distributed across 3 age groups (20-29, 40-49, and 60-69).

The in-vehicle task period with the lowest heart rate was manual Radio Easy task (pressing a station preset button) and the highest heart rate was associated with the manual Radio Hard

tuning task. These physiological measures of demand align relatively well with self-reported workload.

Table 10: Means (and standard deviations) of percent heart rate change.

Task	Younger	Older	(All)
Nav Cancel	1.78 (4.3)	0.88 (2.9)	1.33 (3.6)
Nav Entry	2.96 (4.8)	1.29 (2.9)	2.12 (4.0)
0-Back	2.24 (3.8)	0.99 (3.2)	1.62 (3.5)
1-Back	5.24 (5.6)	3.83 (3.9)	4.54 (4.8)
2-Back	10.95 (8.8)	6.79 (4.9)	8.87 (7.4)
Phone	4.15 (7.0)	2.33 (3.5)	3.25 (5.6)
Radio Manual Easy	0.71 (5.5)	1.27 (3.8)	0.99 (4.7)
Radio Manual Hard	3.18 (5.6)	4.01 (3.1)	3.60 (4.5)
Radio Voice Easy	3.07 (6.0)	0.52 (3.2)	1.80 (4.9)
Radio Voice Hard	2.71 (5.3)	2.18 (8.5)	2.44 (7.1)
Song Fail	2.42 (5.8)	1.27 (3.7)	1.85 (4.9)
Song Select	2.36 (5.4)	0.80 (3.7)	1.58 (4.7)

An inspection of Figure 22 on the previous page and the mean values in Table 10 above show an increase in mean heart rate relative to baseline across the sample during task periods. At the same time, it is clear in Figure 23 (next page) that presents heart rate in terms of individual change scores, that for most tasks, there were many participants who showed modest changes or even decreases in heart rate relative to the baseline. The significance of this pattern will be considered in more depth in a subsequent report. The most notable exception to this general observation occurs with the 1-back and 2-back tasks where almost all participants show some degree of heart rate increase during the secondary task period.

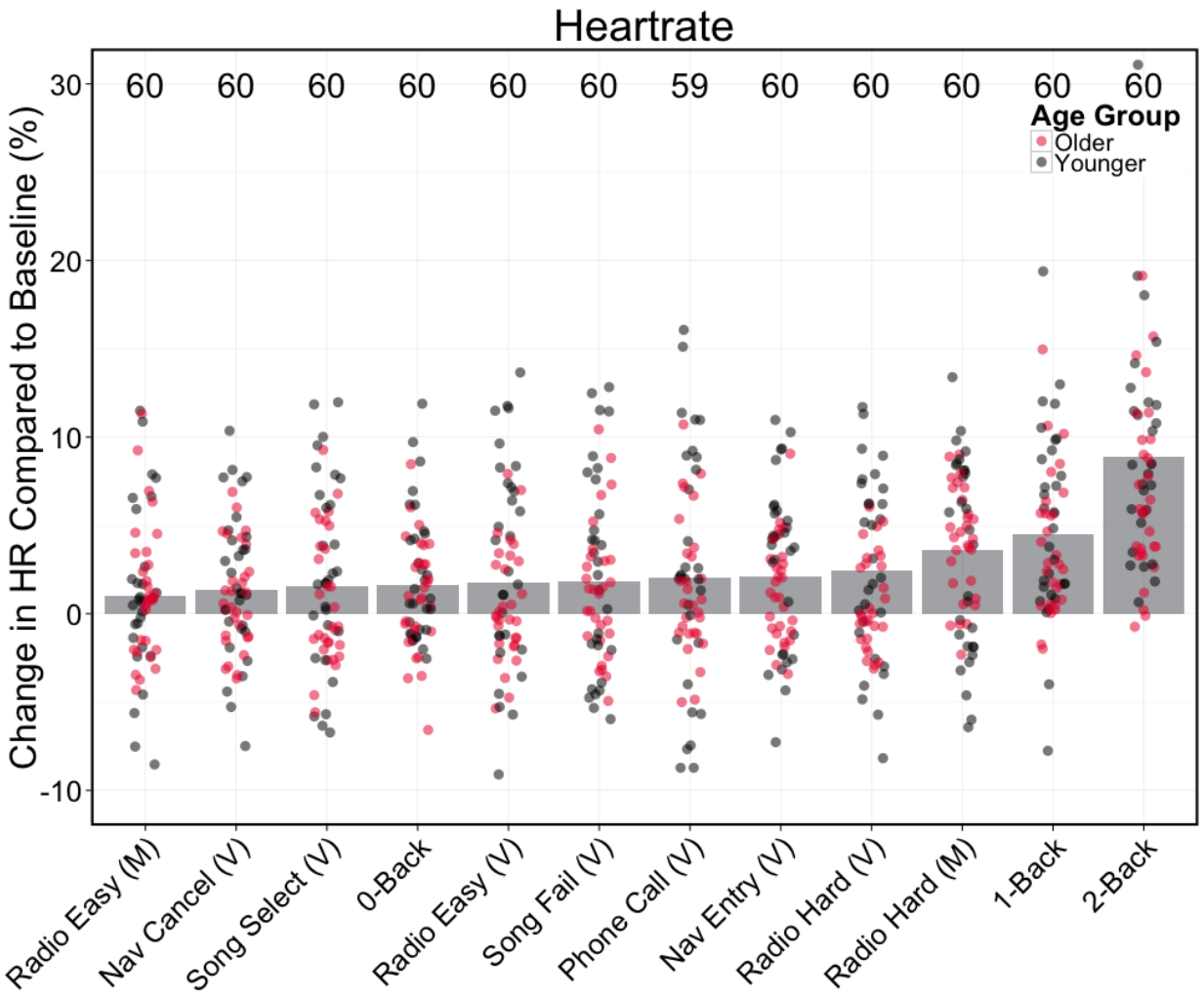


Figure 23: Percent change in heart rate during task periods relative to an averaged baseline period of single-task driving. (Note: Heart rate data were unavailable for one participant during the Phone task.)

It is interesting to note an apparent bimodal distribution in heart rate change scores in the younger participants relative to older participants. Older participants are largely grouped in the center of the distribution for each task while younger participants tend to clump at the two extremes, showing either more prominent drops in mean heart rate or more prominent increases in heart rate relative to the baseline reference.

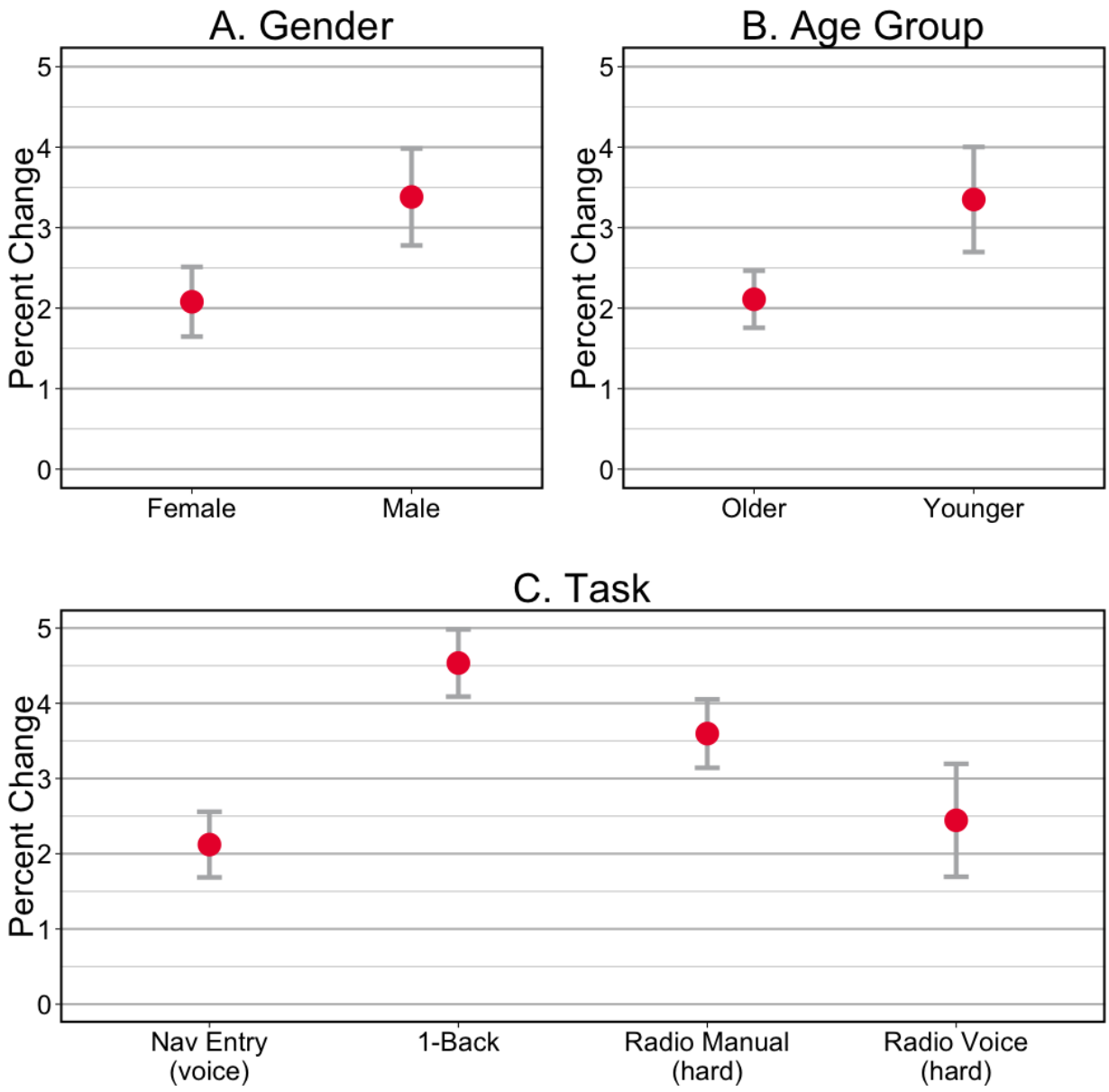


Figure 24: Statistical summary plots for heart rate change relative to an averaged baseline period of single-task driving.

The effects of gender and age on heart rate were not significant (p = .096 & .230, Wilcoxon test). (The apparent bimodal aspect of how younger participants' heart rate changes scores noted on the previous page should be kept in mind in terms of interpreting the apparent lack of an age effect.) The change in heart rate was significantly higher for the manual Radio Hard task than for the using the voice interface (p = .036). Similarly, the change in heart rate for the Nav Entry task was lower than that observed with the manual Radio Hard task (p = .003).

Skin Conductance Level (SCL)

As was done with heart rate, the SCL data is presented by arranging tasks in order of mean absolute SCL (see Figure x). This plot includes a broad baseline that combines mean values across all of the baseline periods. SCL shows a significant main effect of task period ($F(12, 576) = 4.72$, $p < .001$).

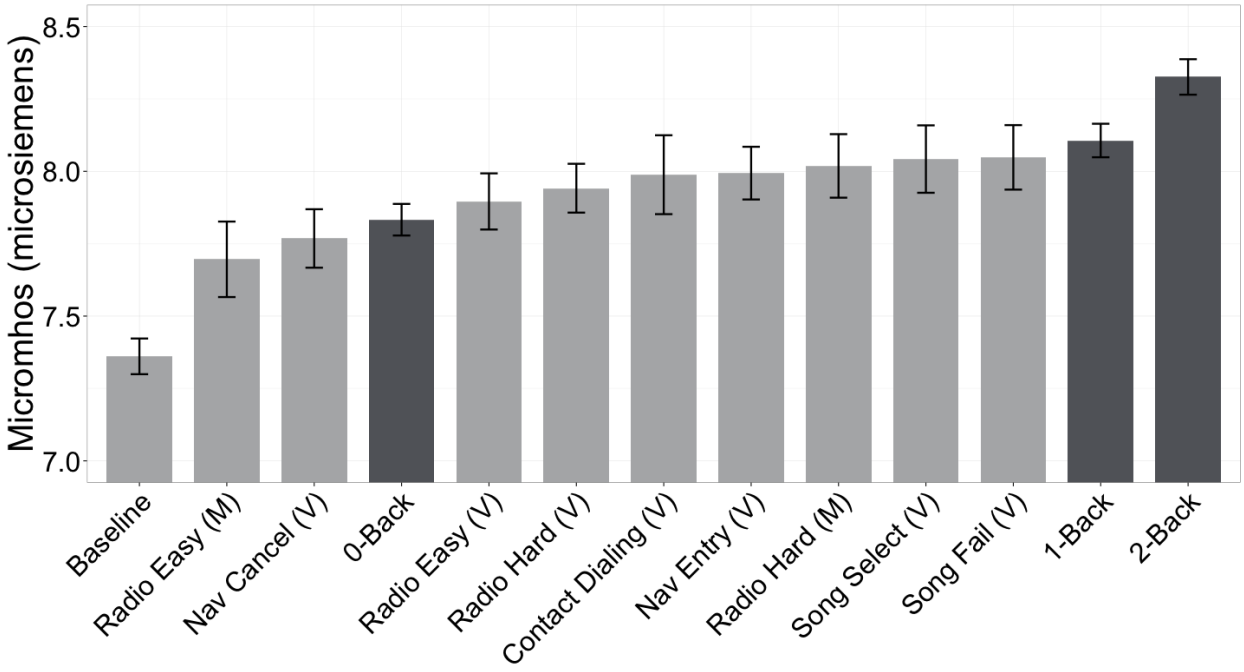


Figure 25: Tasks listed in ascending order for mean SCL. The baseline shown represents the combined baseline periods recorded prior to each task. Error bars represent 1 SEM. (Figure adapted from (Reimer, et al., 2013).)

As was the case with heart rate, a clear, stepwise increase in SCL can be observed across the three levels of the n-back calibration task. SCL showed a 6.8% increase over the baseline period during the 0-back task, an 11.2% increase during the 1-back, and a 15.4% increase during the 2-back. This pattern of response is generally consistent with the findings for SCL responses to randomly ordered n-back tasks in simulation (Mehler & Reimer, 2013; Son, et al., 2011) and a large on-road study (Mehler, et al., 2012) of 108 individuals distributed across 3 age groups (20-29, 40-49, and 60-69). The absolute value change and percentage changes seen here for 30 second long n-back task periods were not as large as those observed during the more sustained, 2 minute long n-back periods used in the previous on-road study.

As was the case with heart rate, the in-vehicle task periods with the lowest SCL values were for the manual Radio Easy task and the Nav Cancel task, with both falling below the 0-back task

level. Also consistent with the heart rate data, all remaining tasks fell below the 1-back task in terms of general physiological arousal as measured by SCL.

Table 11: Means (and standard deviations) of percent SCL change.

Task	Younger	Older	(All)
Nav Cancel	4.20 (10.5)	8.43 (12.4)	6.44 (11.6)
Nav Entry	7.42 (13.6)	13.11 (13.1)	10.43 (13.5)
0-Back	4.16 (7.7)	9.19 (8.6)	6.82 (8.5)
1-Back	5.99 (10.8)	15.79 (13.9)	11.18 (13.4)
2-Back	9.96 (11.9)	20.31 (13.1)	15.44 (13.5)
Phone	13.33 (14.9))	6.86 (23.4)	9.97 (19.8)
Radio Manual Easy	3.98 (13.5)	7.67 (18.1)	5.93 (16.0)
Radio Manual Hard	8.43 (17.0)	15.95 (19.8)	12.41 (18.7)
Radio Voice Easy	3.80 (12.9)	9.33 (12.4)	6.67 (12.8)
Radio Voice Hard	4.96 (12.2)	11.86 (14.5)	8.61 (13.8)
Song Fail	6.37 (16.6)	14.49 (20.2)	10.67 (18.8)
Song Select	6.13 (15.7)	13.08 (21.4)	9.81 (19.1)

With the exception of the phone dialing task, older participants showed greater percentage changes in SCL during each of the tasks.

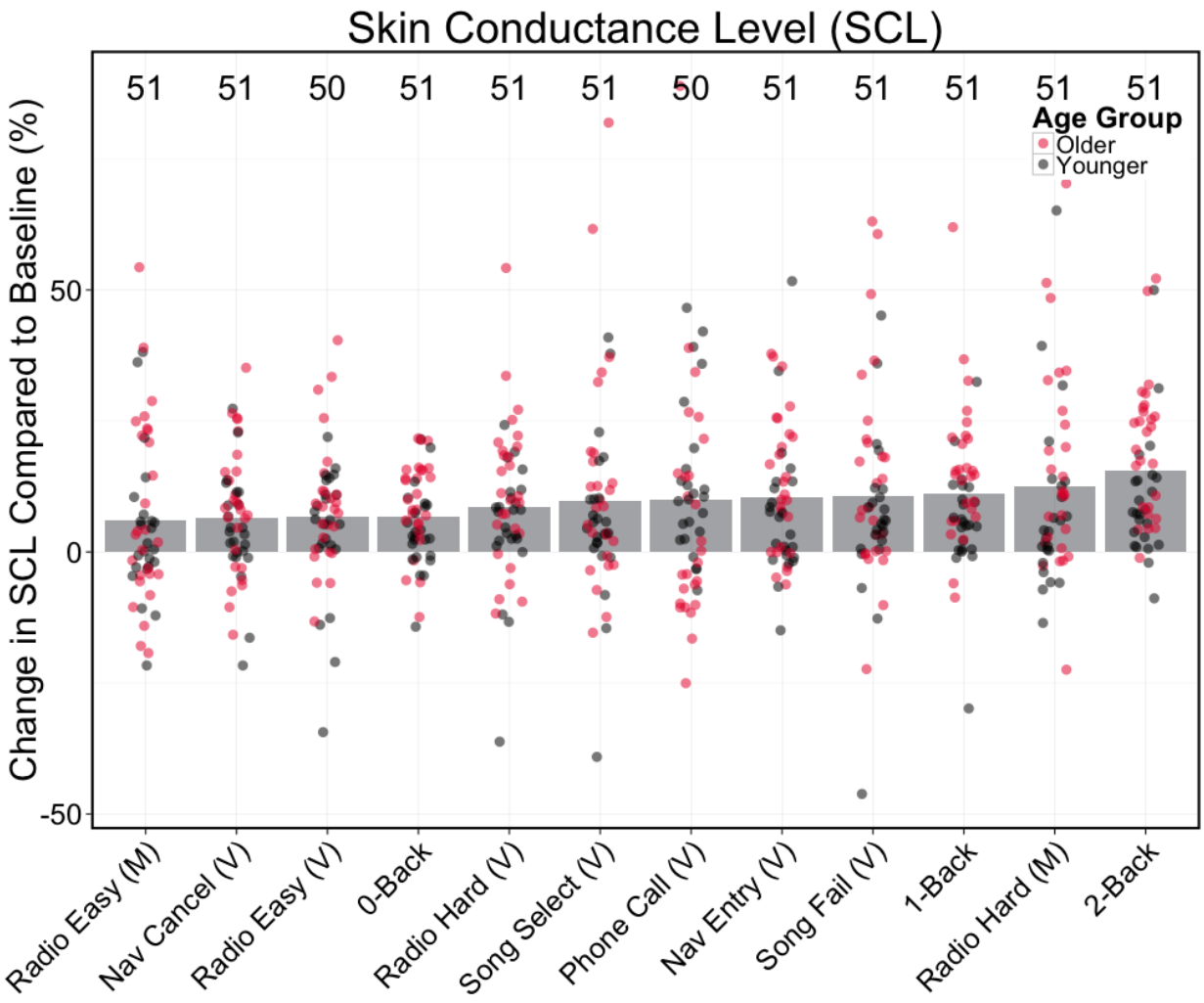


Figure 26: Percent change in skin conductance level (SCL) relative to an averaged baseline period of single-task driving. (Note that SCL data were not available for all participants and tasks as indicated by the numbers at the top of the graph above each task.)

The distribution of SCL change scores is more evenly distributed across the age groups than was seen in heart rate values. In specific, the somewhat bimodal distribution of heart rate changes seen in the younger participants does not appear in the SCL data. Nonetheless, the overall mean percent change in SCL relative to baseline was higher in older participants (see next page).

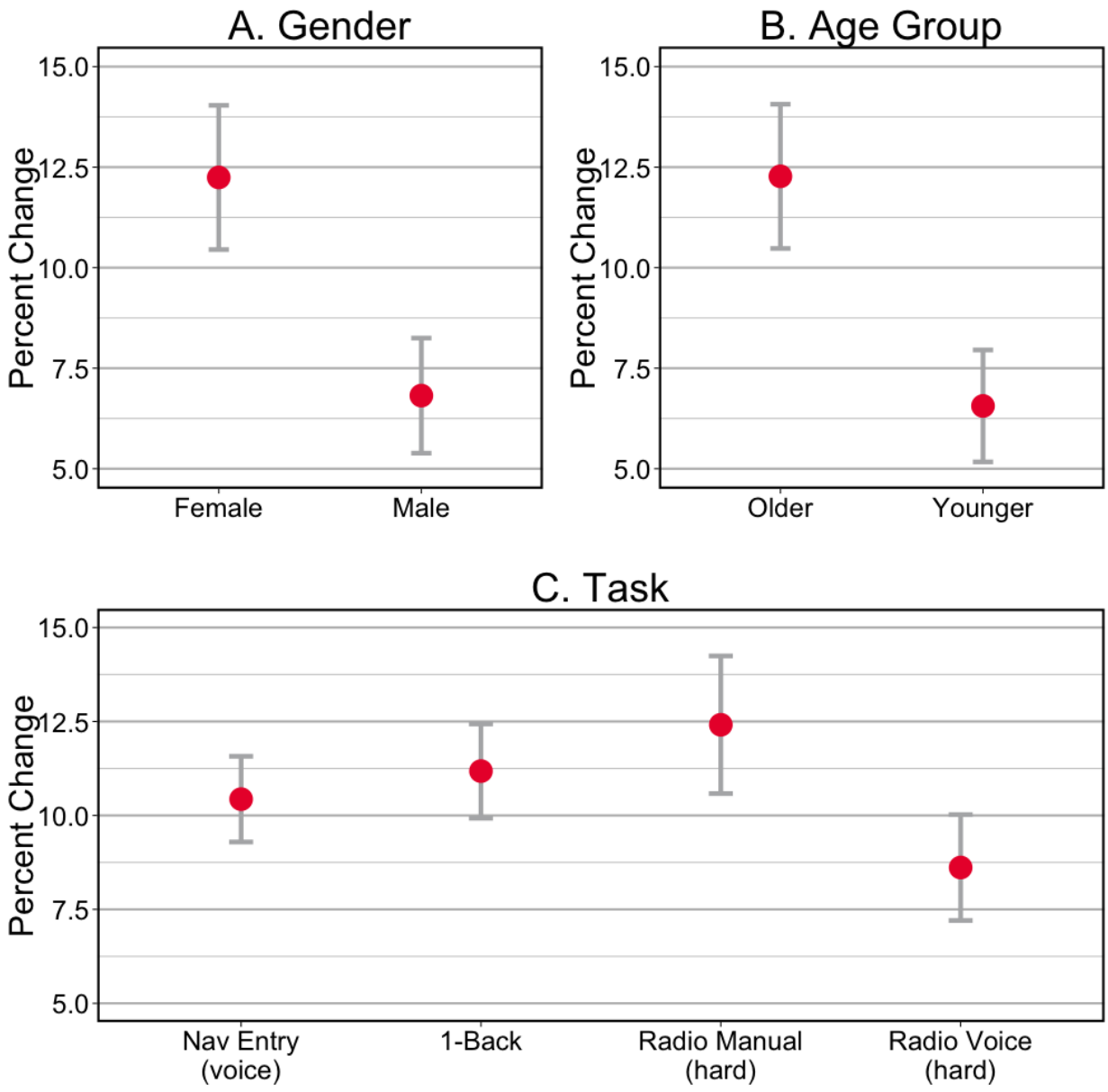


Figure 27: Statistical summary plots for SCL change relative to an averaged baseline period of single-task driving.

The effects of gender and age on SCL were significant, with women showing greater changes in SCL than men ($p = .018$), and older participants showing greater changes in SCL than younger participants ($p < .022$). SCL was not significantly different between the manual Radio Hard task and the voice-command version of the Radio Hard task ($p = .456$). Similarly, the SCL was not significantly different between the Nav Entry task and the manual Radio Hard task ($p = .884$). Thus, to the extent that heart rate and SCL may function as indirect measures of the cognitive

demand experienced by participants, both the voice-command option for the Radio Hard task and voice-command entry of an address into the navigation system might be seen as placing comparable or less demand on the driver along this dimension than the manual Radio Hard tuning task. In making this observation, it should be emphasized that neither heart rate nor SCL are direct measures of cognitive activity and that other aspects of demand upon the driver's attention need to be taken into account in assessing overall demand and distraction considerations. Driving behavior and visual demand characteristics are considered in the sections that follow.

Driving Behavior Measures

Lane Departures

Lane departures—which occur when the vehicle drifts over a lane boundary unintentionally—were extracted from the vehicle's automated lane departure warning system. Lane departure events were extremely rare. There were a total of 76 lane departure events across the entire data sample (see Table 12). Rightward departures were more common than leftward departures (51 vs. 25, respectively). The number of departures appears to be a function of time spent on the road. There were 38 departures during the 14 minutes of Baseline period driving, and 11 departures during the relatively lengthy Navigation Entry task periods (once again, it is emphasized that these counts are not a per-task mean, and instead represent the raw count across the entire study sample).

Table 12: Count of lane departures across all participants by task type as detected by the Iteris AutoVue 3G lane departure warning system

	Left	Right	Total
Baseline	17	21	38
Song Select (voice)		2	2
Radio voice activation (easy)		3	3
Radio voice activation (hard)		6	6
Radio manual input (easy)	1		1
Radio manual input (hard)	1	4	5
Phone call	1	4	5
Navigation entry	3	8	11
Navigation cancel		2	2
1-Back		1	1
2-Back	2		2

Mean Velocity

A repeated measures ANOVA of mean speed during secondary task performance shows a significant effect of task ($F(12, 684) = 5.10$, $p < .001$). Arranging the task periods in ascending order for vehicle velocity (see Figure 28), shows a clear reduction in speed during all task periods, with the exception of the n-back tasks, relative to baseline. Descriptive statistics are presented in Table x (next page) including a breakdown by age group.

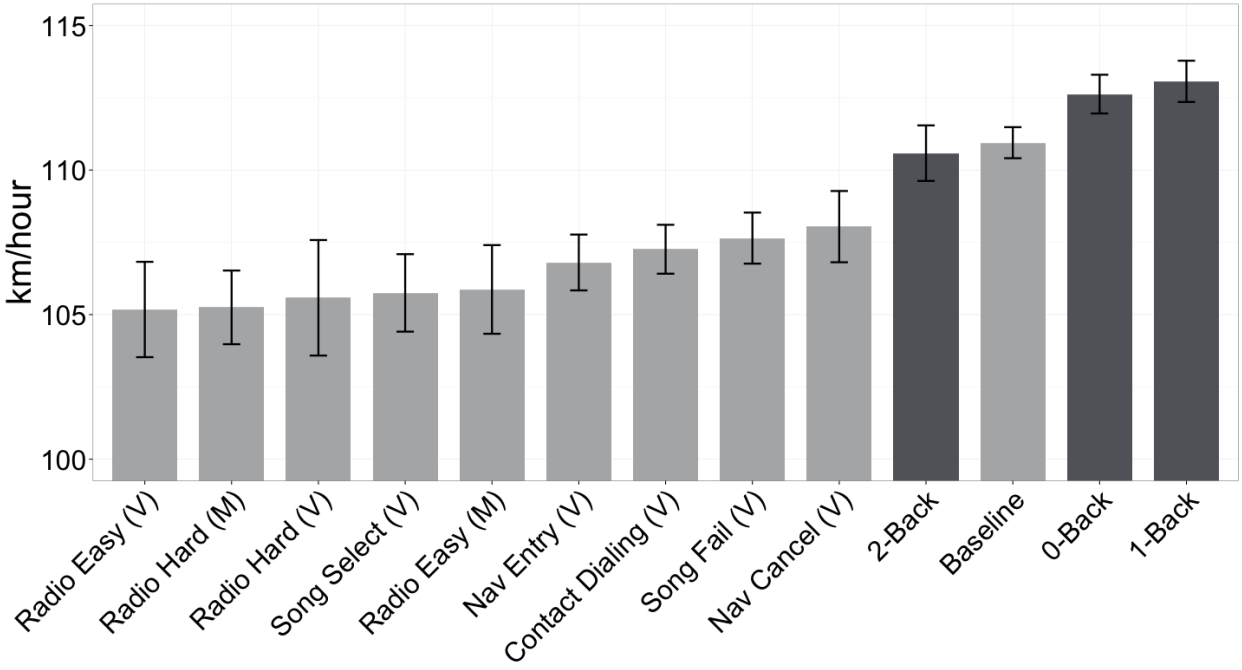


Figure 28: Tasks listed in ascending order for mean velocity. The baseline shown represents the combined baseline periods recorded prior to each task. Error bars represent 1 SEM. (Figure adapted from (Reimer, et al., 2013).)

Reductions in speed relative to single task driving as are seen above for all tasks except the n-backs, are often interpreted as compensatory behaviors to reduce workload and/or increase safety margins (Angell, et al., 2006; Horberry, Anderson, Regan, Triggs, & Brown, 2006; Lerner, Singer, & Huey, 2008).

Table 13: Means (and standard deviations) for velocity in km / hour.

Task	Older	Younger	(All)
Baseline	107.92 (8.4)	113.97 (8.8)	110.95 (9.1)
Nav Cancel	104.92 (14.7)	110.98 (11.3)	107.97 (13.4)
Nav Entry	104.69 (10.8)	108.91 (12.2)	106.80 (11.7)
0-Back	108.95 (8.6)	116.30 (9.5)	112.63 (9.8)
1-Back	110.54 (8.3)	115.59 (9.0)	113.07 (9.0)
2-Back	107.92 (14.9)	113.25 (9.0)	110.58 (12.5)
Phone	104.45 (10.6)	109.98 (8.5)	107.26 (9.9)
Radio Manual Easy	100.78 (19.0)	110.96 (9.6)	105.87 (15.8)
Radio Manual Hard	102.13 (15.9)	108.38 (13.6)	105.25 (15.1)
Radio Voice Easy	102.89 (17.4)	107.47 (18.6)	105.18 (18.1)
Radio Voice Hard	102.32 (18.6)	108.84 (20.1)	105.58 (19.6)
Song Fail	105.98 (7.8)	109.31 (8.7)	107.65 (8.3)
Song Select	101.38 (17.3)	110.12 (8.3)	105.75 (14.2)

As detailed at the end of this section, there was a statistically significant effect of age group on mean velocity across the sample as a whole, with the younger group showing an overall faster driving speed.

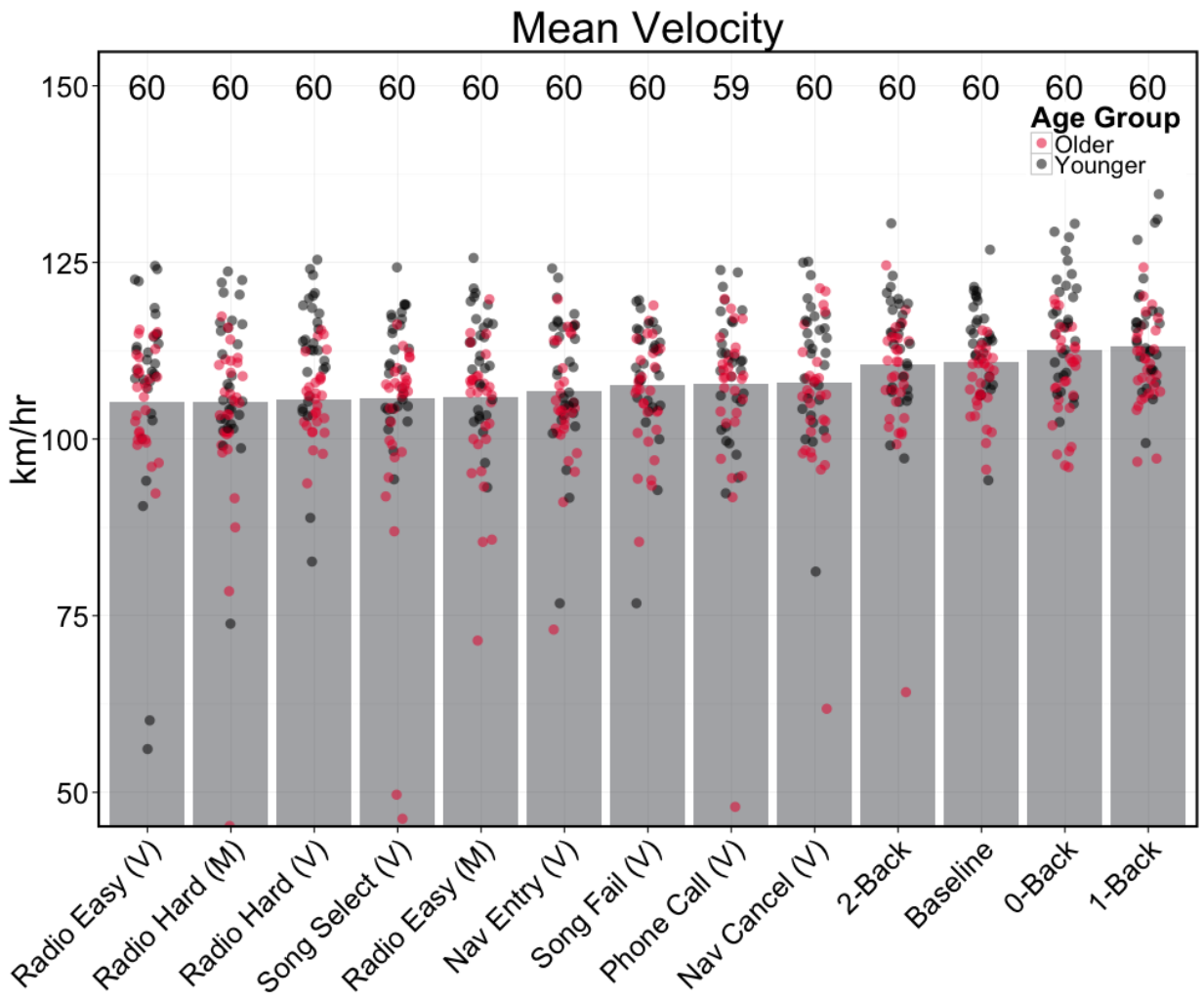


Figure 29: Mean vehicle velocity during each task period.

As noted on the previous page, drivers from the younger group tended to driver faster than drivers from the older group across the different tasks.

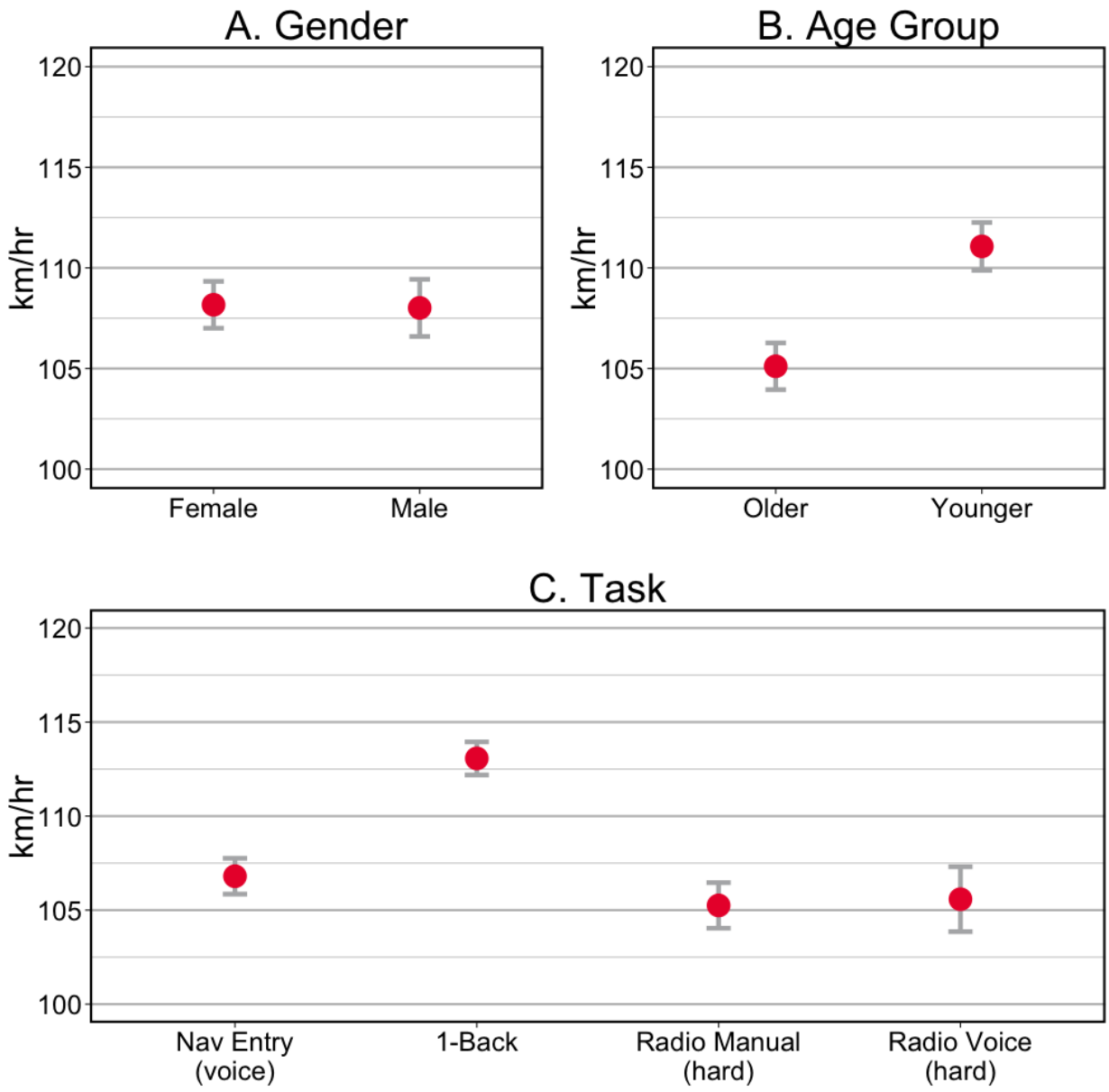


Figure 30: Statistical summary plots for mean vehicle velocity.

Mean vehicle velocity was not affected by gender ($p = 0.877$). However, velocity was affected by age group ($p < .001$), with younger drivers maintaining substantially higher speeds across all tasks. Driving speed did not differ during the manual Radio Hard task and the voice-command version of the Radio Hard task ($p = .106$). Similarly, driving speed was not significantly different between the Nav Entry task and the manual Radio Hard task ($p = .420$).

Variability of Velocity

A Friedman test on speed variability during secondary task performance shows a significant effect of task ($X^2(12) = 304.1$, $p < .001$). Arranging the task periods in ascending order for variability of velocity (see Figure 31), shows a clear reduction in variability during all task periods, with the exception of the navigation entry task, relative to baseline.

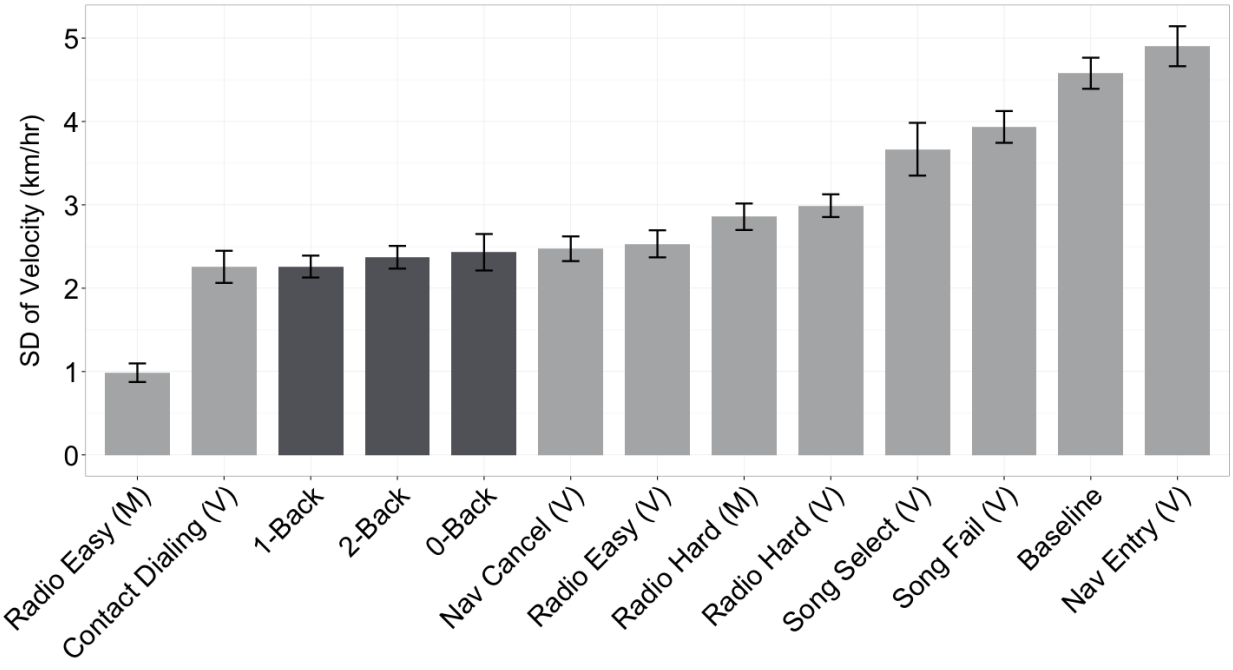


Figure 31: Tasks listed in ascending order for variability of velocity. The baseline shown represents the combined baseline periods recorded prior to each task. Error bars represent 1 SEM.

Table 14: Means (and standard deviations) for variability of velocity.

Task	Older	Younger	(All)
Baseline	4.14 (3.2)	5.02 (3.6)	4.58 (3.4)
Nav Cancel	2.43 (1.8)	2.53 (1.8)	2.48 (1.8)
Nav Entry	4.42 (2.1)	5.39 (3.5)	4.90 (2.9)
0-Back	2.43 (3.2)	2.44 (1.7)	2.43 (2.5)
1-Back	2.26 (1.8)	2.26 (1.3)	2.26 (1.6)
2-Back	2.13 (1.4)	2.61 (1.5)	2.37 (1.5)
Phone	2.87 (1.8)	2.48 (1.5)	2.67 (1.7)
Radio Manual Easy	1.20 (1.4)	0.77 (0.7)	0.99 (1.1)
Radio Manual Hard	3.05 (2.2)	2.67 (1.8)	2.86 (2.0)
Radio Voice Easy	2.64 (1.8)	2.43 (1.5)	2.53 (1.7)
Radio Voice Hard	3.13 (1.7)	2.85 (1.4)	2.99 (1.5)
Song Fail	3.93 (1.8)	3.94 (1.7)	3.93 (1.8)
Song Select	4.24 (3.9)	3.10 (2.1)	3.67 (3.2)

As detailed at the end of this section, there was a statistically significant effect of age group on mean velocity across the sample as a whole, with the younger group showing an overall faster driving speed.

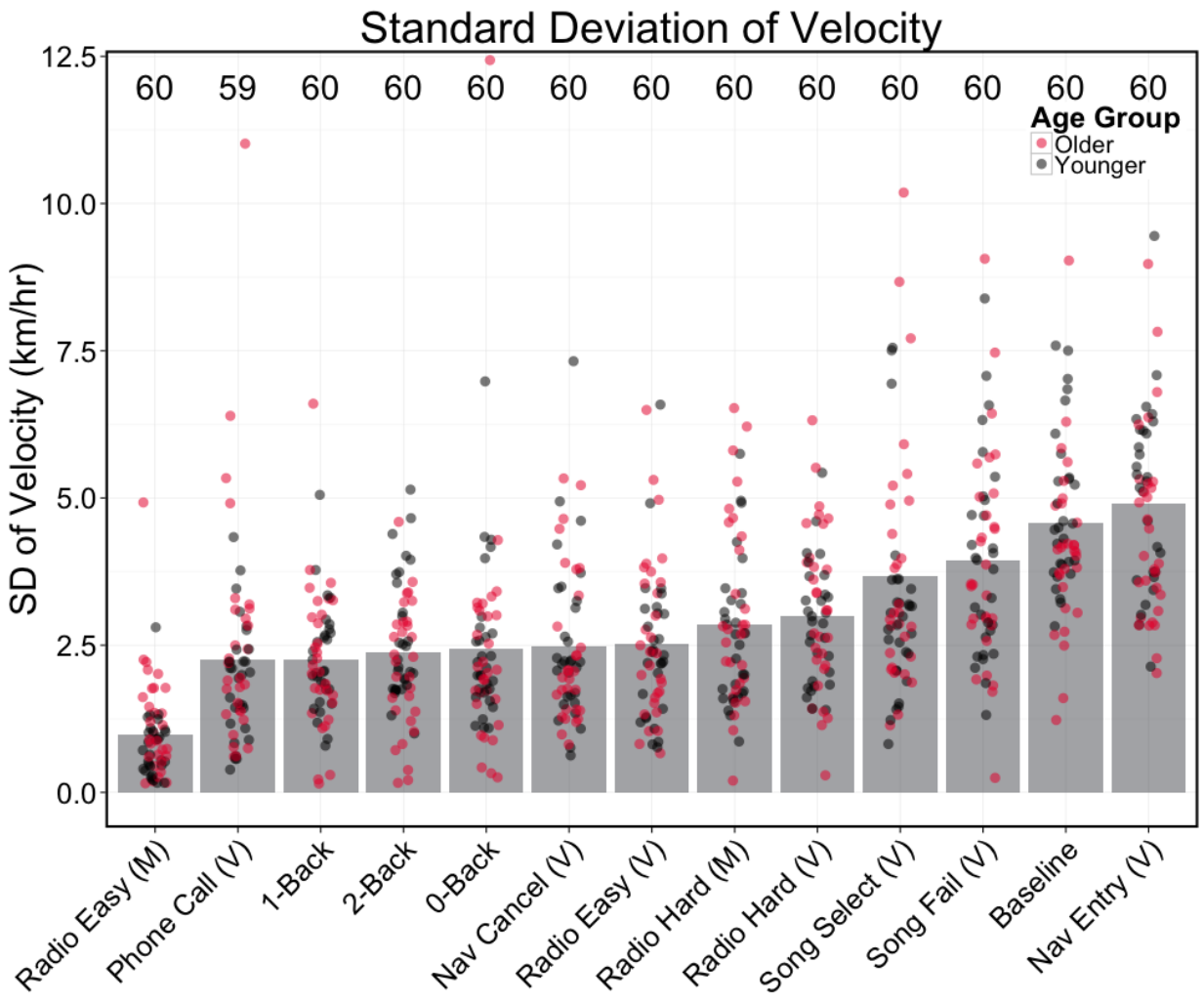


Figure 32: Variability of velocity during each task period, measured as the standard deviation of vehicle velocity over the period.

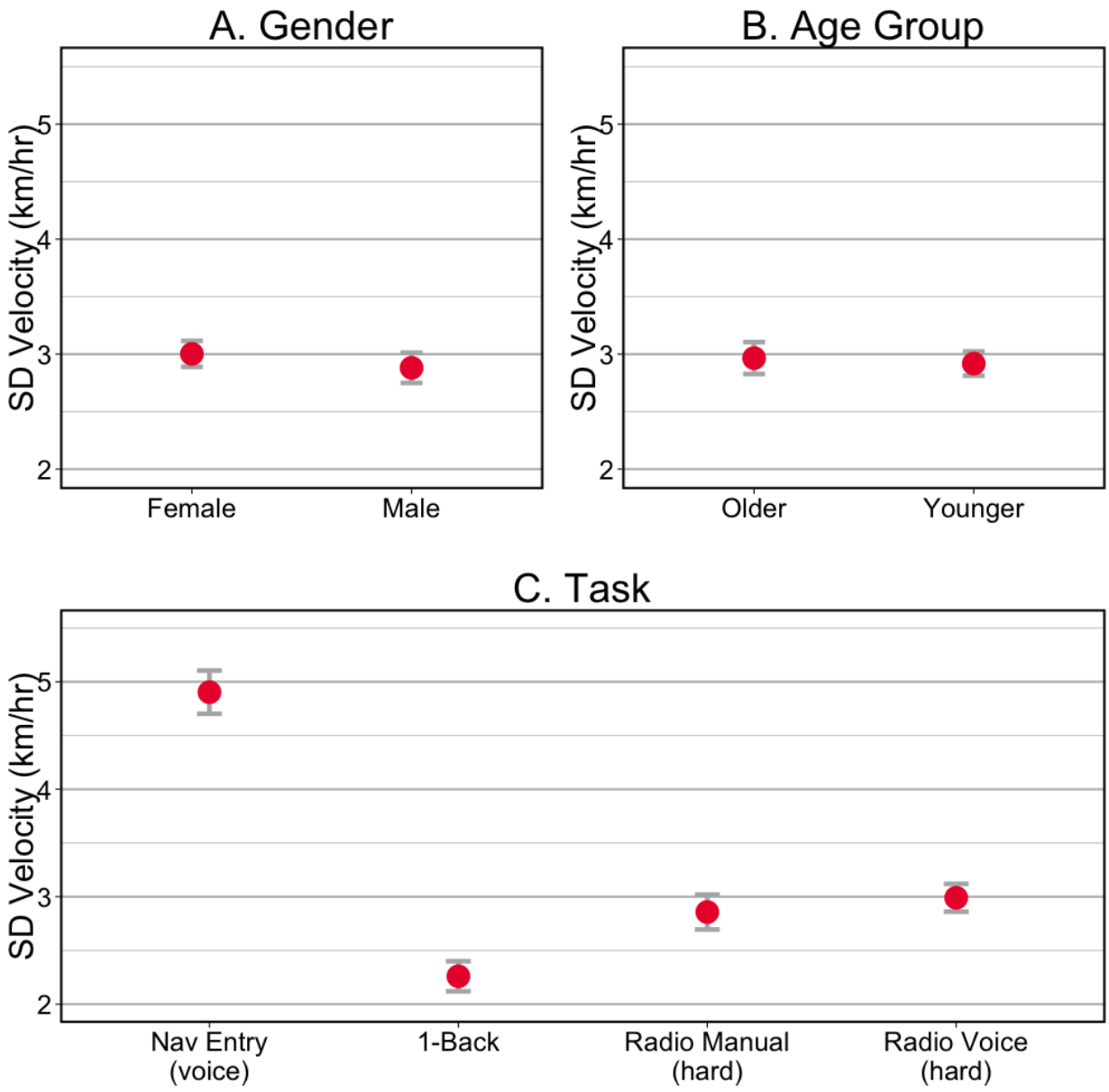


Figure 33: Statistical summary plots for standard deviation of velocity.

Standard deviation of velocity was not affected by gender or age group ($p = .665$ and $p = .775$, respectively). The standard deviation of velocity did not differ significantly between the manual Radio Hard task and doing the same task using the voice interface ($p = .237$). However, the standard deviation of velocity was significantly higher during use of the voice interface for the Nav Entry task than during the manual Radio Hard task ($p < .001$).

Acceleration Events

As detailed in the Methods section, a minimum threshold of 0.1g (0.98m/s^2) and a temporal separation of 2 seconds between independent events were applied in defining acceleration events for this report. Following this metric, Figure 34 displays acceleration events per minute for all task periods.

A Friedman test considering the frequency of acceleration events shows a significant main effect of task period ($X^2(12) = 108.2$, $p < .001$).

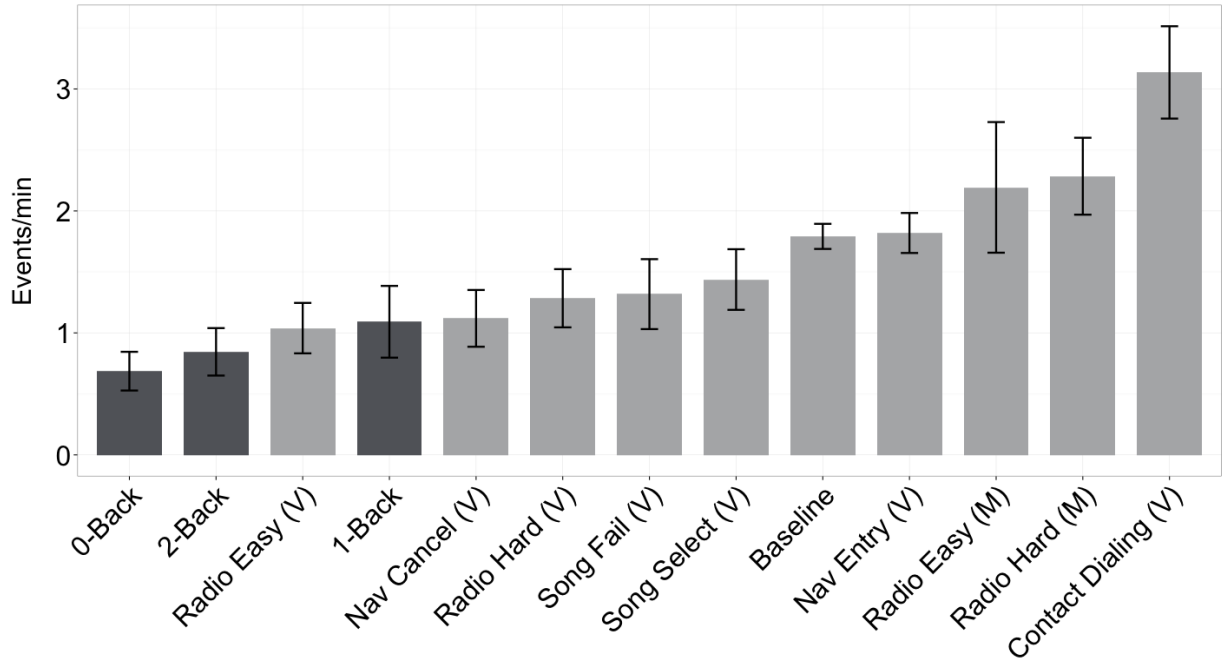


Figure 34: Tasks listed in ascending order for acceleration events. The baseline shown represents the combined baseline periods recorded prior to each task. Error bars represent 1 SEM.

Due to the safety relevance of acceleration events, in addition to the standard figures and tables provided in preceding sections, extended detail on acceleration events including a listing of all individual acceleration events equal to or greater than .2g follows at the end of this section. As detailed there, the largest individual event recorded during the formal assessment period was .31 g. Typical thresholds used to define near crash events include braking at greater than 0.50g or lateral acceleration greater than 0.40g (Fitch, et al., 2013); thus, all of the events considered here provide a window on a potentially useful metric of vehicle control but also would typically be considered as sub-critical events.

In a previous on-road study (Reimer, Mehler, Wang, et al., 2012) involving the auditory-vocal n-back task, the frequency of acceleration events during all three levels of the cognitive task dropped below the event rate observed during single task driving (baseline). This was taken as suggesting that a reduction in the frequency of low to moderate level acceleration events might provide sensitivity to the presence of cognitively loading activities. (This could be seen as parallel to the reduction in standard deviation of lane position that is sometimes observed during periods of low to moderate cognitive demand and where drivers appear more oriented toward the roadway directly ahead.) As can be observed in Figure x above, all three levels of the auditory-vocal n-back task again fall below baseline driving in terms of the frequency of low to moderate acceleration events. It is interesting to observe that the voice-command tasks that involve minimal interaction with the display screen also fall to the left of the baseline (lower acceleration event rate). The two classic visual-manual tasks (manual Radio) fall to the right, while Nav Entry, which is the task with the longest duration, is virtually indistinguishable from baseline driving on this metric. Interpretation of the data on the Phone Contact Dialing is more complex and will be addressed at another time.

Table 15: Means (and standard deviations) of minor acceleration events.

Task	Older	Younger	(All)
Baseline	1.24 (1.9)	2.37 (2.4)	1.80 (2.2)
Nav Cancel	0.99 (2.0)	1.28 (2.8)	1.13 (2.5)
Nav Entry	1.59 (2.0)	2.06 (2.0)	1.82 (2.0)
0-Back	0.31 (0.8)	1.08 (2.8)	0.69 (2.1)
1-Back	1.49 (3.4)	0.68 (1.6)	1.09 (2.7)
2-Back	0.45 (1.1)	1.25 (3.2)	0.84 (2.4)
Phone	2.33 (2.8)	3.63 (4.9)	2.98 (4.1)
Radio Manual Easy	1.74 (4.2)	2.66 (8.1)	2.19 (6.4)
Radio Manual Hard	2.03 (2.7)	2.55 (4.5)	2.28 (3.7)
Radio Voice Easy	0.58 (1.2)	1.51 (2.9)	1.04 (2.2)
Radio Voice Hard	0.60 (1.1)	2.00 (3.9)	1.28 (2.9)
Song Fail	1.49 (2.1)	1.14 (2.6)	1.32 (2.3)
Song Select	1.05 (1.8)	1.83 (4.2)	1.44 (3.2)

As can be observed in the entries for the individual tasks above, the frequency of acceleration events was higher in the younger group. As detailed later, the overall effect of age on frequency of acceleration events was statistically significant.

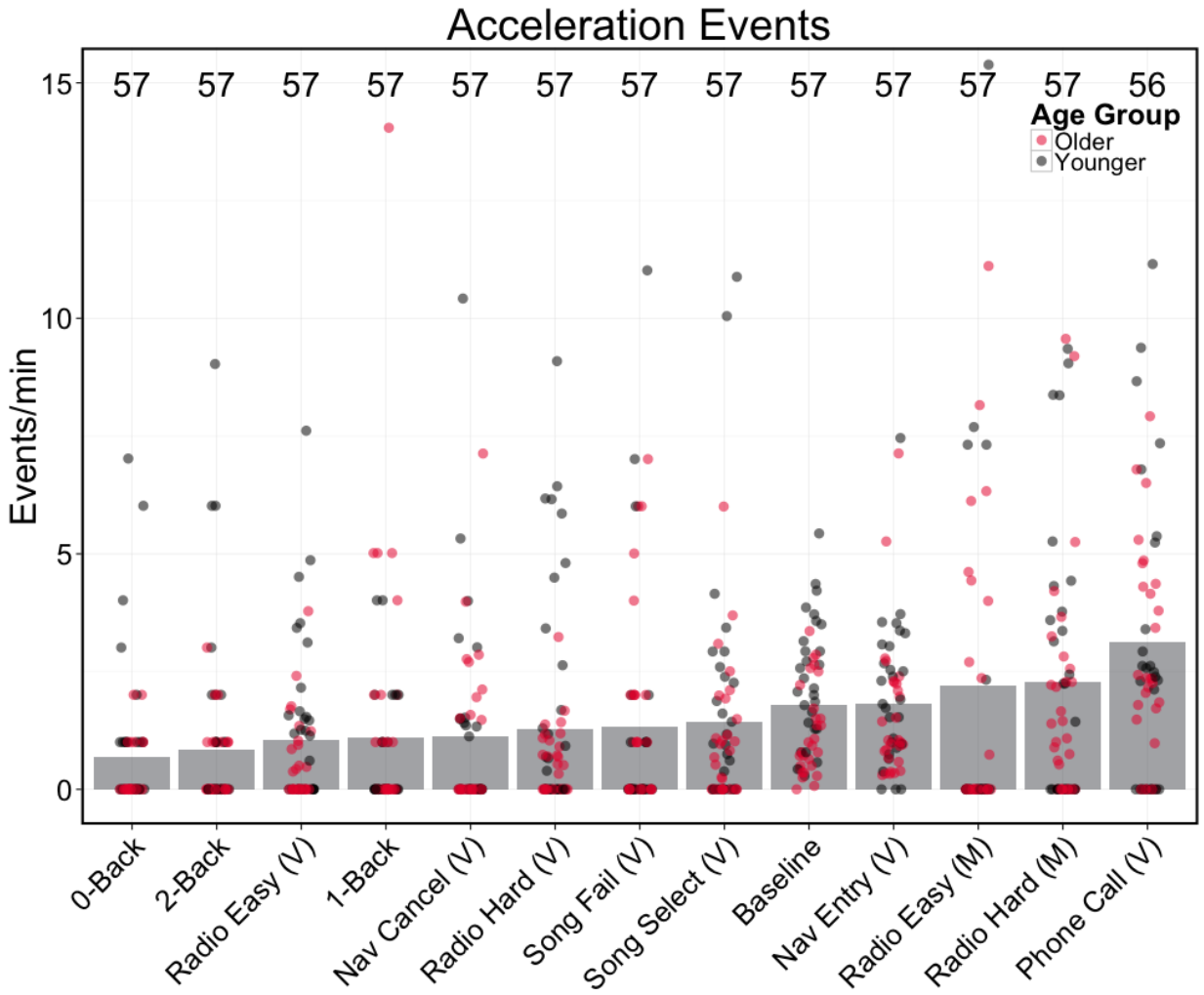


Figure 35: Acceleration events per minute for all task periods.

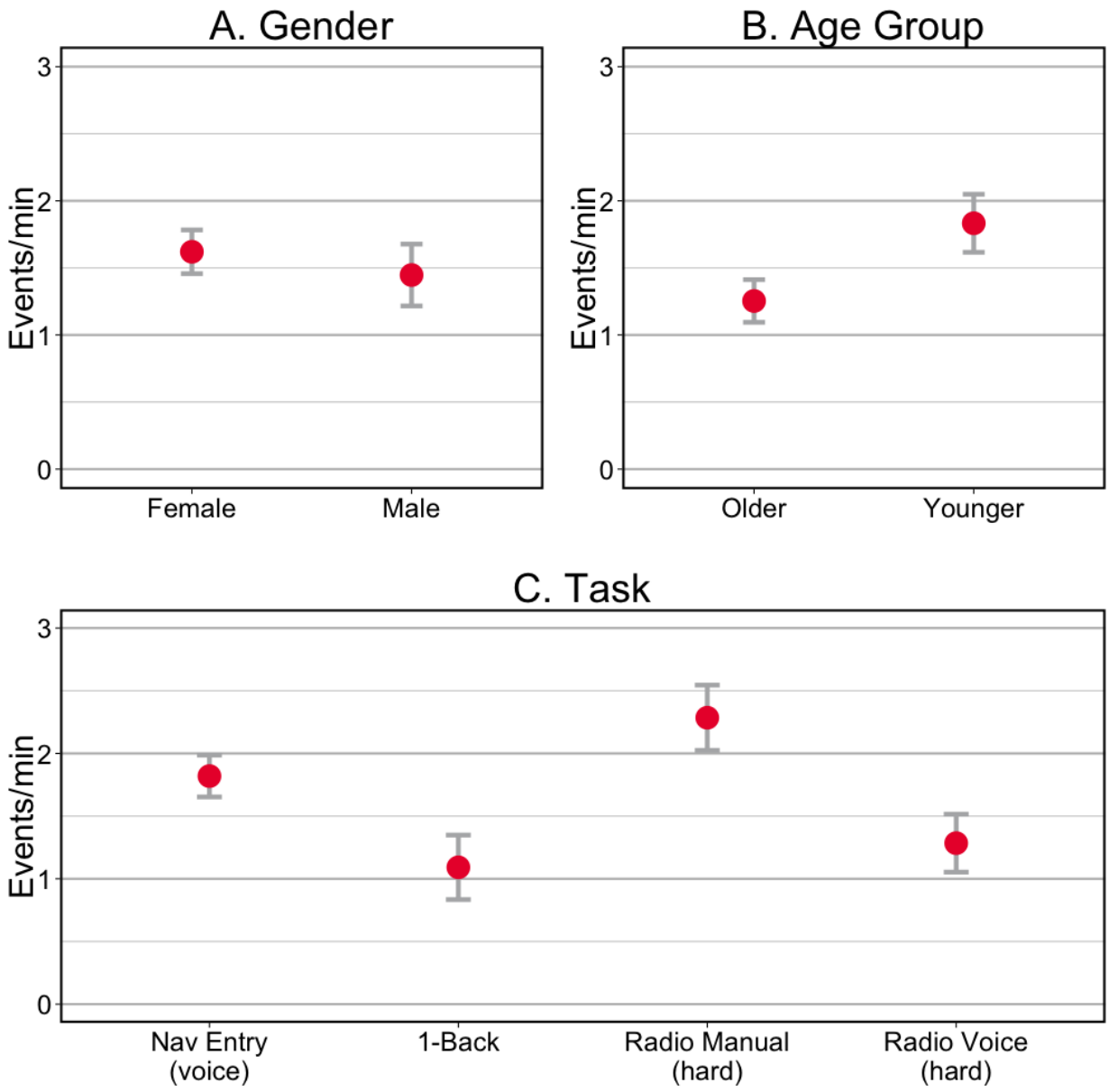


Figure 36: Statistical summary plots for acceleration events.

The rate of acceleration events was not affected by gender (p = .149). However, younger drivers had significantly more acceleration events than older drivers (p = .025). In line with presumed design goals, the frequency of acceleration events was significantly less with the voice-comand control in the Radio Hard task than when using the manual interface (p = .004). The frequency rate for the voice-based Nav Entry task was nominally lower than for the manual Radio Hard task; however, this difference was not statistically significant (p = .589).

Acceleration Events – Extended Detail

Table 16: Mean count (& std. errors) of acceleration events by age group (N=57)

Acceleration	Younger	Older	Event Count
0.10 g	56.07 ± 5.81	36.62 ± 4.63	2632
0.15 g	4.21 ± 0.68	3.41 ± 0.58	217
0.20 g	0.43 ± 0.14	0.48 ± 0.16	26
0.25 g	0.04 ± 0.04	0.10 ± 0.06	4
0.30 g	None	0.03 ± 0.03	1
0.35 g	None	None	None

Table 17: Count of acceleration events ≥ 0.20g by task type

	0.20g	0.25g	0.30g
Baseline	11	3	
Song Select (voice)	3	1	1
Radio voice activation (easy)	1		
Radio voice activation (hard)	1		
Radio manual input (easy)	1		
Radio manual input (hard)	2		
Navigation entry	5		
Navigation cancel	1		
2-Back	1		

Table 16 summarizes unidirectional acceleration events across all participants by age group. No events in exceedance of 0.35g were encountered in the primary analysis portions of the dataset. For a threshold of 0.3g, only one independent acceleration event was found in the dataset. The 0.3g event was a lateral acceleration executed by an older driver during the song selection task. Looking at acceleration events greater than 0.25g, three additional events were observed during the 14 minutes of baseline driving. For a threshold of 0.20g, 26 events were observed in the primary analysis periods of the dataset. Eleven of these events were during baseline driving periods and 15 during periods involving secondary activities. Since the total analysis period durations were quite similar for both baseline and task intervals (baseline periods totaled 14 minutes and the mean total task time across the sample was 14.6 minutes), the data suggest that

engaging in the secondary voice control tasks resulted in only a nominal incidence of moderate 0.20g or greater acceleration events relative to single task (baseline) driving behavior.

Table 18 provides further contextual details for acceleration events in excess of 0.20g. Only one of the four acceleration events exceeding 0.25g was caused by apparent loss of lateral control during task execution. The three remaining events can be attributed to traffic conditions such as merging maneuvers or slow traffic ahead.

Table 18: Details of acceleration events ≥ 0.20g

Acceleration	Direction	Magnitude	Participant	Period	Reason
0.30g	Lateral	0.314g	104 Older Female	Song select	Loss of lateral control while looking at HMI during task
0.25g	Longitudinal	-0.262g	7 Younger Male	Baseline	Braking during merge at lane end
	Longitudinal	-0.248g	74 Older Male	Baseline	Braking due to traffic jam ahead
	Longitudinal	-0.276g	77 Older Female	Baseline	Braking due to merging traffic
0.20g	Lateral	0.202g	7	Navigation Entry	Loss of lateral control while looking at HMI during task
	Lateral	0.221g	7	Baseline	Loss of lateral control
	Lateral	-0.212g	25 Younger Male	Navigation Entry	Loss of lateral control while looking at HMI during task
	Lateral	-0.219g	27 Younger Female	Radio voice (hard)	Abrupt lane change
	Lateral	0.236g	29 Younger Male	Baseline	Avoiding merging traffic
	Longitudinal	-0.176g	41 Younger Male	Baseline	Braking due to traffic ahead
	Lateral	-0.212g	49	Navigation	Abrupt lane change

		Younger Female	Cancel	
N/A	N/A	49	Baseline	Damaged road surface
N/A	N/A	60 Younger Female	Baseline	Damaged road surface
N/A	N/A	62 Older Male	Navigation entry	Damaged road surface
Longitudinal	-0.201g	67 Older Male	Radio Manual (hard)	Braking due to merging traffic arriving at traffic jam
Lateral	0.219g	77	Radio Manual (hard)	Loss of lateral control while looking at HMI during task
Lateral	0.236g	77	Radio Voice (easy)	Loss of lateral control while looking at HMI during task
Longitudinal	-0.205g	79 Older Male	Song Select	Braking due to merging traffic
Longitudinal	-0.212g	84 Younger Female	Navigation Entry	Braking due to traffic jam ahead
Lateral	-0.202g	89 Older Female	Baseline	Loss of lateral control
Longitudinal	-0.234g	89	Song Select	Braking due to traffic jam ahead
Lateral	-0.217g	91 Older Male	Radio Manual (easy)	Loss of lateral control while looking at HMI during task
Lateral	0.216g	94 Younger Male	2-Back	Loss of lateral control during task
Lateral	-0.207g	99 Older Female	Navigation Entry	Damaged road surface
Longitudinal	-0.212g	104	Baseline	Braking due to lane change
Lateral	-0.212g	104	Baseline	Loss of lateral control

Steering Wheel Angle

Increases in the variability in steering wheel angle is commonly related to reduced lateral control and associated with the driver's need to concurrently manage the additional workload of secondary activates (Östlund et al., 2004). Under normal driving conditions, small steering wheel corrections are made to adjust the vehicle heading for variations in roadway conditions (Liu, Schreiner, & Dinges, 1999). These variations can be looked at using a number of different methods including the standard deviation of wheel angle, and counts of minor wheel reversals and major wheel reversals. In situations of increased cognitive workload, the number of small steering wheel adjustments tend to increase, while secondary activates that involve visual attention demands often impact large reversal (Östlund, et al., 2005).

A Friedman test of steering wheel angle during secondary task performance shows a significant effect of task ($X^2(12) = 335.8$, $p < .001$). Arranging the task periods in ascending order for variability of velocity (see Figure 37), shows a clear reduction in variability during all task periods, with the exception of the navigation entry task, relative to baseline.

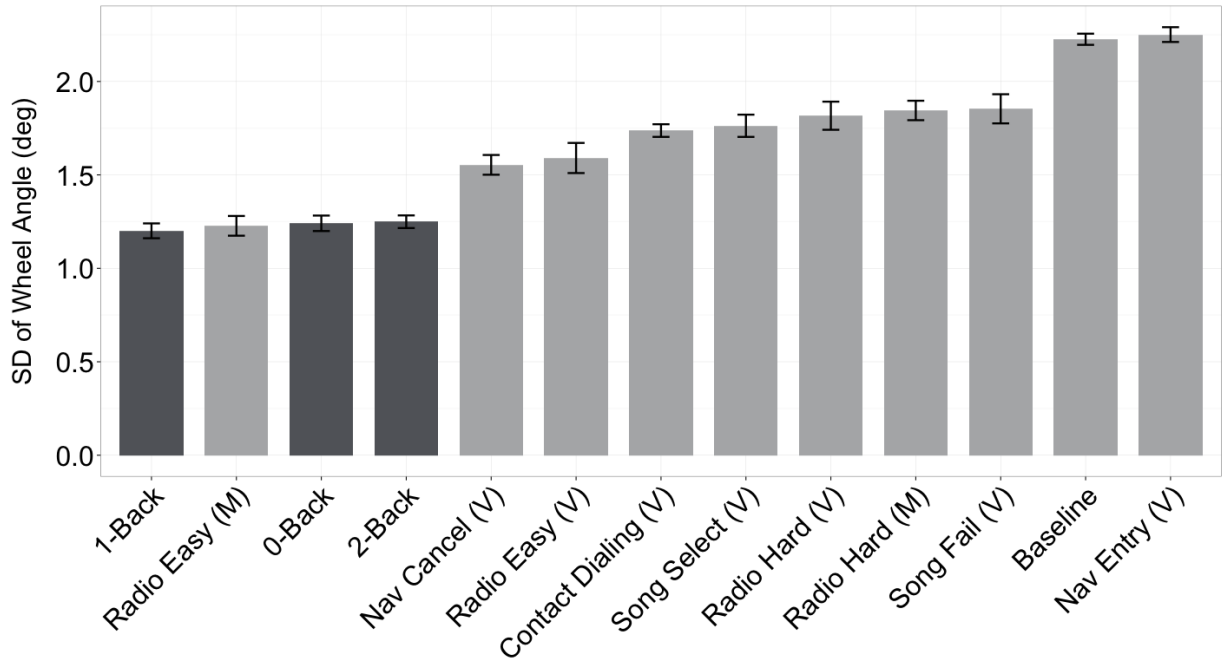


Figure 37: Tasks listed in ascending order for variability of steering wheel angle. The baseline shown represents the combined baseline periods recorded prior to each task. Error bars represent 1 SEM.

Table 19: Means (and standard deviations) of steering wheel angle.

Task	Older	Younger	(All)
Baseline	2.15 (0.6)	2.30 (0.6)	2.23 (0.6)
Nav Cancel	1.56 (0.6)	1.55 (0.6)	1.55 (0.6)
Nav Entry	2.25 (0.4)	2.25 (0.6)	2.25 (0.5)
0-Back	1.19 (0.4)	1.30 (0.5)	1.24 (0.5)
1-Back	1.15 (0.5)	1.25 (0.5)	1.20 (0.5)
2-Back	1.24 (0.4)	1.26 (0.5)	1.25 (0.4)
Phone	1.85 (0.6)	1.70 (0.5)	1.77 (0.5)
Radio Manual Easy	1.43 (0.7)	1.03 (0.5)	1.23 (0.6)
Radio Manual Hard	1.95 (0.5)	1.74 (0.6)	1.84 (0.6)
Radio Voice Easy	1.54 (0.5)	1.64 (1.2)	1.59 (0.9)
Radio Voice Hard	1.88 (1.1)	1.76 (0.6)	1.82 (0.9)
Song Fail	1.90 (0.7)	1.81 (0.6)	1.85 (0.7)
Song Select	1.95 (0.7)	1.58 (0.6)	1.76 (0.7)

As detailed below, there was no overall main effect of age group on steering wheel angle variability.

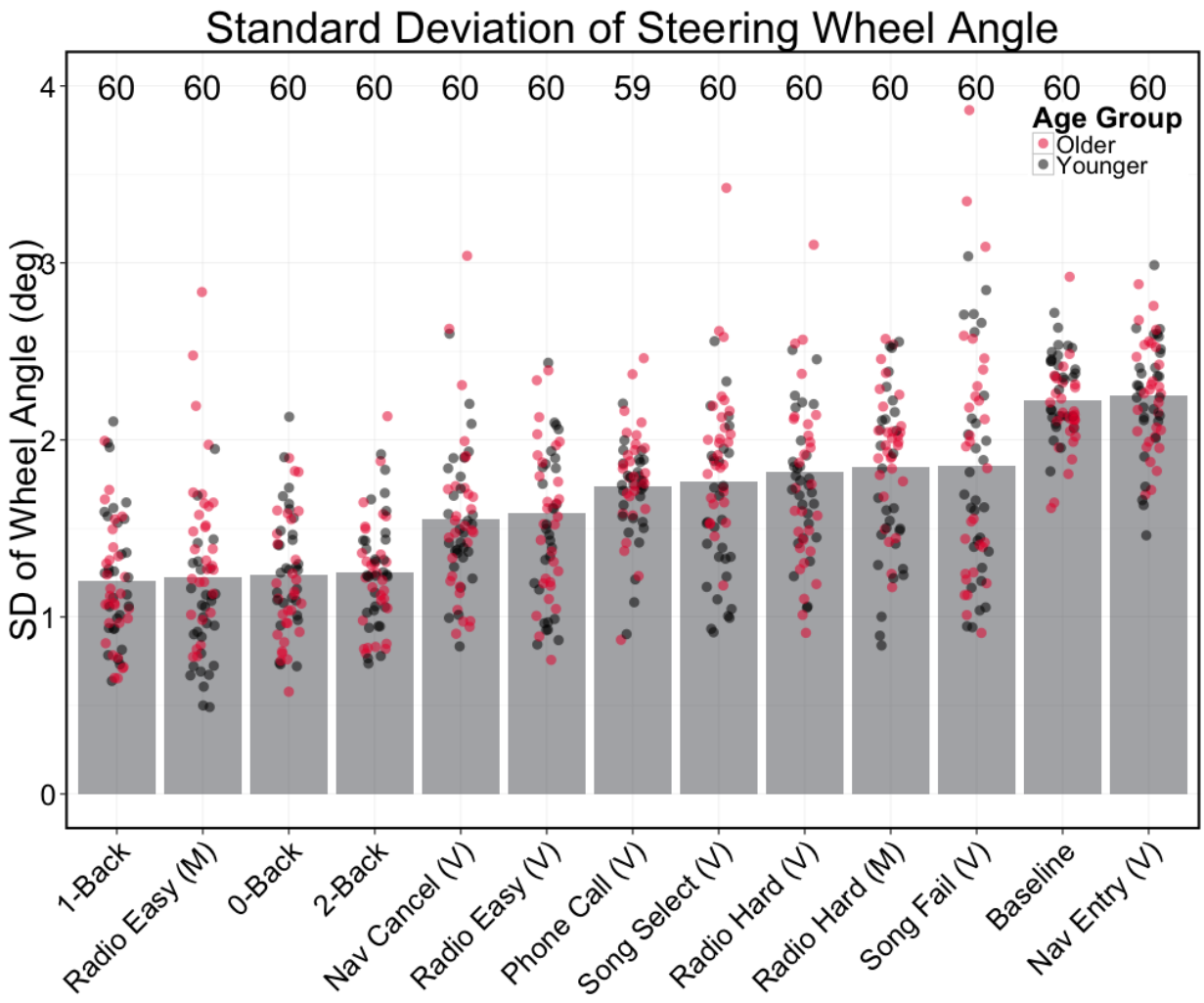


Figure 38: Standard deviation (SD) of steering wheel angle for all task periods. Note that two data points are cut off from the plot for display purposes.

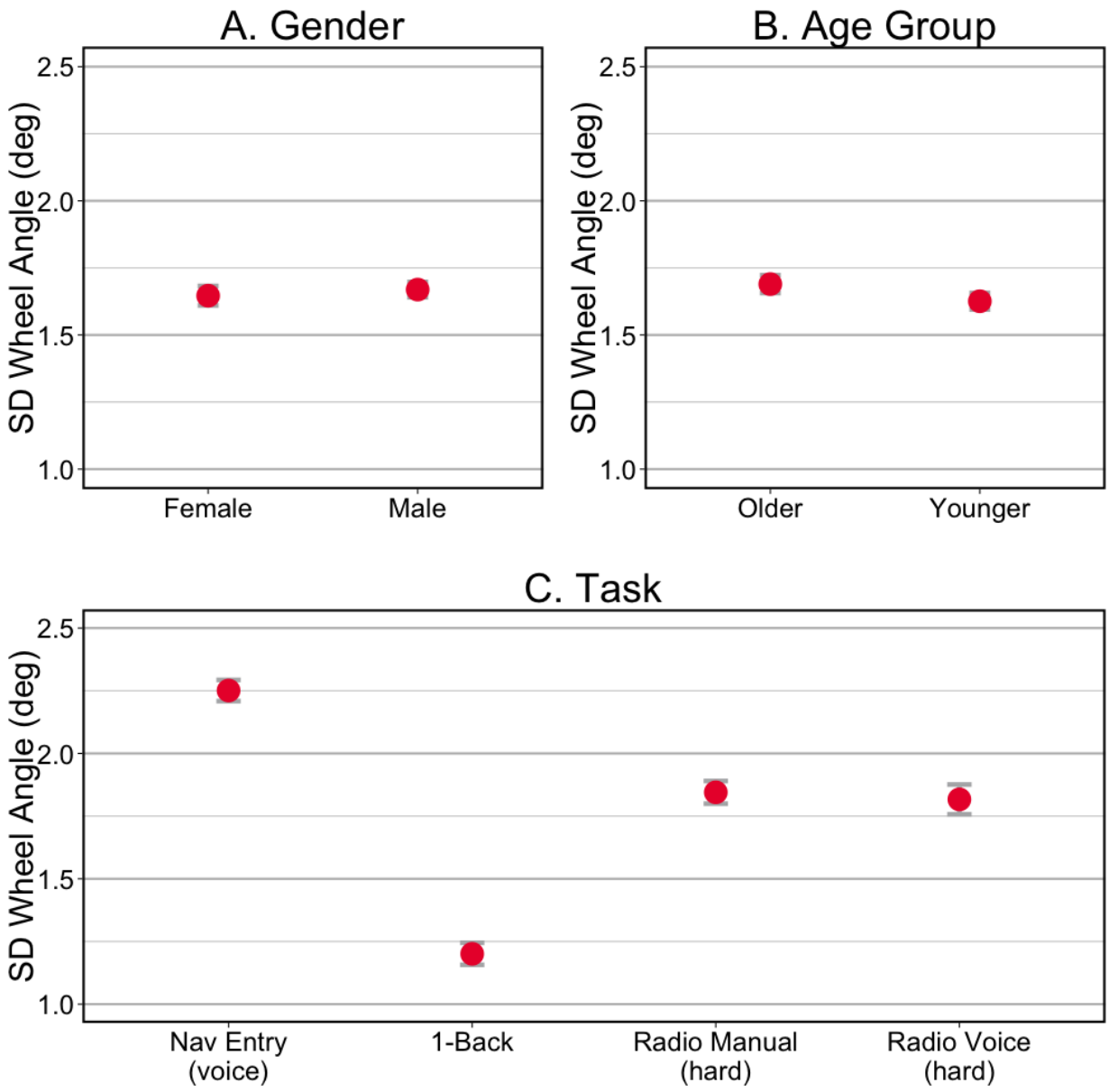


Figure 39: Statistical summary plots for SD of steering wheel angle.

Variability in steering wheel angle was not affected by gender or age group ($p = .582$ and $p = .159$, respectively). Standard deviation of steering wheel angle did not differ significantly between the manual Radio Hard task and the voice-command version of the Radio Hard task ($p = .223$). However, the variability in steering wheel angle was higher during the voice-command based Nav Entry task vs. the manual Radio Hard task ($p < .001$).

Minor Steering Wheel Reversals

A Friedman test of minor steering wheel reversal rate during secondary task performance shows a significant effect of task ($X^2(12) = 109.4$, $p < .001$).

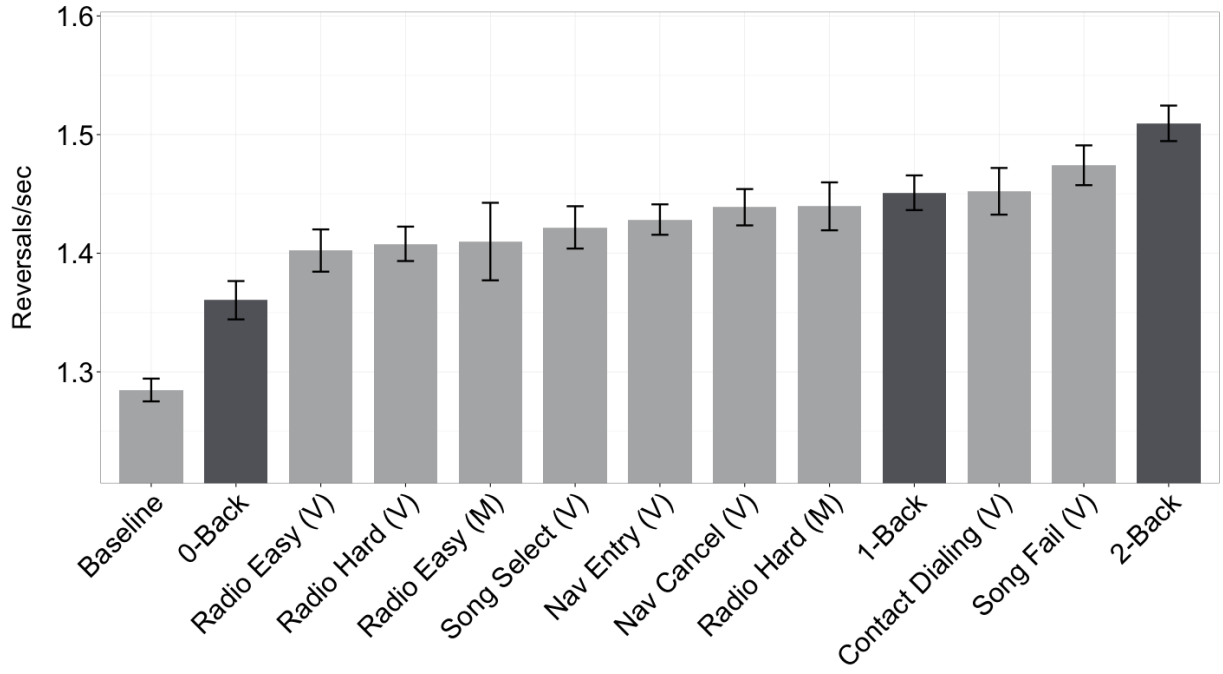


Figure 40: Tasks listed in ascending order for minor steering wheel reversal rate. The baseline shown represents the combined baseline periods recorded prior to each task. Error bars represent 1 SEM.

As can be observed in Figure 40 above, minor wheel reversals increased during the task periods relative to the single task driving period (baseline). The three levels of the auditory-vocal calibration task (n-back) appear to scale well along this metric, with steering wheel reversals increasing with each step in cognitive demand. This clear ordering across demand levels is more consistent than what was seen in our previous on-road study (Reimer, Mehler, Wang, et al., 2012) that employed single instances of 2 minute long n-back periods for each demand level. Since this study averages across two instances of 30 second n-back periods for each demand level, it is possible that is better estimation of the underlying behavior pattern.

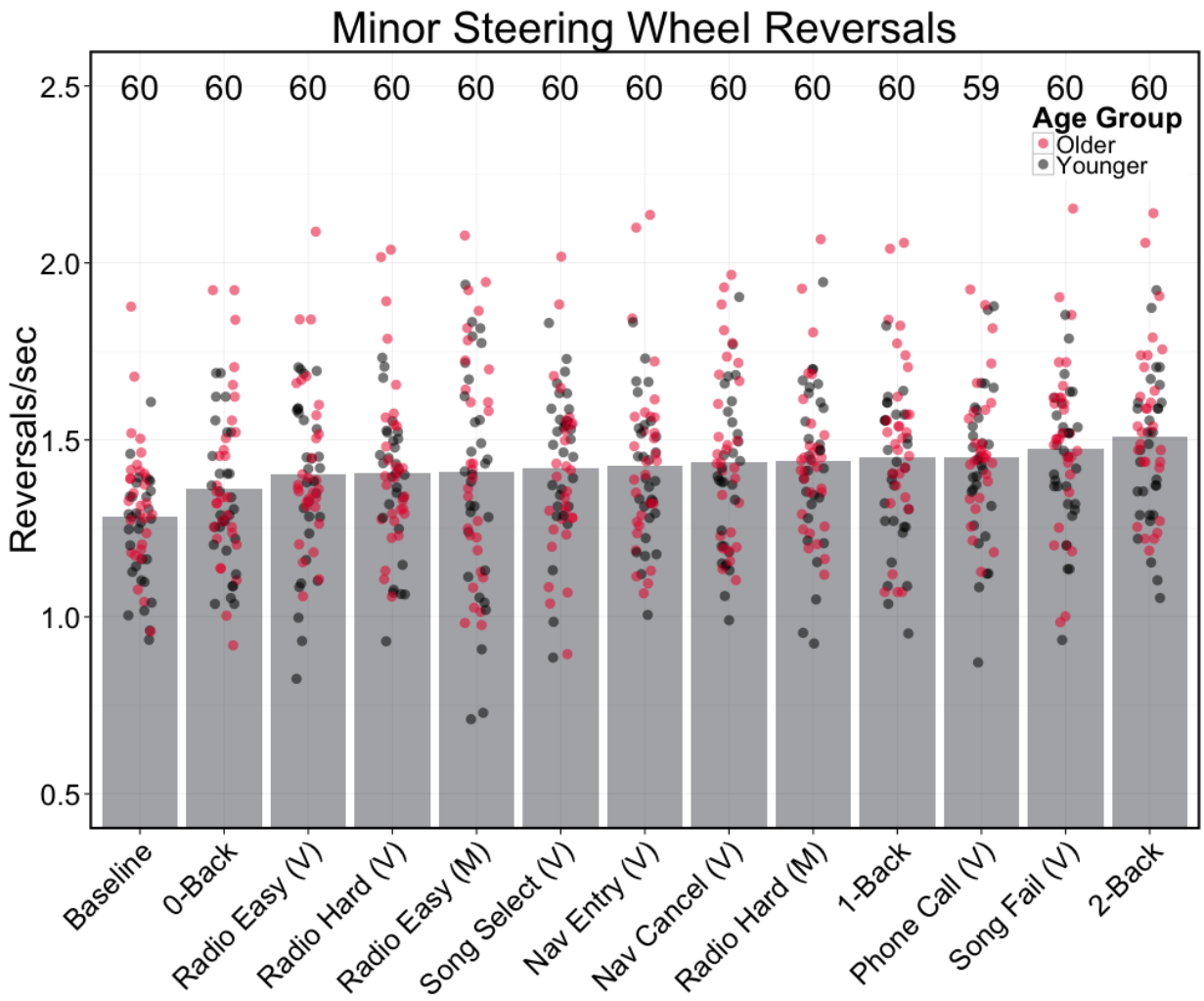


Figure 41: Number of minor steering wheel reversals (SWRs) per second for each task.

Minor SWRs were counted and classified using a 0.1° gap size (see Methods for details). As detailed at the end of this section, there was no overall statistically significant effect of age group on major steering wheel reversal rates across the sample as a whole.

Table 20: Means (and standard deviations) of minor wheel reversal rate.

Task	Older	Younger	(All)
Baseline	1.33 (0.2)	1.24 (0.2)	1.29 (0.2)
Nav Cancel	1.48 (0.3)	1.40 (0.3)	1.44 (0.3)
Nav Entry	1.45 (0.3)	1.41 (0.2)	1.43 (0.2)
0-Back	1.39 (0.3)	1.33 (0.2)	1.36 (0.2)
1-Back	1.51 (0.3)	1.39 (0.2)	1.45 (0.3)
2-Back	1.55 (0.3)	1.47 (0.2)	1.51 (0.3)
Phone	1.47 (0.2)	1.42 (0.2)	1.44 (0.2)
Radio Manual Easy	1.44 (0.4)	1.38 (0.4)	1.41 (0.4)
Radio Manual Hard	1.45 (0.2)	1.43 (0.3)	1.44 (0.3)
Radio Voice Easy	1.43 (0.3)	1.37 (0.3)	1.40 (0.3)
Radio Voice Hard	1.44 (0.3)	1.38 (0.2)	1.41 (0.2)
Song Fail	1.50 (0.2)	1.45 (0.2)	1.47 (0.2)
Song Select	1.40 (0.3)	1.45 (0.2)	1.42 (0.2)

As detailed at the end of this section, there was no overall statistically significant effect of age group on minor steering wheel reversal rates across the sample as a whole.

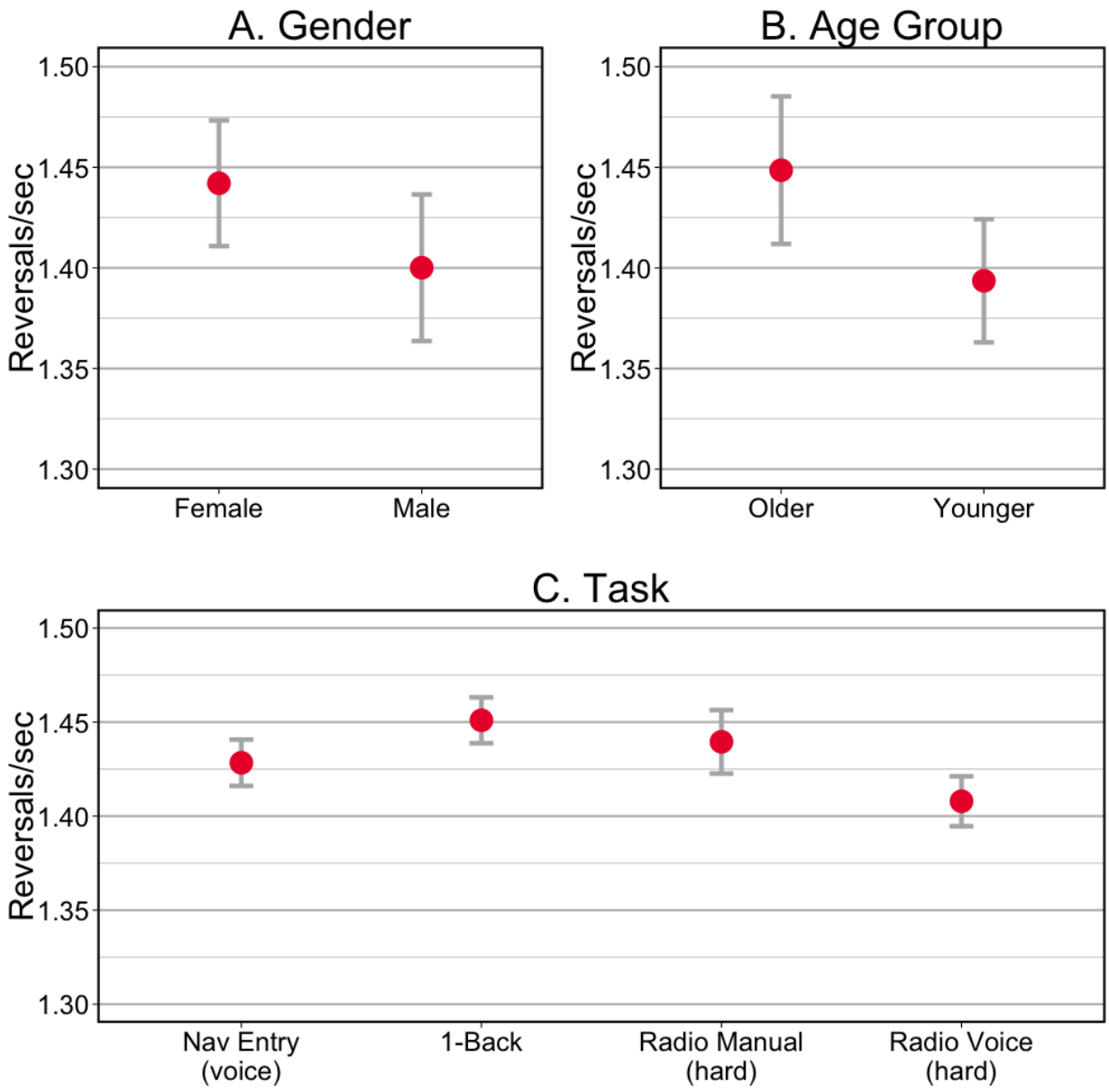


Figure 42: Statistical summary plots for minor SWR.

There were no significant differences by gender or age group on the minor steering wheel reversal rate measure ($p = .406$ and $p = .398$, respectively). The minor SWR did not differ significantly between the manual Radio Hard tuning task and the voice Radio Hard task or the voice Nav Entry task ($p = .274$ and $p = .777$, respectively). Thus, to the extent that the manual Radio Hard task is used as a reference point for vehicle control, both tasks compare favorably on this vehicle control metric.

Major Steering Wheel Reversals

A Friedman test of major steering wheel reversal rate during secondary task performance shows a significant effect of task ($X^2(12) = 222.4$, $p < .001$).

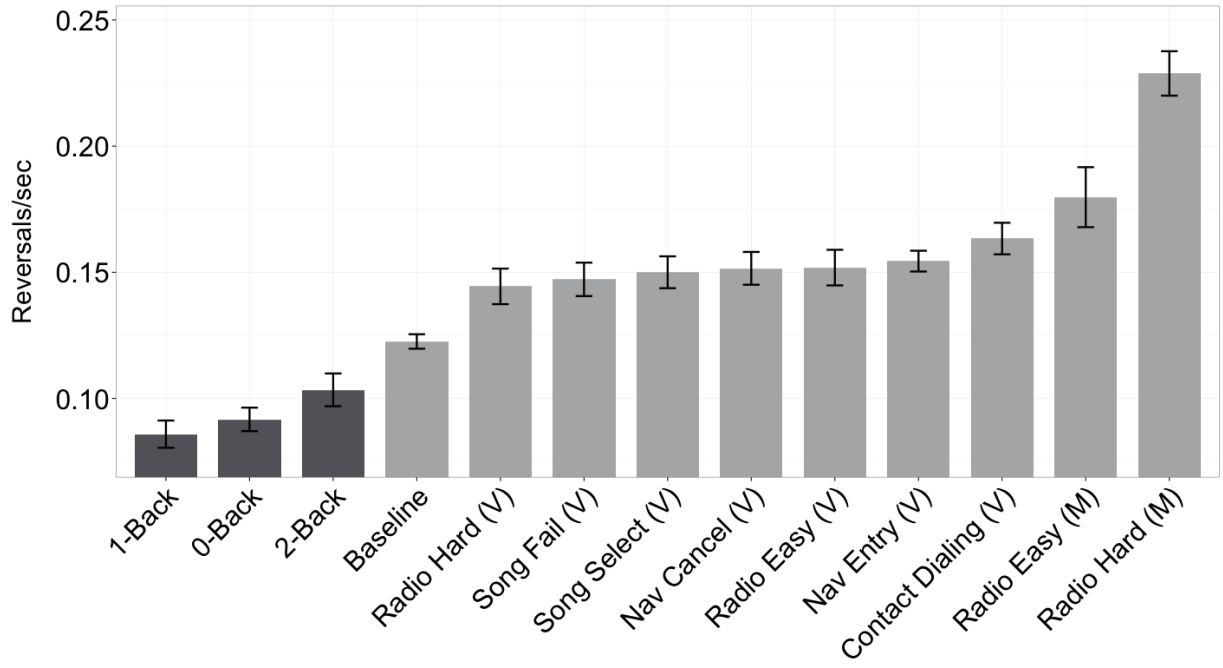


Figure 43: Tasks listed in ascending order for major steering wheel reversal rate. The baseline shown represents the combined baseline periods recorded prior to each task. Error bars represent 1 SEM.

All of the in-vehicle system tasks show a higher major steering wheel reversal rate than single task baseline driving, with the manual Radio Hard task showing the highest value. As detailed at the end of this section, using the manual Radio Hard task is associated with a significantly higher major steering wheel reversal rate than using the voice interface to complete the Radio Hard task. Thus, in assessing the voice interface along this dimension of vehicle control, it could be seen as having advantages over the traditional manual control interface. Interestingly, all three levels of the auditory-vocal n-back task are associated with a lower major steering wheel reversal rate than single task (baseline) driving.

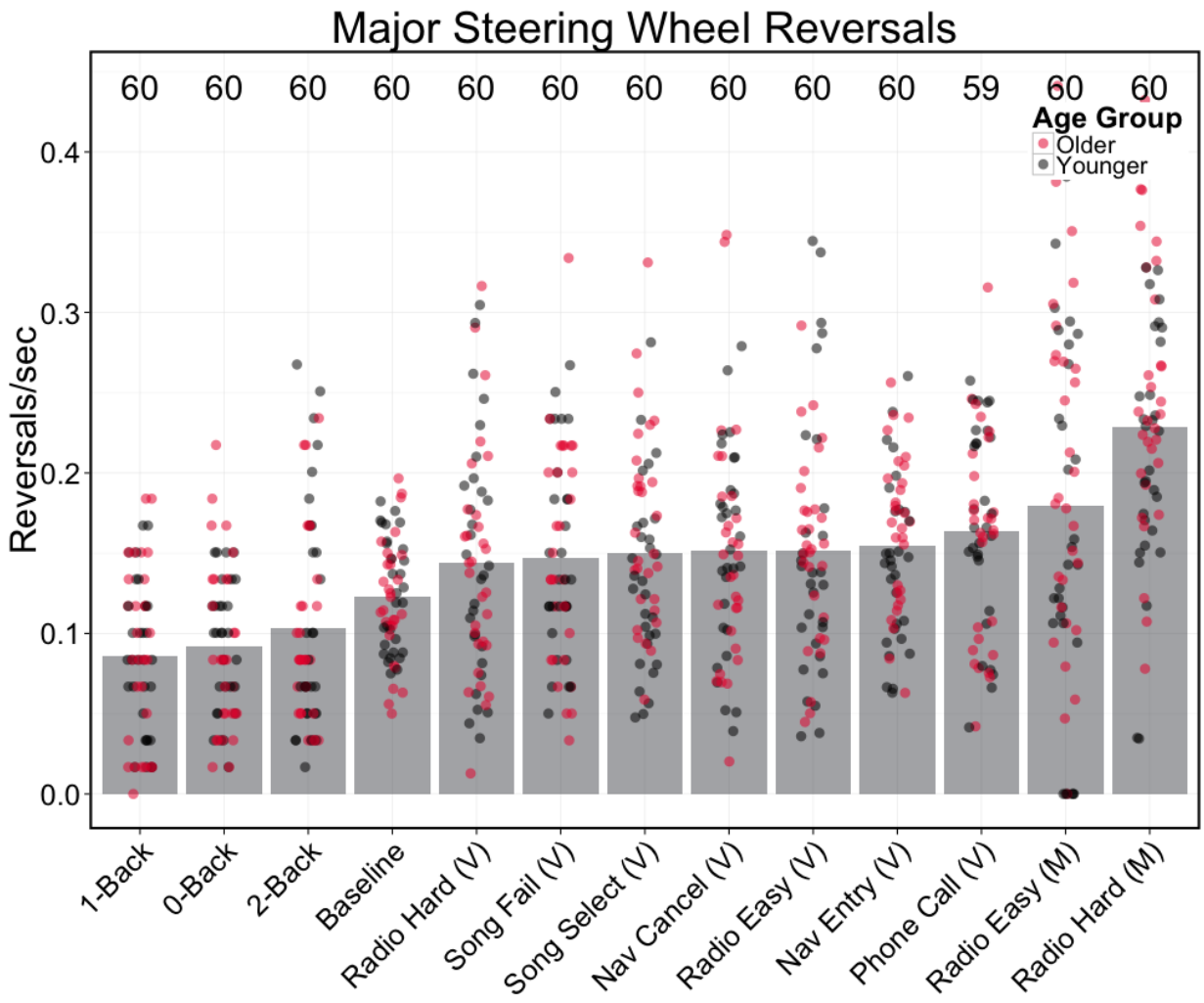


Figure 44: Number of major steering wheel reversals (SWRs) per second for each task.

Major SWRs were counted and classified using a 3° gap size (see Methods for details). As detailed at the end of this section, there was no overall statistically significant effect of age group on major steering wheel reversal rates across the sample as a whole.

Table 21: Means (and standard deviations) of major steering wheel reversal rates.

Task	Older	Younger	(All)
Baseline	0.12 (0.0)	0.12 (0.0)	0.12 (0.0)
Nav Cancel	0.15 (0.1)	0.15 (0.1)	0.15 (0.1)
Nav Entry	0.16 (0.1)	0.14 (0.1)	0.15 (0.1)
0-Back	0.09 (0.1)	0.09 (0.1)	0.09 (0.1)
1-Back	0.09 (0.1)	0.09 (0.1)	0.09 (0.1)
2-Back	0.10 (0.1)	0.11 (0.1)	0.10 (0.1)
Phone	0.17 (0.1)	0.16 (0.1)	0.17 (0.1)
Radio Manual Easy	0.20 (0.2)	0.16 (0.2)	0.18 (0.2)
Radio Manual Hard	0.25 (0.1)	0.21 (0.1)	0.23 (0.1)
Radio Voice Easy	0.15 (0.1)	0.15 (0.1)	0.15 (0.1)
Radio Voice Hard	0.14 (0.1)	0.15 (0.1)	0.14 (0.1)
Song Fail	0.15 (0.1)	0.15 (0.1)	0.15 (0.1)
Song Select	0.17 (0.1)	0.13 (0.1)	0.15 (0.1)

As detailed on the next page, there was no significant statistically significant effect of age group on the major steering wheel reversal rate measure across the sample as a whole.

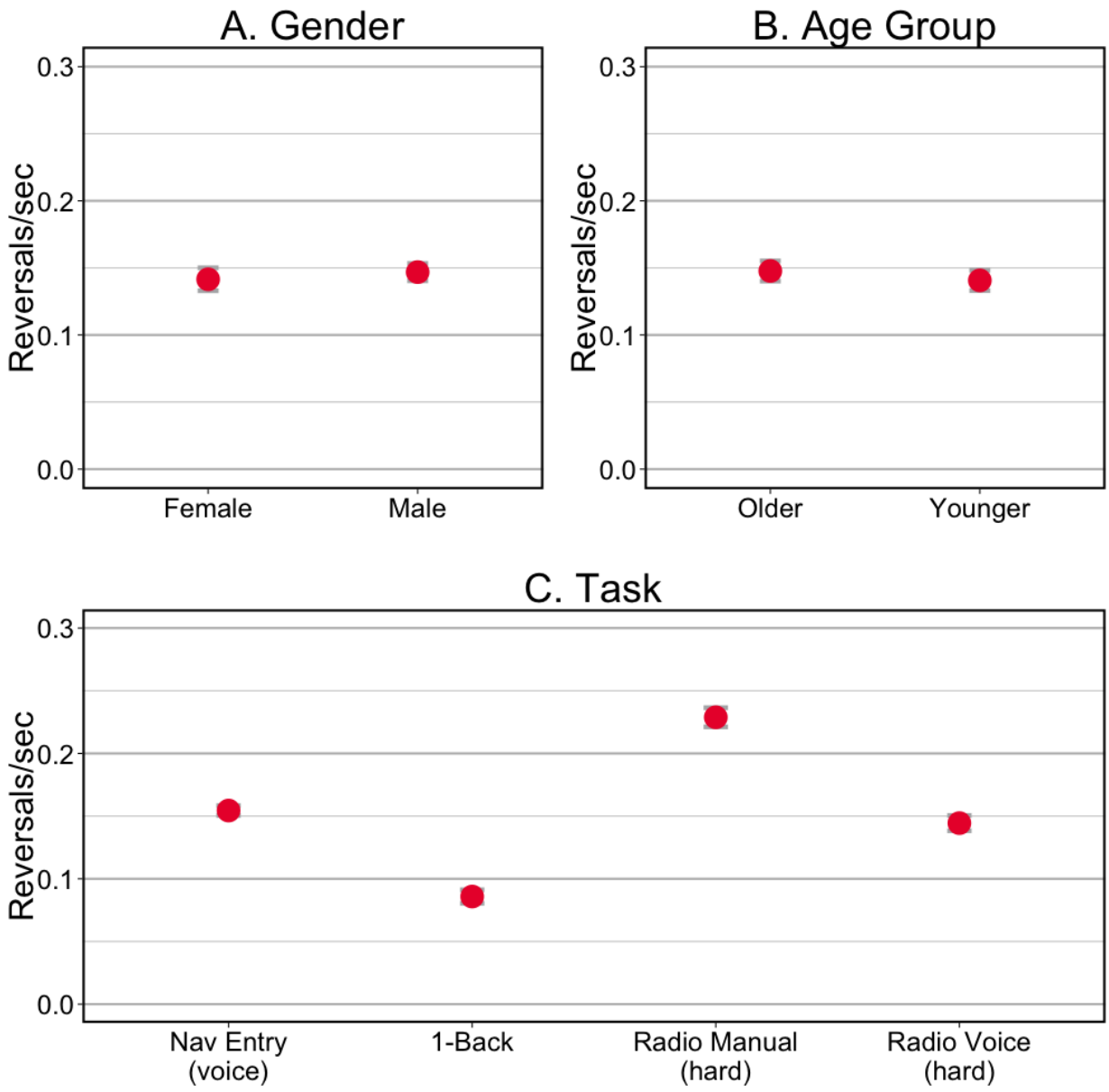


Figure 45: Statistical summary plots for major SWR.

There were no significant differences by gender or age group on the major steering wheel reversal rate metric, ($p = .612$ and $p = .581$, respectively). The major SWR rate for voice control of the Radio Hard task and the voice-command Nav Entry task were significantly lower than the manual Radio Hard task ($p < .001$ for both). Thus, to the extent that the manual Radio Hard task is used as a reference point for vehicle control, both tasks compare favorably on this vehicle control metric.

Glance Analyses

Manual Glance Coding

A significant portion of the data from the instrumented vehicle's automated eye-tracker proved unreliable, and as a result, we opted to code glance data manually (see Methods: Data Reduction & Analysis and Appendix F). Glances for 7 of the 60 participants could not be reliably coded, resulting in an analysis sample of 53 cases. Of the 7 participant cases that were not coded, 5 were due to persistent problems with the ambient lighting (video too bright or dark), 1 participant was not coded because his/her eyes were out of the video frame, and 1 participant was not coded because his/her height resulted in a high frequency of ambiguous glances.

Glance Measures & Off-Road Glance Metrics

As discussed in the methods section, traditional automotive device assessment has generally considered visual demand in terms of glances made to the device / interface under study. In developing criteria and evaluation procedures to aid automotive and telematics manufacturers during the product development process, the Alliance of Automobile Manufacturers utilized this approach (Driver Focus-Telematics Working Group, 2006). In specific, The Alliance guidelines (criterion 2.1 A) specify that a visual or visual-manual task intended for use by a driver while the vehicle is in motion should be designed such that: 1) single glance durations generally should not exceed 2 seconds; and b) task completion should require no more than 20 seconds of total glance time to task display(s) and controls (p. 39). NHTSA recently released visual-manual distraction guidelines for in-vehicle electronic devices (National Highway Traffic Safety Administration, 2013) in which they specify assessing several aspects of glance behavior in terms of off-road glance metrics as opposed to quantifying glances to the device. In a recent simulation study considering visual interaction with two nomadic devices, we compared the approaches to calculating glance metrics and found that the two approaches produced generally parallel patterning of the data but with definite effects on absolute values (Dopart et al., 2013 in press). Based on our experience with that dataset, we believe that it may be useful in the ongoing review and development of distraction criteria to further develop data that allows for a comparison of these two approaches to considering glance behavior. For purposes of the current report, we present glance measures in terms of off-road glance metrics following the current NHTSA recommendations. Alternate analyses employing the glance to device approach are presented in Appendix A.

Since the Manual Radio tasks are classic visual-manual device interactions, it is appropriate to evaluate them in terms of visual demand. As part of this assessment, we have taken the approach of looking at the dataset using metrics recommended in the new NHTSA visual-manual guidelines, i.e. mean glance duration, percentage of glances greater than 2 seconds in duration, and total off-road glance time. We also consider the number of glances associated with each task period, and look at the distribution of glances as a function of age, gender, and task type.

In addition, in presenting these variables, we have added reference points on relevant graphs showing how this analysis sample would fair if the NHTSA criteria for each of these metrics were applied. It should be emphasized that some aspects of our methodology and our sample do not conform fully to NHTSA's guidelines for age distribution since this study was initiated prior to the release of the guidelines. Second, the guidelines assume that sampling is carried out in a simulator while our assessment was carried out during actual highway driving. (NHTSA quite reasonably makes the argument that testing in a simulator is specified since a device should not be tested under actual driving conditions if it has not been established that it is relatively safe to do so.) There is some question as the extent to which metrics collected in the simulator and field correspond, although previous work in our own work has shown a fairly close correspondence when comparing the distribution pattern of glances across different device interfaces in the vehicle (Wang et al., 2010).

To meet its guidelines, NHTSA specifies a minimum sample size of 24 participants (with specified age, gender, and experience characteristics) and mandates that at least 21 out of the 24 participants meet each of the following criteria while performing the "testable task" one time (see p. 272):

- **Percentage of Long Duration Glances**. No more than 15 percent (rounded up) of the total number of eye glances away from the forward road scene have durations of greater than 2.0 seconds.
- **Mean Off-Road Glance Duration**. The mean duration of all eye glances away from the forward road scene is less than or equal to 2.0 seconds.
- **Total Off-Road Glance Time**. The sum of the durations of each individual participant's eye glances away from the forward road scene is less than or equal to 12.0 seconds.

For samples larger than 24, the same proportional relationship is to be applied such that 85% (rounded up) or more of the participants meet the criteria. Note that NHTSA defines "off-road"

as any glance off of the forward roadway, which classifies glances to the rear- and side-view mirrors as "off-road". We adhere to this definition in the following analyses. In the interest of improving the data's reliability, we had each participant perform each task twice, and have averaged their performance across the two trials. We have also produced a series of parallel analyses and summaries that examine subsets of the data:

- A "to device" analysis that computes the glance metrics based only on glances to the device, rather than any glance off the forward roadway (Appendix A).
- An "error-free" analysis that considers only task trials that were successfully completed without error or assistance (Appendix B).
- A "trial comparison" analysis that examines the two trials of each task separately (Appendix C).

Selected Glance Metrics Summary Table (Off-Road Glance Analysis)

If one were to apply the NHTSA distraction cutpoints to younger, older, and overall cohorts, **Table 22** below shows the percentage who would meet each of the off-the-forward-roadway glance criteria. Entries for situations where less than 85% of a group meet a threshold are bolded and shown in red.

Task	Age Group	Long Duration Glances	Mean Glance Duration	Total Off-road Glance Time
Nav Cancel	Younger	96.70%	96.70%	100.00%
	Older	100.00%	100.00%	91.30%
	(all)	98.10%	98.10%	96.23%
Nav Entry	Younger	100.00%	100.00%	**13.33%**
	Older	100.00%	100.00%	**0.00%**
	(all)	100.00%	100.00%	**7.55%**
Radio Manual Easy	Younger	90.00%	100.00%	100.00%
	Older	**82.60%**	100.00%	86.96%
	(all)	86.80%	100.00%	94.34%
Radio Manual Hard	Younger	96.70%	100.00%	**73.33%**
	Older	87.00%	100.00%	**8.70%**
	(all)	92.50%	100.00%	**45.28%**
Radio Voice Easy	Younger	100.00%	100.00%	100.00%
	Older	100.00%	100.00%	**78.26%**
	(all)	100.00%	100.00%	90.57%
Radio Voice Hard	Younger	100.00%	100.00%	90.00%
	Older	100.00%	100.00%	**65.22%**
	(all)	100.00%	100.00%	**79.25%**
Song Select	Younger	96.70%	100.00%	86.67%
	Older	100.00%	100.00%	**47.83%**
	(all)	98.10%	100.00%	**69.81%**
Song Fail	Younger	93.30%	100.00%	**26.67%**
	Older	100.00%	100.00%	**52.17%**
	(all)	96.20%	100.00%	**37.74%**
Phone	Younger	100.00%	100.00%	96.67%
	Older	100.00%	100.00%	**63.64%**
	(all)	100.00%	100.00%	**82.69%**

Mean Off-Road Glance Duration

This metric considers the mean duration of all eye glances away from the forward road scene. Following NHTSA's (2013) definition of this measure, glances to the rear and side mirrors are coded as glances away from the forward road scene. A Friedman test of mean off-road glance time during secondary task performance shows a significant effect of task ($X^2(12) = 330.4$, $p < .001$).

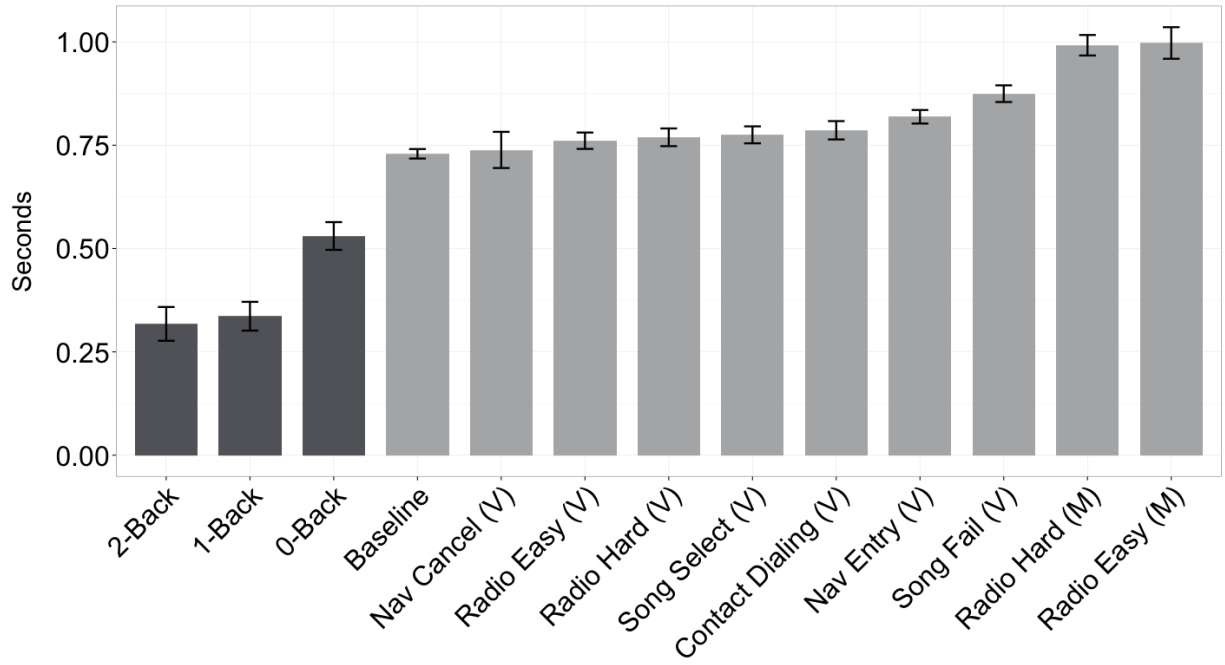


Figure 46: Tasks listed in ascending order for mean off-road glance time. The baseline shown represents the combined baseline periods recorded prior to each task. Error bars represent 1 SEM.

Consistent with previous findings that pure auditory-verbal cognitive tasks are associated with a concentration of gaze toward the forward roadway (Reimer, Mehler, Wang, et al., 2012), mean off-road glance time for the three levels of the n-back task were all notably lower than single task (baseline) driving. Mean off-road glance time was highest for the two classic visual-manual radio tasks.

As noted elsewhere, this study was not designed as a test of the test-vehicle user interface under either The Alliance or the new NHTSA visual-manual distraction guidelines. Rather, we are considering here how the different interface tasks studied compare to each other if the new NHTSA criteria were applied to them. The primary figures for Mean Off-Road Glance Time, Long Glance Rate, and Total Off-Road Glance Time use a horizontal dashed line to represent the critical threshold established in the NHTSA visual-manual distraction guidelines. The short horizontal bars above each task category represent the point at which 85% of the participants fall. If NHTSA testing conditions were employed, the bar representing this 85% level must fall at or below the dashed line for an age and gender compliant sample to meet NHTSA visual-manual guidelines.

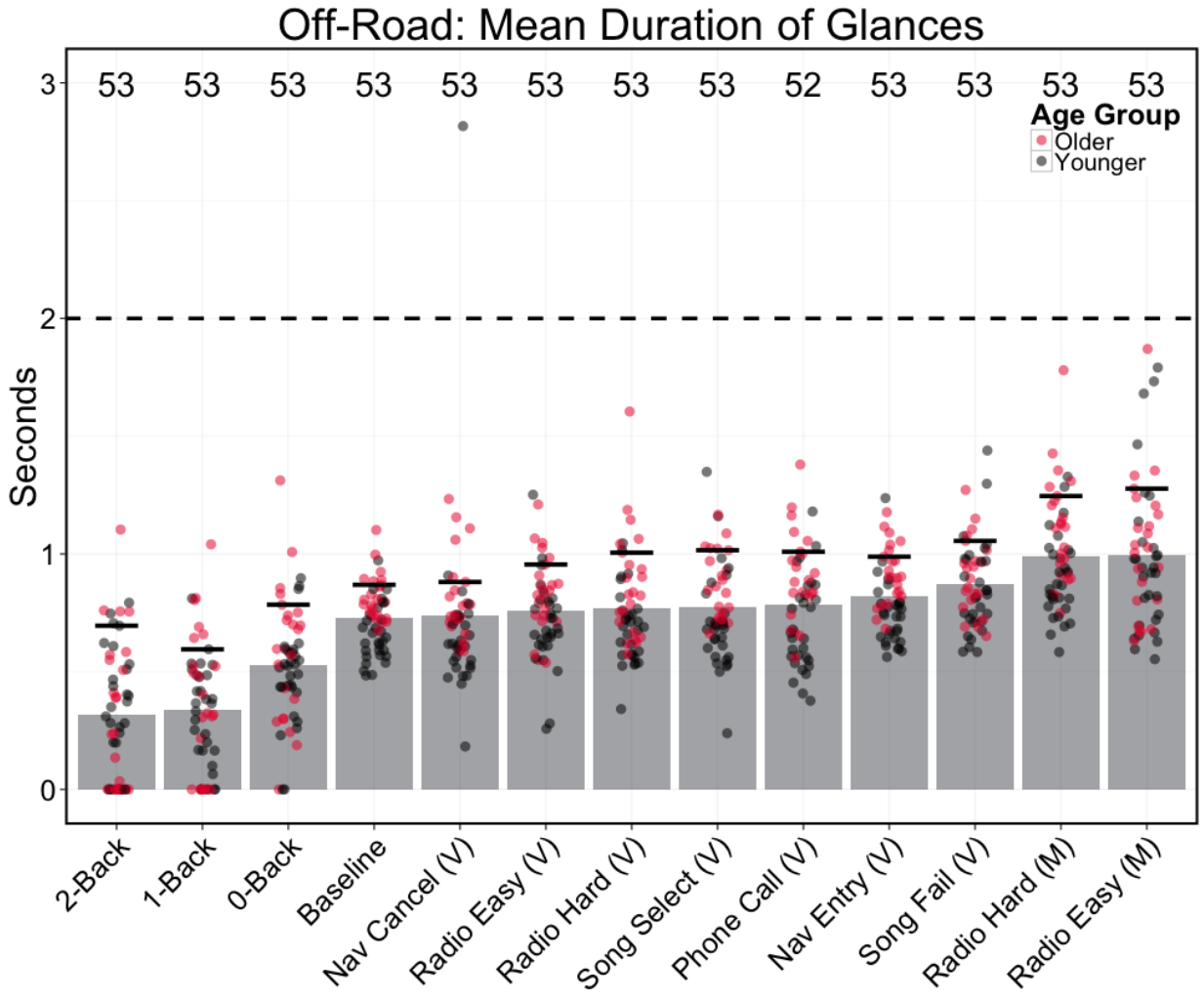


Figure 47: Mean glance time off the forward roadway for each participant.

Participants almost always maintained a mean glance duration of under 2.0 seconds when performing these secondary tasks. Based on the data collected in this study, it appears highly likely that a NHTSA age compliant sample would meet the guidelines for mean duration of off-road glances for each of the interactions studied in the test-vehicle.

Table 23: Means (and standard deviations) for mean glance time.

Task	Younger	Older	(All)
Baseline	0.66 (0.1)	0.82 (0.1)	0.73 (0.1)
Nav Cancel	0.69 (0.4)	0.80 (0.2)	0.74 (0.3)
Nav Entry	0.74 (0.1)	0.92 (0.1)	0.82 (0.2)
Phone	0.67 (0.1)	0.84 (0.2)	0.74 (0.2)
0-Back	0.49 (0.2)	0.59 (0.3)	0.53 (0.3)
1-Back	0.31 (0.2)	0.37 (0.3)	0.34 (0.3)
2-Back	0.31 (0.3)	0.32 (0.3)	0.32 (0.3)
Radio Manual Easy	0.98 (0.3)	1.02 (0.3)	1.00 (0.3)
Radio Manual Hard	0.90 (0.2)	1.11 (0.2)	0.99 (0.2)
Radio Voice Easy	0.70 (0.2)	0.84 (0.2)	0.76 (0.2)
Radio Voice Hard	0.69 (0.2)	0.87 (0.2)	0.77 (0.2)
Song Select	0.71 (0.2)	0.86 (0.1)	0.78 (0.2)
Song Fail	0.85 (0.2)	0.90 (0.2)	0.87 (0.2)

As detailed on the next page, there was a main effect of age on mean glance time, with older participants overall showing a somewhat longer mean glance time.

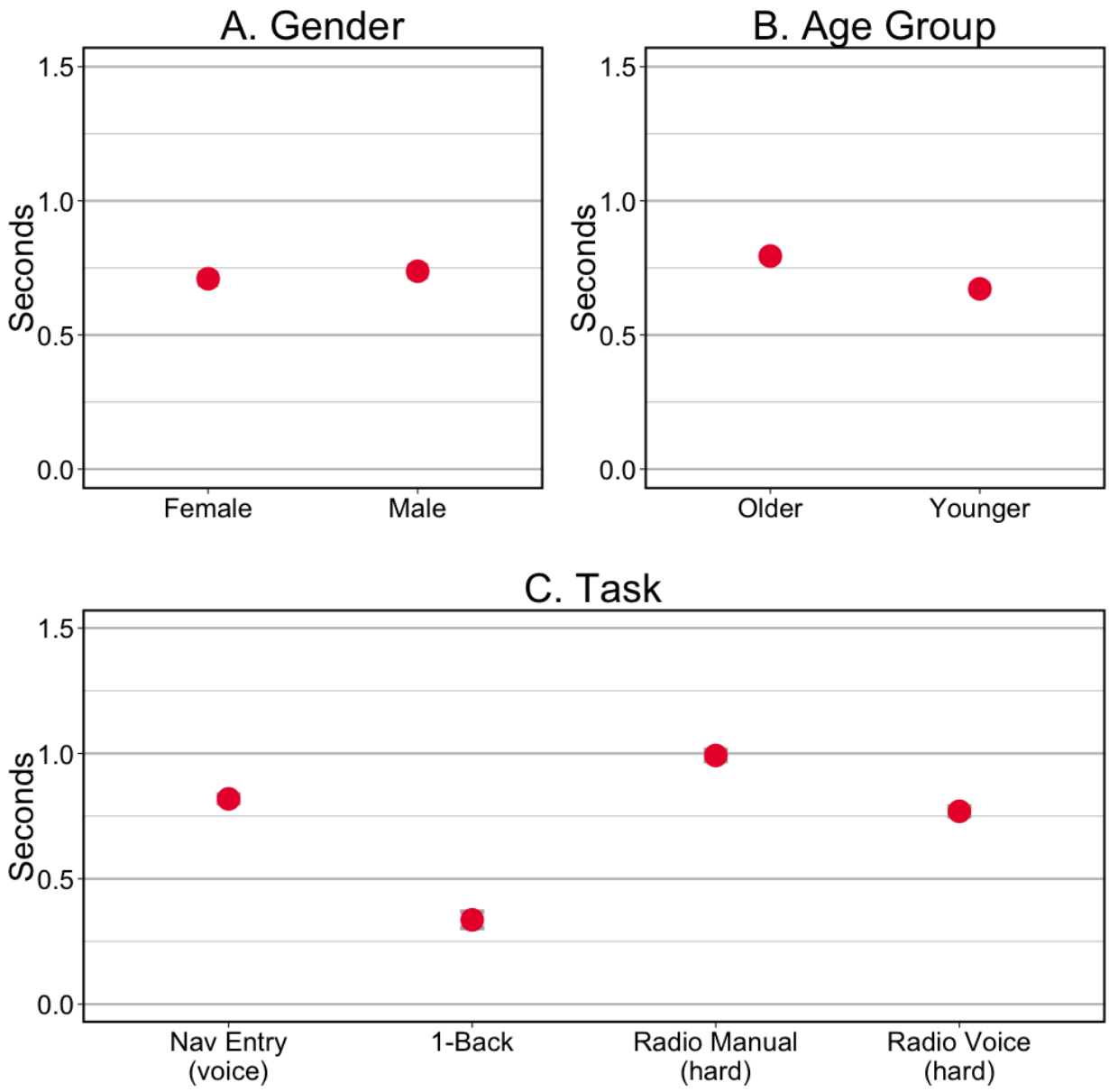


Figure 48: Statistical summary plots for mean duration of off-road glances.

Gender did not significantly affect mean glance duration ($p = .210$). Age group significantly affected mean glance duration ($p < .001$), with older participants showing slightly elevated glance durations. In line with the design goals for a voice-command interface, mean glance time for the voice Radio Hard task was significantly lower than with the manual Radio Hard tasks. ($p < .001$). Similarly, mean glance time for the Nav Entry task was lower than that observed with the manual Radio Hard task ($p < .001$).

Percentage of Long Duration (> 2s) Glances

This measure considers the percentage of off-road glances during a task that are in excess of 2.0 seconds in duration. These values are based on determining the percentage of long duration glances per participant as a base datum. A Friedman test of long glance rate during secondary task performance shows a significant effect of task period ($X^2(12) = 104.1$, $p < .001$).

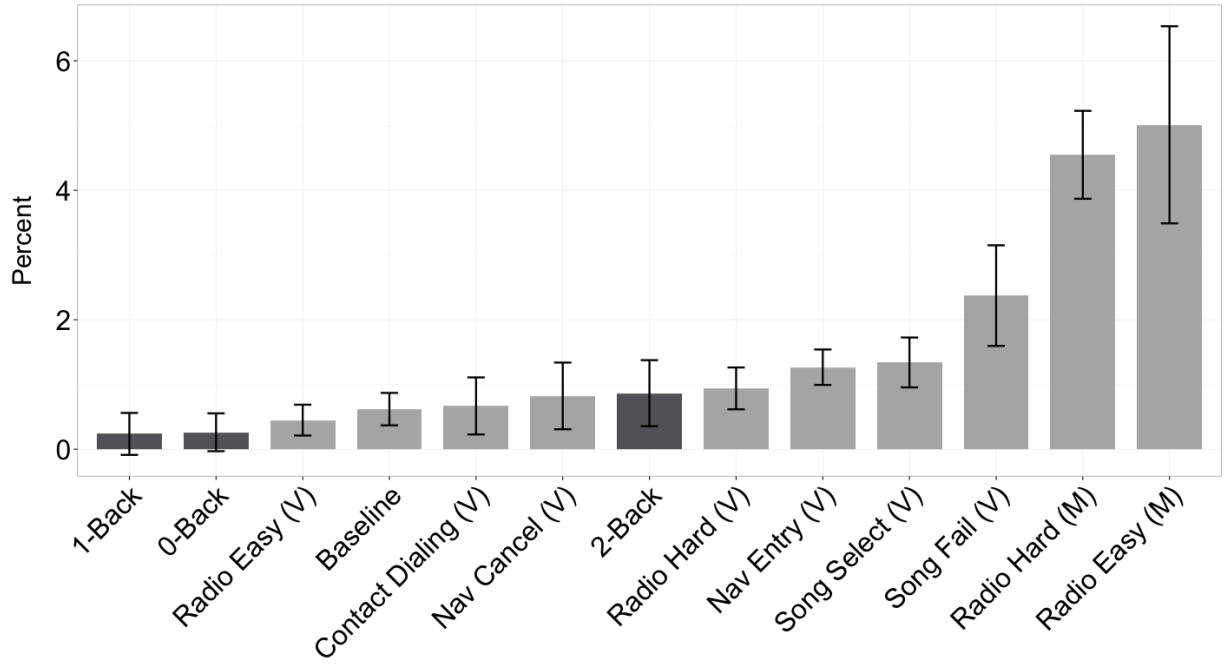


Figure 49: Tasks listed in ascending order for percentage of off-road glances in excess of 2.0 seconds. The baseline shown represents the combined baseline periods recorded prior to each task. Error bars represent 1 SEM.

As expected, the highest percentage of long glances occurred during the two classic visual-manual tasks involving manual radio tuning. It can be observed that the nominally highest value for this long glance metric appears when participants are engaged with the single key-press task of selecting a pre-set radio station.

The short horizontal bars above each task category represent the point at which 85% of the participants fall. If one were to apply the NHTSA visual-manual distraction criteria to the tasks studied here, the bar representing this 85% level must fall at or below the dashed line for an age and gender compliant sample.

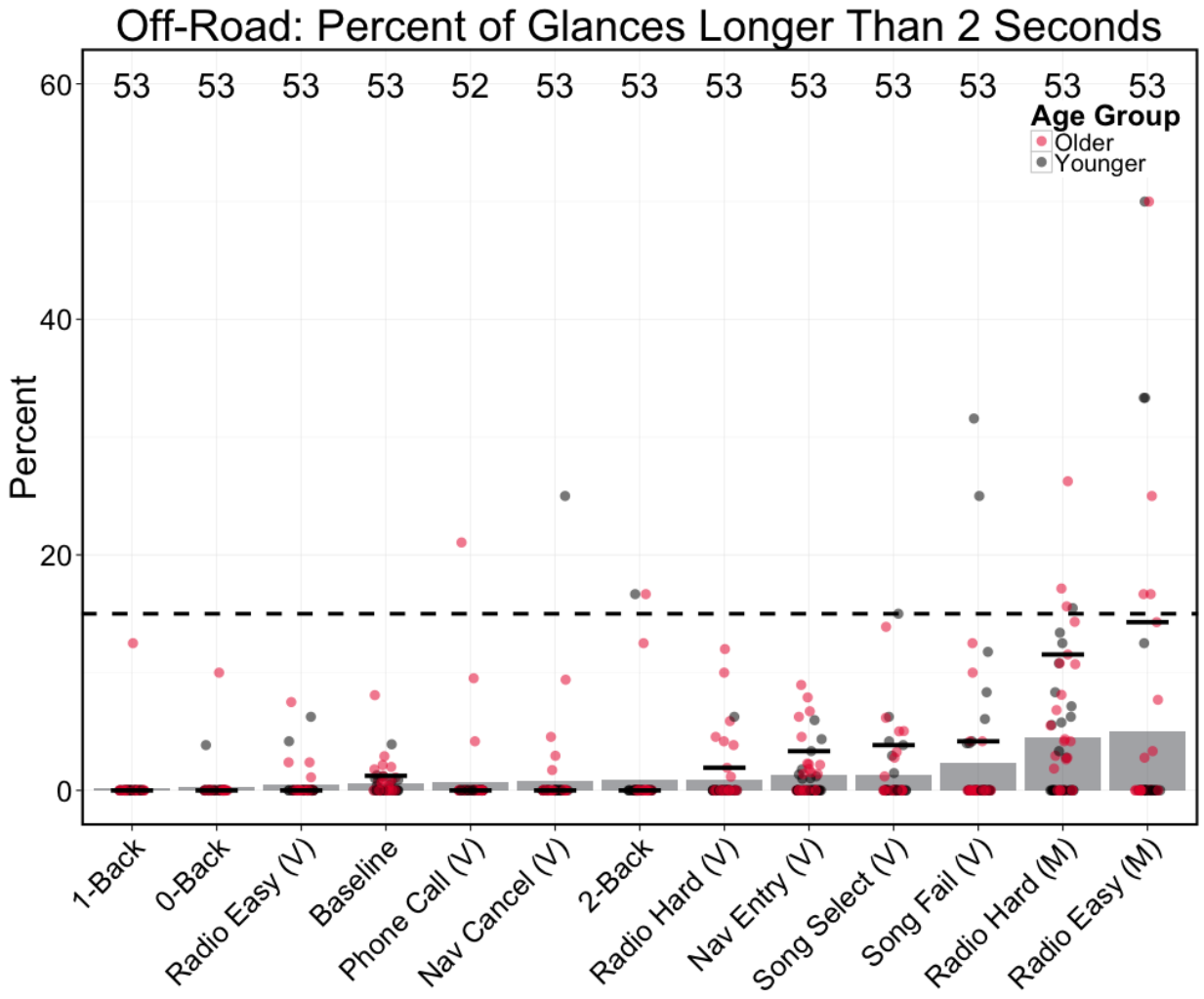


Figure 50: Percentage of glances longer than 2 seconds while performing each secondary task.

Most participants maintained a very low long glance rate. Note that the manual Radio Easy task was typically completed very quickly, which may distort the presence of a relatively small number of longer off-road glances (see Task Completion Time, above).

Table 24: Means (and standard deviations) of glances longer than 2 seconds (percentages).

Task	Younger	Older	(All)
Baseline	0.33 (0.8)	0.99 (1.8)	0.62 (1.3)
Nav Cancel	0.83 (4.6)	0.81 (2.2)	0.82 (3.7)
Nav Entry	0.67 (1.4)	2.05 (2.8)	1.27 (2.2)
Phone	0.13 (0.7)	0.43 (2.1)	0.26 (1.5)
0-Back	0.00 (0.0)	0.54 (2.6)	0.24 (1.7)
1-Back	0.56 (3.0)	1.27 (4.2)	0.86 (3.6)
2-Back	0.00 (0.0)	1.58 (4.9)	0.67 (3.2)
Radio Manual Easy	4.31 (12.2)	5.93 (12.0)	5.01 (12.0)
Radio Manual Hard	3.04 (4.8)	6.51 (6.9)	4.55 (6.0)
Radio Voice Easy	0.35 (1.3)	0.58 (1.7)	0.45 (1.5)
Radio Voice Hard	0.21 (1.1)	1.89 (3.4)	0.94 (2.5)
Song Select	3.16 (7.5)	1.34 (3.4)	2.37 (6.1)
Song Fail	1.12 (3.0)	1.62 (3.3)	1.34 (3.1)

As detailed on the next page, there was a main effect of age with an overall pattern of older participants have a higher percentage of longer duration glances greater than 2.0 seconds.

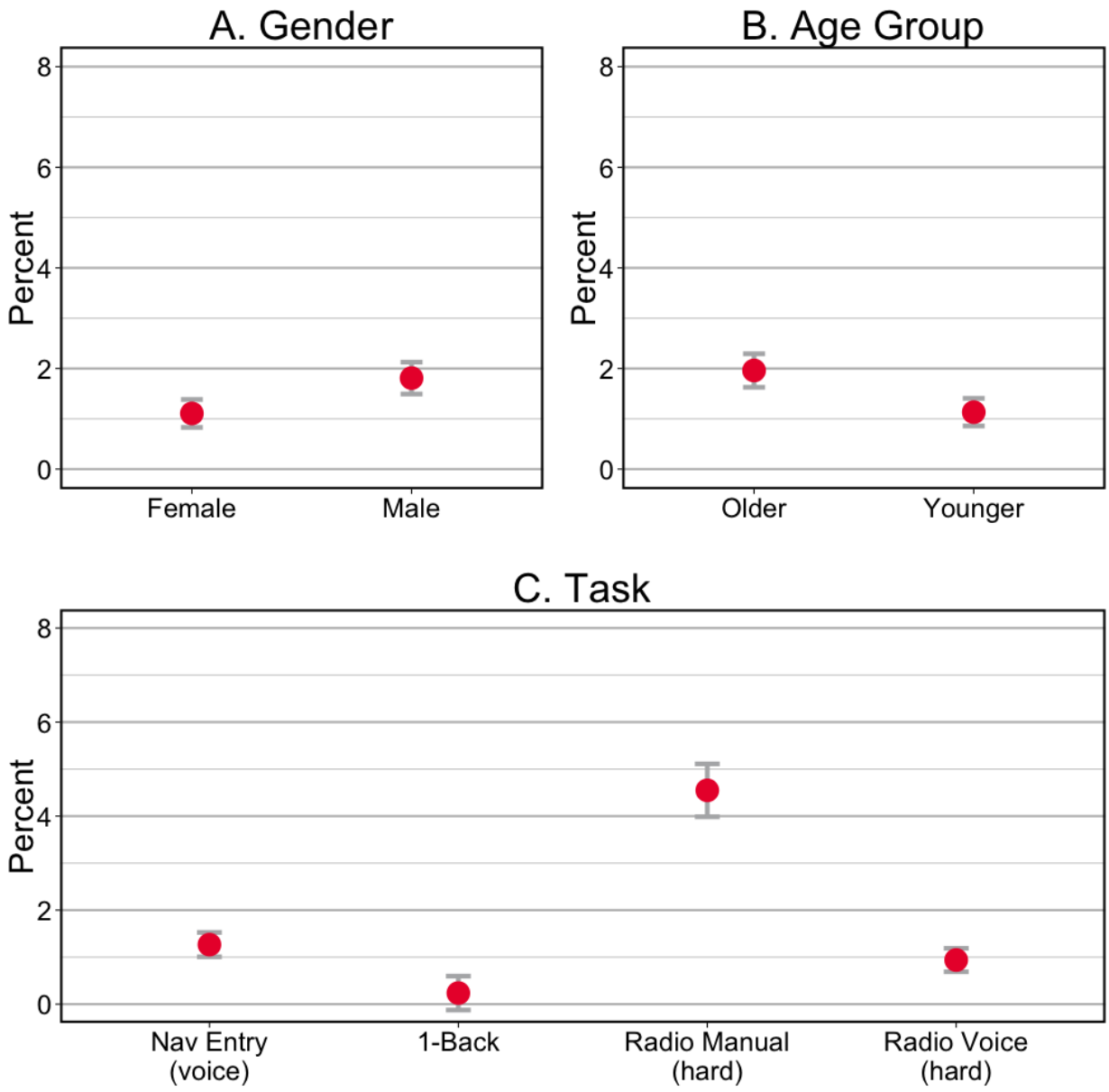


Figure 51: Statistical summary plots for glances longer than 2 seconds.

The percentage of long glances was not significantly affected by gender ($p = 0.128$), but was affected by age group ($p = .031$). In line with the design goals for a voice-command interface, the percentage of long glances for the voice Radio Hard task was significantly lower than with the manual Radio Hard tasks. ($p < .001$). Similarly, the percentage of long glances for the Nav Entry task was lower than that observed with the manual Radio Hard task ($p < .001$).

Total Off-Road Glance Time

This measure considers the sum of the durations of each individual participant's eye glances away from the forward road scene. A repeated measures ANOVA shows a significant effect of task period for the NHTSA (2013) definition of total eyes-off-road-time (TEORT) ($F(12, 612) = 68.70$, $p < .001$).

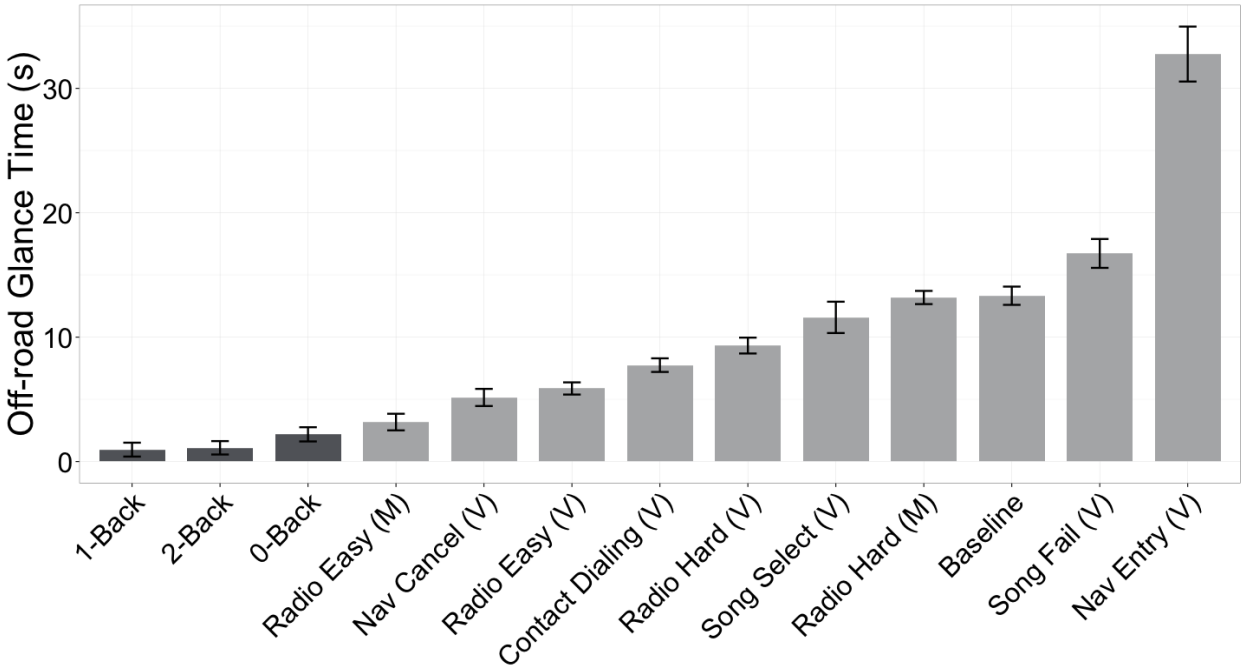


Figure 52: Tasks listed in ascending order for the amount of glance time away from the forward road scene that occurred during the completion of each task. Error bars represent 1 SEM. Tasks marked (V) used the voice interface. Tasks marked (M) utilized traditional manual/tactile interactions. Baseline represents the mean off-road glance time for 2 minute periods averaged across all 7 baselines collected. The n-backs represent mean values for 30 second periods.

The Navigation Entry task clearly is associated with a much longer off-road glance time than the other tasks. It should be noted that the Song Fail task was deliberately designed to be an impossible task to complete successfully. Baseline period glance time is relatively high, likely because of its duration (two minutes) relative to most traditional in-vehicle HMI tasks. Thus, the fact that a short duration task has a lower TEORT value does not necessarily indicate that less visual distraction is present. As pure auditory-vocal cognitive tasks, it can be observed that off-road glance time is very low for each level of the 30 second long n-back task periods.

Table 25: Means (and standard deviations) of total off-road glance time.

Task	Younger	Older	(All)
Baseline	11.70 (5.4)	15.45 (5.5)	13.33 (5.7)
Nav Cancel	3.03 (2.1)	7.92 (10.5)	5.15 (7.5)
Nav Entry	25.89 (10.9)	41.71 (26.0)	32.76 (20.4)
Phone	5.07 (3.1)	11.40 (8.1)	7.75 (6.5)
0-Back	2.19 (1.9)	2.19 (1.8)	2.19 (1.8)
1-Back	0.79 (0.9)	1.19 (1.7)	0.96 (1.3)
2-Back	0.99 (1.2)	1.25 (1.8)	1.10 (1.5)
Radio Manual Easy	1.98 (1.2)	4.73 (6.0)	3.17 (4.3)
Radio Manual Hard	10.19 (2.1)	17.10 (4.3)	13.19 (4.7)
Radio Voice Easy	3.84 (2.5)	8.53 (7.2)	5.88 (5.6)
Radio Voice Hard	6.87 (4.2)	12.52 (10.0)	9.32 (7.7)
Song Select	6.94 (5.9)	17.66 (14.3)	11.59 (11.6)
Song Fail	18.52 (9.3)	14.40 (9.2)	16.73 (9.4)

As detailed at the end of this section, there was a main effect of age across the tasks with older participants having longer off-road glance times overall.

The short horizontal bars above each task category represent the point at which 85% of the participants fall. If one were to apply the NHTSA visual-manual distraction criteria to the tasks studied here, bars representings the 85% level must fall at or below the dashed line for an age and gender compliant sample.

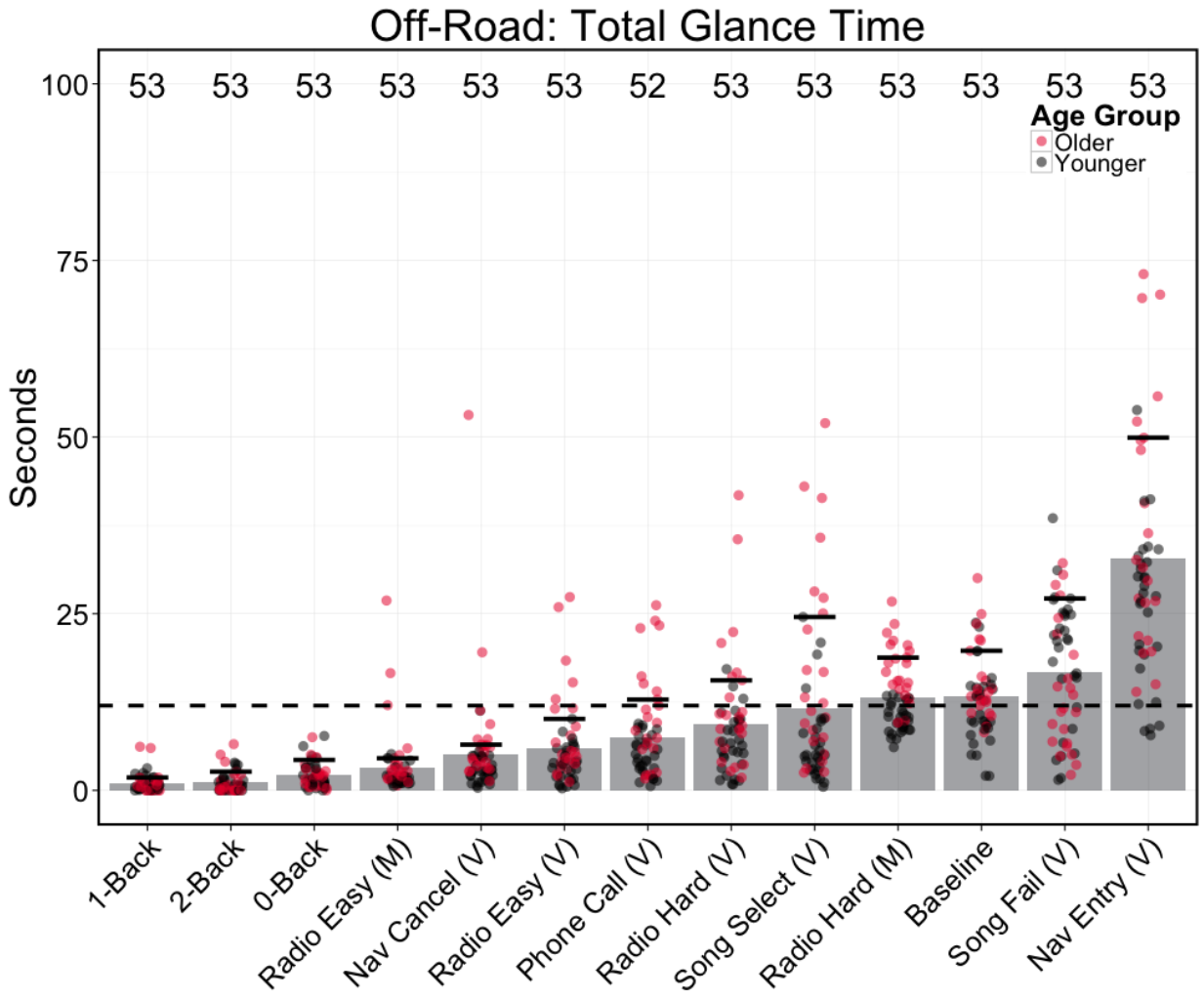


Figure 53: Total off-road glance time for each task. One outlier data point in the Nav Entry task is excluded from view to improve the readability of the plot.

Based on the data presented here, it appears highly unlikely that a NHTSA age compliant sample would meet the guidelines for total glance time off-the-forward-roadway for the navigation device address entry task studied in the test-vehicle. A number of other tasks also have relatively long total glance times relative to this criterion, in particular the Radio Hard tasks and the Song Selection task. As detailed in Appendix C, TEORT is lower for some tasks the second time participants engage with it on-road. Both the manual and voice based Radio

Hard tasks are much closer to the threshold for the second trials. However, looking only at the second trial would not bring the TEORT value for Nav Entry closer to the criterion line. It is also informative to look at the impact of using The Alliance (2006) metrics when looking at the Radio Hard task (see Appendix A). When using the glance-to-device measure, greater than 85% of our sample meets a 12 second criterion when considering the voice Radio Hard task; it would not meet the criterion if the eyes-off-the-forward-roadway measure is applied. If the manual Radio Hard task employed in this study is evaluated using the glance-to-device metric and The Alliance 20 second reference (criterion 2.1 A), then greater than 85% of our sample meets this reference point; if the NHTSA 12 second threshold is applied, it would not. The Nav Entry task, however, would fail to meet either criteria if applied.

We wish to emphasize again that the Baseline condition shown in the graph represents glances off the forward roadway for two minutes of "just driving" and the data is presented for reference purposes only regarding the distribution of glance time across participants for that length of time. The TEORT value for the baseline period should not be interpreted as "failing" the total glance time criterion.

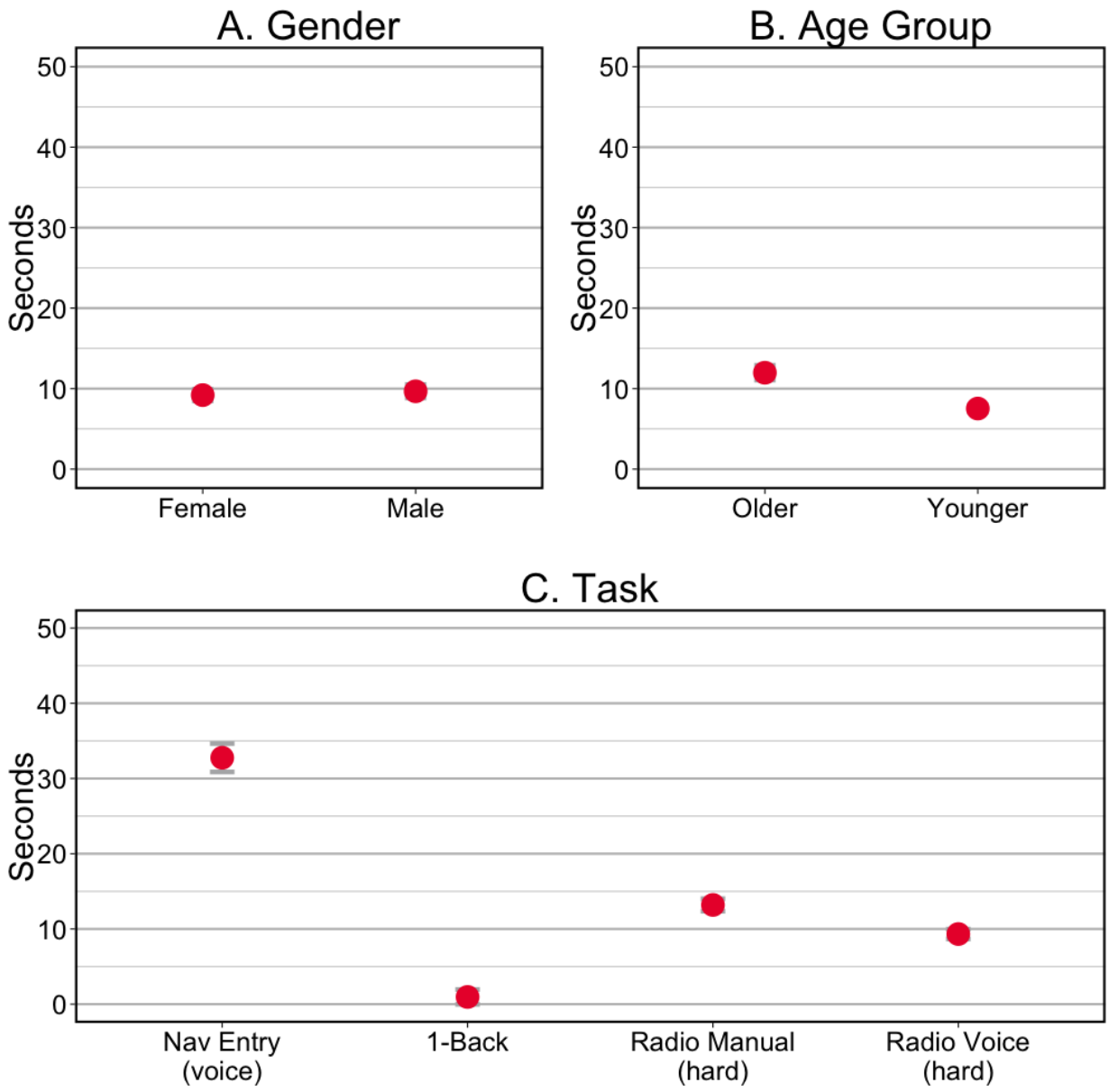


Figure 54: Statistical summary plots for total off-road glance time.

Gender did not significantly affect off-road glance time ($p = .730$). However, there was a significant effect of age group ($p < .001$), with the older group showing an overall longer total time for eyes off the forward roadway. In line with the design goals for a voice-command interface, the TEORT value for the voice version of the Radio Hard task was significantly lower than for the manual Radio Hard task ($p < .001$). For the Nav Entry task, however, TEORT was significantly higher than for the manual Radio Hard task ($p < .001$).

Number of Glances

A Friedman test of number of off-the-forward-roadway glances during secondary task performance shows a significant effect of task ($X^2(12) = 465.4$, $p < .001$).

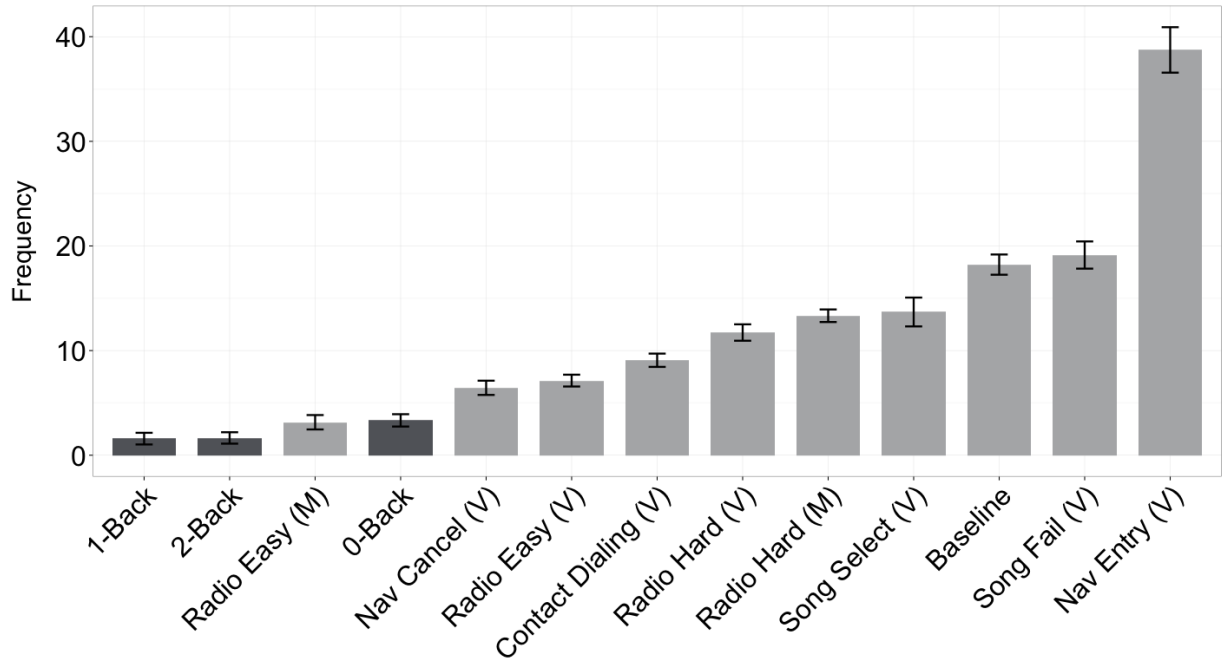


Figure 55: Tasks listed in ascending order for number of off-road glances. The baseline shown represents the combined baseline periods recorded prior to each task. Error bars represent 1 SEM.

The number of glances off-the-forward-roadway during a task is not currently used as an analysis metric in either The Alliance or the NHTSA visual-manual distraction guidelines. It does, however, provide an interesting point of comparison with the other metrics. The tasks line-up in a pattern quite similar to that seen in the total off-road glance time plot. In specific, the Nav Entry task and the Song Fail condition involve the largest total number of glances while the pure auditory-vocal n-back tasks and the discrete manual Radio Easy task involve the fewest number of glances off the forward roadway.

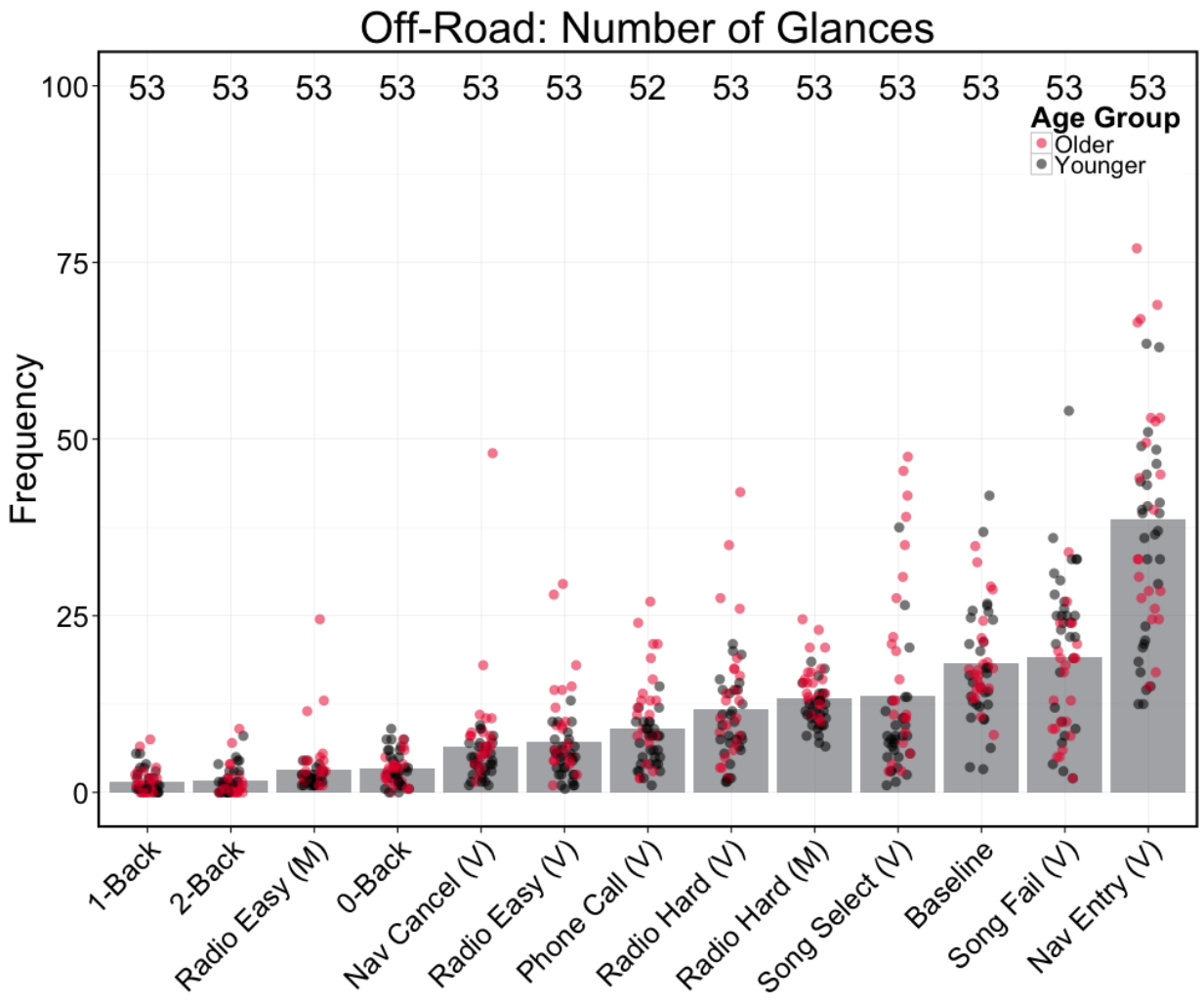


Figure 56: Number of off-road glances during each task. One outlier data point in the Nav Entry task is excluded from view to improve the readability of the plot.

Although not formally part of NHTSA's visual-manual distraction criteria, the number of off-road glances is implicit in the measures of mean glance duration and long glance rate, and it may be useful to visualize the number of glances separately. The number of off-road glances made during the N-Back tasks is quite small, particularly compared to the Baseline period, indicative of a concentration of gaze effect that occurs as cognitive demand (but not visual demand) increases. See Reimer et al. (2012) for further discussion of the concentration of gaze phenomena.

Table 26: Means (and standard deviations) of off-road glance frequency.

Task	Younger	Older	(All)
Baseline	17.67 (8.8)	18.93 (6.8)	18.22 (7.9)
Nav Cancel	4.43 (2.3)	9.07 (9.1)	6.44 (6.6)
Nav Entry	34.53 (14.3)	44.22 (22.8)	38.74 (18.9)
Phone	7.37 (4.1)	12.94 (7.5)	9.73 (6.4)
0-Back	3.55 (2.5)	3.04 (2.0)	3.33 (2.3)
1-Back	1.48 (1.6)	1.72 (2.0)	1.58 (1.8)
2-Back	1.65 (1.9)	1.65 (2.4)	1.65 (2.1)
Radio Manual Easy	2.07 (1.0)	4.57 (5.3)	3.15 (3.7)
Radio Manual Hard	11.68 (2.9)	15.48 (3.8)	13.33 (3.8)
Radio Voice Easy	5.07 (3.1)	9.83 (7.4)	7.13 (5.9)
Radio Voice Hard	9.72 (5.6)	14.35 (10.2)	11.73 (8.2)
Song Select	9.32 (7.6)	19.39 (14.2)	13.69 (11.9)
Song Fail	21.93 (11.5)	15.48 (8.2)	19.13 (10.6)

As detailed at the end of this section, the older group generally showed more glances across the periods.

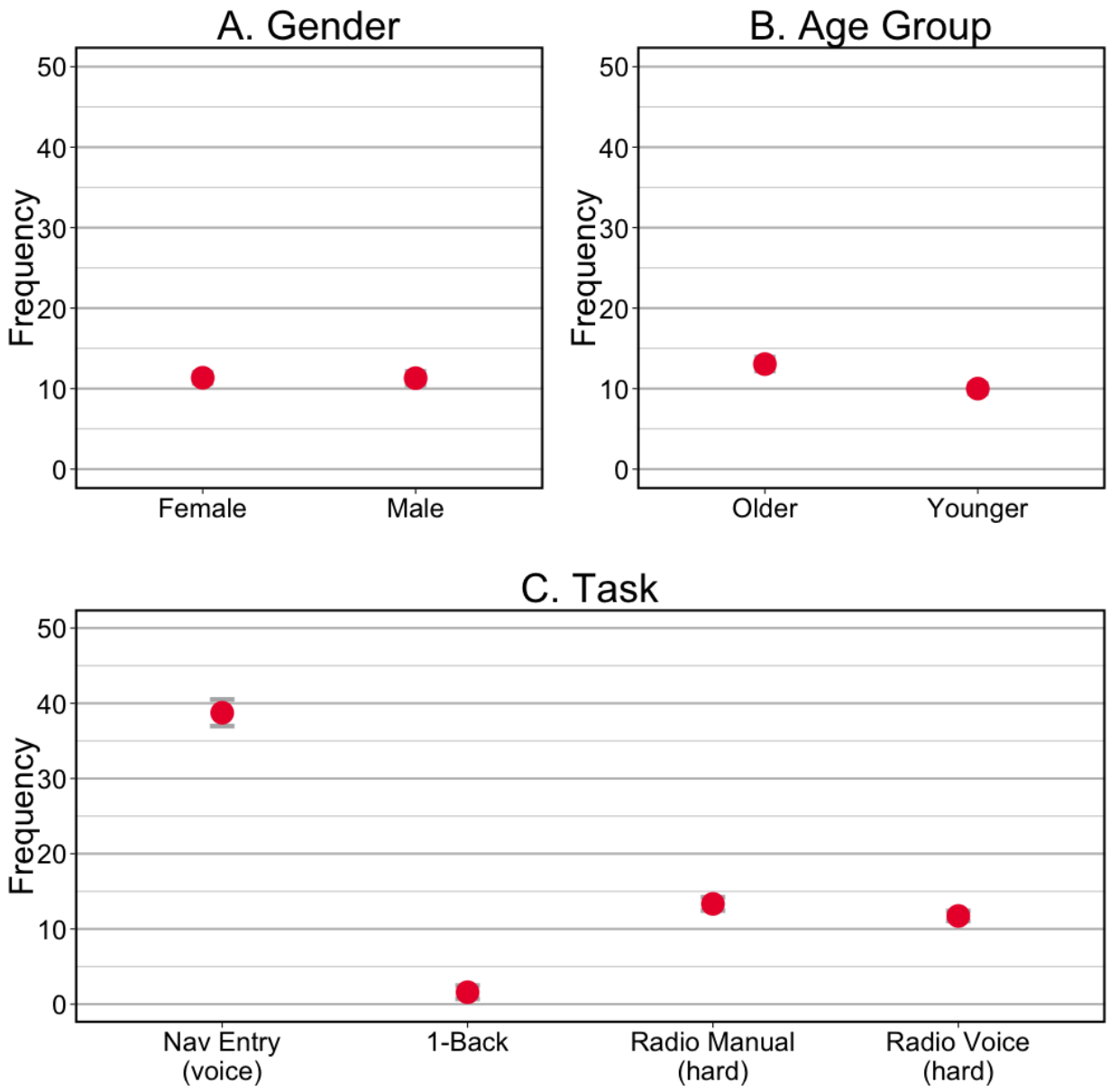


Figure 57: Statistical summary plot for number of off-road glances.

Number of glances was not significantly affected by gender (p = 0.48). However, the number of glances was affected by age group (p = .003), with the older group showing more glances overall. In contrast with what might have been expected, while the number of glances for the voice version of the Radio Hard task was nominally lower than for the manual Radio Hard version, the difference was not statistically significant (p = .079). Relative to the version of the

manual Radio Hard reference task studied here, the Nav Entry task involved a significantly higher number of glances off-the-forward roadway ($p < .001$).

Bootstrap Analysis Sampling Sets of 24 Participants

NHTSA recommends that a minimum sample of 24 participants (equally balanced between men and women, and between age groupings of 18-24, 25-39, 40-54, and 55+ years) be used to gauge the distraction potential of an in-vehicle device. As noted earlier, our sample does not include participants in these specific demographic ranges. However, our sample size of 60 is substantially larger than the recommended minimum of 24. One may wonder how likely a study would be to replicate the pass/fail results shown in the preceding table using a sample size of 24 participants.

We conducted a bootstrap analysis to address this concern (see Table 27 on next page). Six participants were randomly selected from each age*gender group to produce a sub-sample of 24 participants. Pass/fail results were then calculated and compared to the results of the full sample. This random selection and calculation was then repeated 10,000 times. The percentages of bootstrap matches to our full data set are shown in the table below. Of note, all replications passed the mean glance duration criterion for all tasks. Conversely, no replications produced a sample with a passing total off-road glance time for the Navigation Entry task (100% agreement with the overall sample's failing result).

Table 27: The percentage of 10,000 24-participant samples that produced pass/fail criteria results identical to the Primary Analysis. Note that the closer the sample was to the 85% pass threshold, the more variable the replication becomes.

Task	Age Group	Long Duration Glances	Mean Glance Duration	Total Off-Road Glance time
Nav Cancel	Younger	100.00%	100.00%	100.00%
	Older	100.00%	100.00%	**83.50%**
	(all)	100.00%	100.00%	100.00%
Nav Entry	Younger	100.00%	100.00%	100.00%
	Older	100.00%	100.00%	100.00%
	(all)	100.00%	100.00%	100.00%
Radio Manual Easy	Younger	**66.40%**	100.00%	100.00%
	Older	**59.20%**	100.00%	**38.00%**
	(all)	**69.30%**	100.00%	100.00%
Radio Manual Hard	Younger	100.00%	100.00%	92.10%
	Older	**36.90%**	100.00%	100.00%
	(all)	92.90%	100.00%	100.00%
Radio Voice Easy	Younger	100.00%	100.00%	100.00%
	Older	100.00%	100.00%	92.60%
	(all)	100.00%	100.00%	**74.60%**
Radio Voice Hard	Younger	100.00%	100.00%	**65.50%**
	Older	100.00%	100.00%	99.70%
	(all)	100.00%	100.00%	93.50%
Song Select	Younger	100.00%	100.00%	**46.70%**
	Older	100.00%	100.00%	100.00%
	(all)	100.00%	100.00%	99.90%
Song Fail	Younger	86.00%	100.00%	100.00%
	Older	100.00%	100.00%	100.00%
	(all)	100.00%	100.00%	100.00%
Phone	Younger	100.00%	100.00%	100.00%
	Older	100.00%	100.00%	99.10%
	(all)	100.00%	100.00%	**79.70%**

Glance Distribution Analyses

We observed that glance strategies seemed to vary significantly between drivers. An analysis of overall glance distributions, visualized in **Figure X**, shows all off-road glances made during fourteen minutes of baseline driving (top panels) and across all secondary task periods (bottom panels).

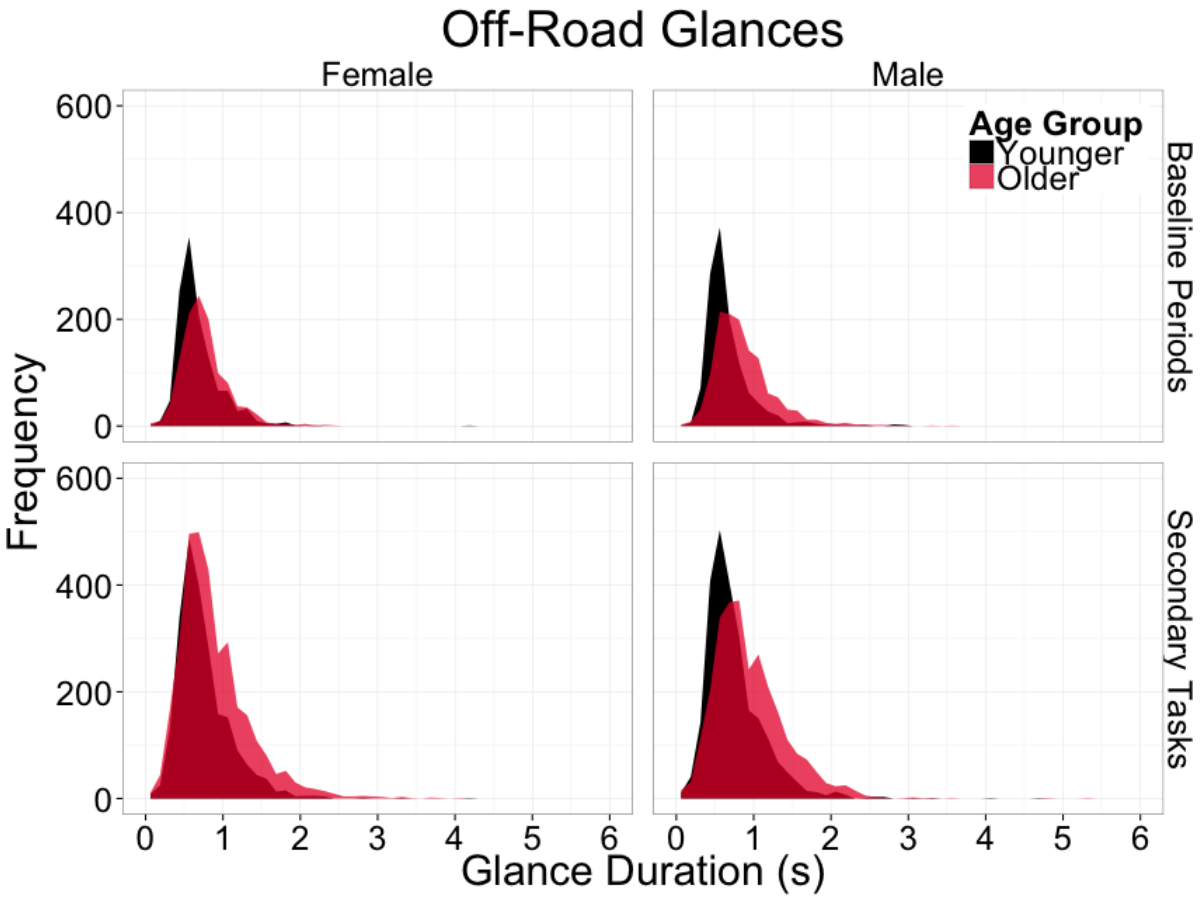


Figure 58: Glance frequency distributions for baseline and task periods.

Since glance data were not available for all participants, a subset of the coded data were randomly selected such that nine participants remained in each Age * Gender subgroup. A statistical test (ANOVA) on each participant's number of task period off-road glances shows that younger participants make an mean of 8.7 off-road glances during tasks, whereas older participants make 11.1 ($F_{(1, 32)} = 4.98$, $p = .033$).

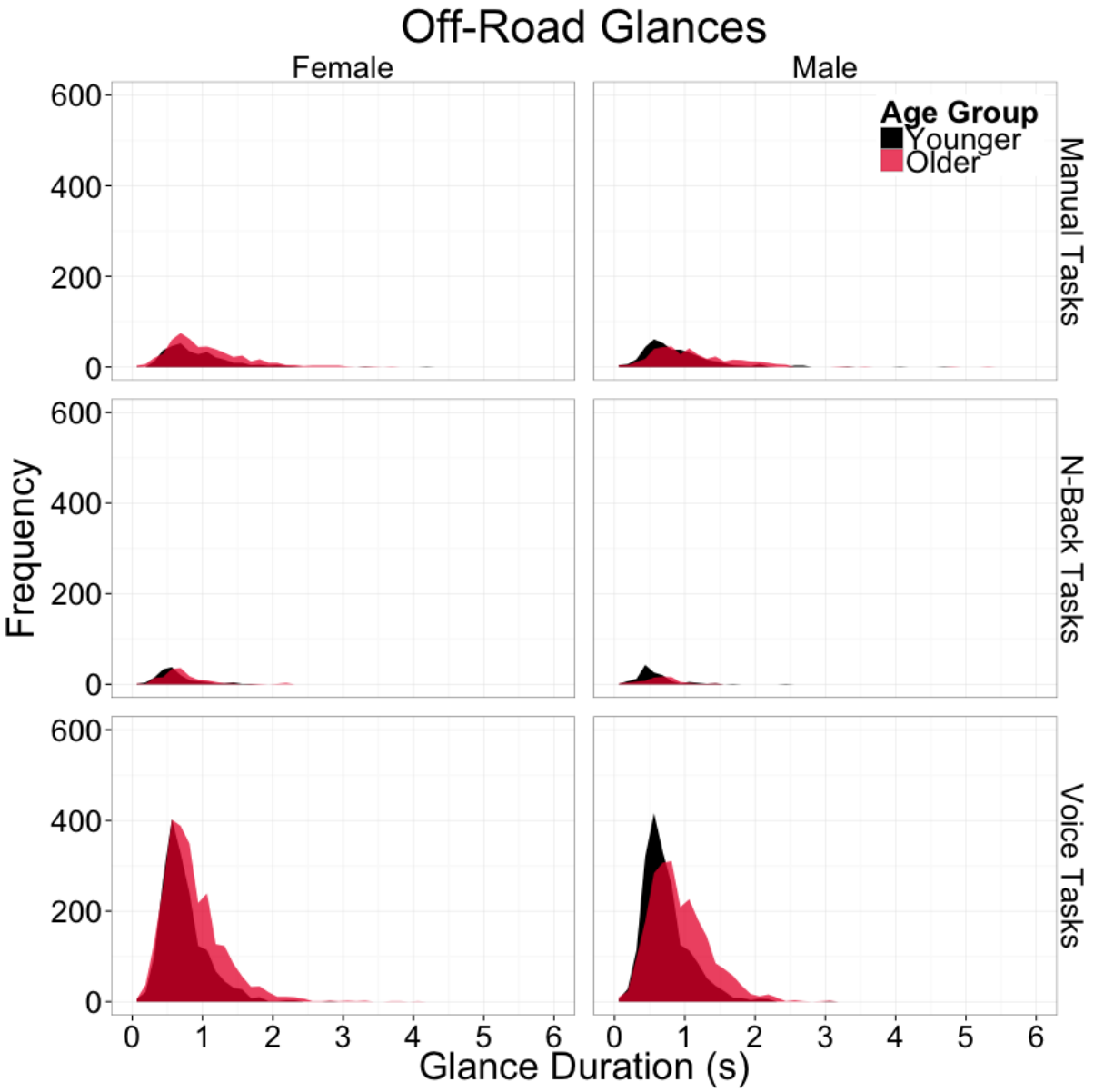


Figure 59: Glance frequency distributions by task type.

Orienting Response

The Orienting Response (OR) rating reflects a subjective judgment of the extent to which participants appeared to engage directly with the center stack display screen at some point while performing a voice-command enabled task (see Appendix D for details on how OR was defined for rating purposes). It should be noted that this analysis does not explicitly distinguish glances for visual confirmation from glances associated with OR behavior, and it is recognized that this is a partial confounding factor in assessing this behavior pattern.

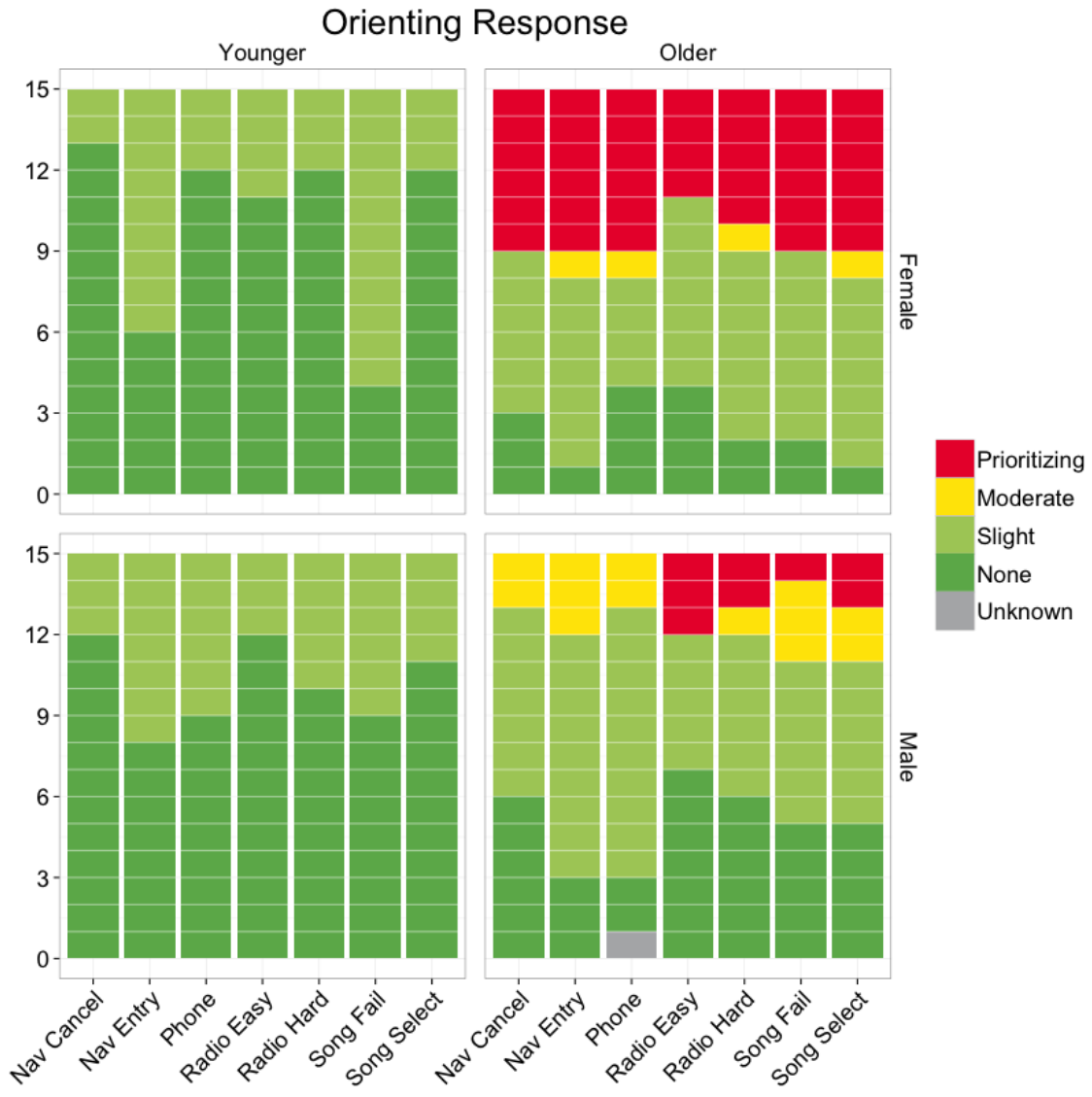


Figure 60: Displays OR for 7 voice control tasks, split by age group and gender. Participants performed each task twice (except Song Fail).

Each cell in Figure 60 represents the participant's stronger OR for the task. For example, if a participant performed two trials of the Nav Entry task and was scored as displaying a Moderate and Prioritizing OR, the Prioritizing OR is the one plotted. Factorial ANOVA using mean OR code per participant as a dependent measure reveals a significant effect of age group ($F_{(1, 56)}$ = 29.1, $p < .001$), and a borderline effect of gender ($F_{(1, 56)}$ = 3.9, $p = .053$). Age and gender also show a borderline interaction effect ($F_{(1, 56)}$ = 3.7, $p = .060$). The interaction is somewhat limited due to the small sample size, though the plot makes the age * gender effect clear. This data strongly suggests that the older adults, and in particular older adult women, are orienting themselves towards the interface. This tendency, seen most strongly here in older adults, may be a potential limitation associated with use of these technologies among older adults since this behavior tends to move the eyes off the forward roadway and may place the driver in a body position that impacts the speed with which they can respond to unexpected events.

Summary Comparison of Manual & Voice-Based Radio Hard Tuning Task

Table 28 below summarizes the results of the statistical tests reported in the previous sections that compare the manual (m) engagement with the Radio Hard tuning task and the voice-command (v) based method of engaging with the same task on various measures. The "+ Voice" column indicated measures for which the voice-command method showed a level of demand or impact on the driver of the task that might reasonably be interpreted as being less demanding or having a smaller impact on that variable than is measured when engaging in the task using the manual interface. The "- Voice" column indicates measures for which the manual method might be interpreted as being less demanding.

Table 28: Manual vs. Voice-Based Radio Hard Tuning Tasks (Wilcoxon tests)

Measure	+ Voice	- Voice	V statistic	p value
Self-Reported Workload	m > v		619	**p = .036**
Task Completion Time		m < v (?)	39	p **< .001***
Heart Rate	m > v		1205	p **= .033***
SCL			743	p = .456
Mean Velocity			695	p = .106
SD Velocity			754	p =.237
Acceleration Events	m > v		654	**p = .004**
SD Steering Wheel Angle			1081	p = .223
Minor SWR			1064	p = .274
Major SWR	m > v		1636	p **< .001***
EOFR Mean Glance Duration	m > v		1347	p **< .001***
EOFR % Glances > 2s	m > v		426	p **< .001***
EOFR Total Glance Time	m > v		1175	p **< .001***
EOFR Number of Glances			915	p = .079
GTD Mean Glance Duration	m > v		1392	p **< .001***
GTD % Glances > 2s	m > v		398	p **< .001***
GTD Total Glance Time	m > v		1349	p **< .001***
GTD Number of Glances	m > v		1283	p **< .001***

m = Manual Radio Hard tuning task; v = voice-based based Radio Hard tuning task
EFOR = Eyes-Off-Forward-Roadway; GTD = Glance-to-Device

Note: The distribution of the V statistic for the Wilcoxon test is such that smaller and larger values are associated with the two ends of the distribution (and hence statistical significance).

With the exception of task completion time, which is marked in the table with a question mark (?), use of the voice-command interface to engage carryout the Radio Hard tuning task appeared to have a clear advantage on a number of metrics over the manual method in terms of the level of demand / distraction imposed on the driver. Compared to the manual method of completing the Radio Hard tuning task, the voice option was given a lower self-reported workload rating, was associated with lower heart rate and SCL levels, and was associated with a lower rate of acceleration events and major steering wheel reversals. In terms of glance metrics, voice control of the Hard Radio tuning task was found to have a lower mean glance duration, a lower percentage of glances longer than 2 seconds, and a lower total eyes-off-road-time. For the remaining measures collected, the level of demand / impact of the driver was indistinguishable. The implications of this pattern of results will be considered further in the discussion. Regarding task completion time, when all other factors are equal, the ability to complete a task quickly is generally seen as advantageous. However, while the use of the voice interface to engage in the Radio Hard tuning task took significantly longer than the manual tuning method, it could be argued that total time to complete a task needs to be considered within a broader context of various demand features of the task.

For the glance metrics, the pattern of the relationships is the same regardless of whether the EFOR or the GTD method of characterizing glance behavior is used. See Appendix A for a more detailed consideration of the significance of the GTD vs. EFOR metrics.

Summary Comparison of Manual Radio Tuning and Voice Nav. Entry

Table 29 below summarizes the results of the statistical tests reported in the previous sections that compare the manual Radio Hard (mRH) tuning task and the voice-command based entry of addresses into the navigation system (Nav E) on various measures. The "+ Voice" column indicated measures for which the voice-command system showed a level of demand or impact on the driver of the task that might reasonably be interpreted as being less demanding or having a smaller impact on that variable than is measured when engaging in the manual Radio Hard tuning task. The "- Voice" column indicates measures for which the manual Radio Hard task might be interpreted as being less demanding.

Table 29: Manual Radio Hard vs. Voice-Based Nav. Entry (Wilcoxon tests)

Measure	+ Voice	- Voice	V statistic	p value
Self-Reported Workload			402	p = .275
Task Completion Time		mRH < Nav E (?)	1830	p **< .001***
Heart Rate	mRH > Nav E		506	p **= .003***
SCL			679	p = .884
Mean Velocity			1025	p = .420
SD Velocity		mRH < Nav E	1636	p **< .001***
Acceleration Events			705	p = .589
SD Steering Wheel Angle		mRH < Nav E	1564	p **< .001***
Minor SWR			876	p = .777
Major SWR	mRH > Nav E		113	p **< .001***
EOFR Mean Glance Duration	mRH > Nav E		87	p **< .001***
EOFR % Glances > 2s	mRH > Nav E		64	p **< .001***
EOFR Total Glance Time		mRH < Nav E	1417	p **< .001***
EOFR Number of Glances		mRH < Nav E	1430	p **< .001***
GTD Mean Glance Duration	mRH > Nav E		232	p **< .001***
GTD % Glances > 2s	mRH > Nav E		76	p **< .001***
GTD Total Glance Time		mRH < Nav E	1165	p **< .001***
GTD Number of Glances		mRH < Nav E	1137	p **< .001***

mRH = Manual Radio Hard tuning task; Nav E = voice-based address entry
EFOR = Eyes-Off-Forward-Roadway; GTD = Glance-to-Device

Note: The distribution of the V statistic for the Wilcoxon test is such that smaller and larger values are associated with the two ends of the distribution (and hence statistical significance).

In contrast with what was found in comparing the voice-command interface and the manual interface for engaging with the Radio Hard tuning task, comparing the voice-based address entry task with the manual Radio Hard tuning task used in this study as a relative reference point for a maximal acceptable level of demand on the driver results in a series of findings that raise concerns about the level of visual demand associated with address entry interface. As can be seen in the table, total glance time and total number of glances is significantly higher for the destination entry task. It is also associated with higher variability in the velocity control and steering wheel angle control metrics.

The categorization of task completion time is marked in the table with a question mark (?). When all other factors are equal, the ability to complete a task quickly is generally seen as advantageous. However, while the use of the voice interface to engage in the Radio Hard tuning task took significantly longer than the manual tuning method, it could be argued that total time to complete a task needs to be considered within a broader context of various demand features of the task.

For the glance metrics, the pattern of the relationships is the same regardless of whether the EFOR or the GTD method of characterizing glance behavior is used. See Appendix A for a more detailed consideration of the significance of the GTD vs. EFOR metrics.

Task Completion Data

Participants were coded on their ability to complete the task, and how much assistance they required to do so (see Appendix G for coding details). As in the Orienting Response section, each cell in Figure 61 represents the participant's worse performance between the two trials of each task. ANOVA using the mean task completion code as a dependent variable reveals a significant effect of age group ($F_{(1, 56)} = 49.9$, $p < .001$). There were no other significant effects (gender, $F_{(1, 56)} = 0.81$, $p = .372$; age group * gender, $F_{(1, 56)} = 1.82$, $p = .183$).

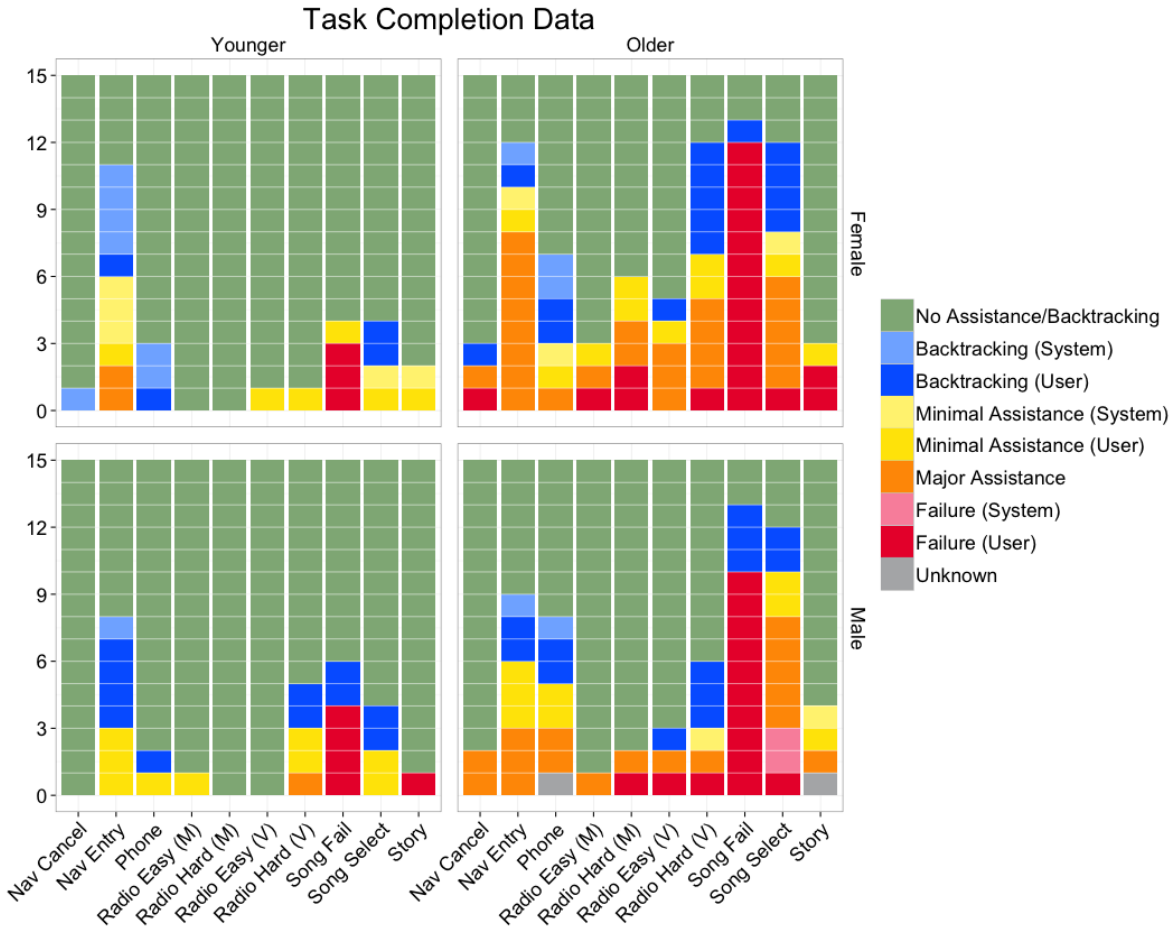


Figure 61: visualizes task completion data for the sample ("System" and "User" refer to system error and user error, respectively).

Effect of Task Completion Time

Figure 62 presents correlation plots between total task completion time and total off-road glance time for the hard radio tuning task (manual and voice), as well as the Navigation Entry task. There is a strong correlation between task completion time and total glance time for all three tasks (Pearson R > 0.74 for all three tasks, all p < .001).

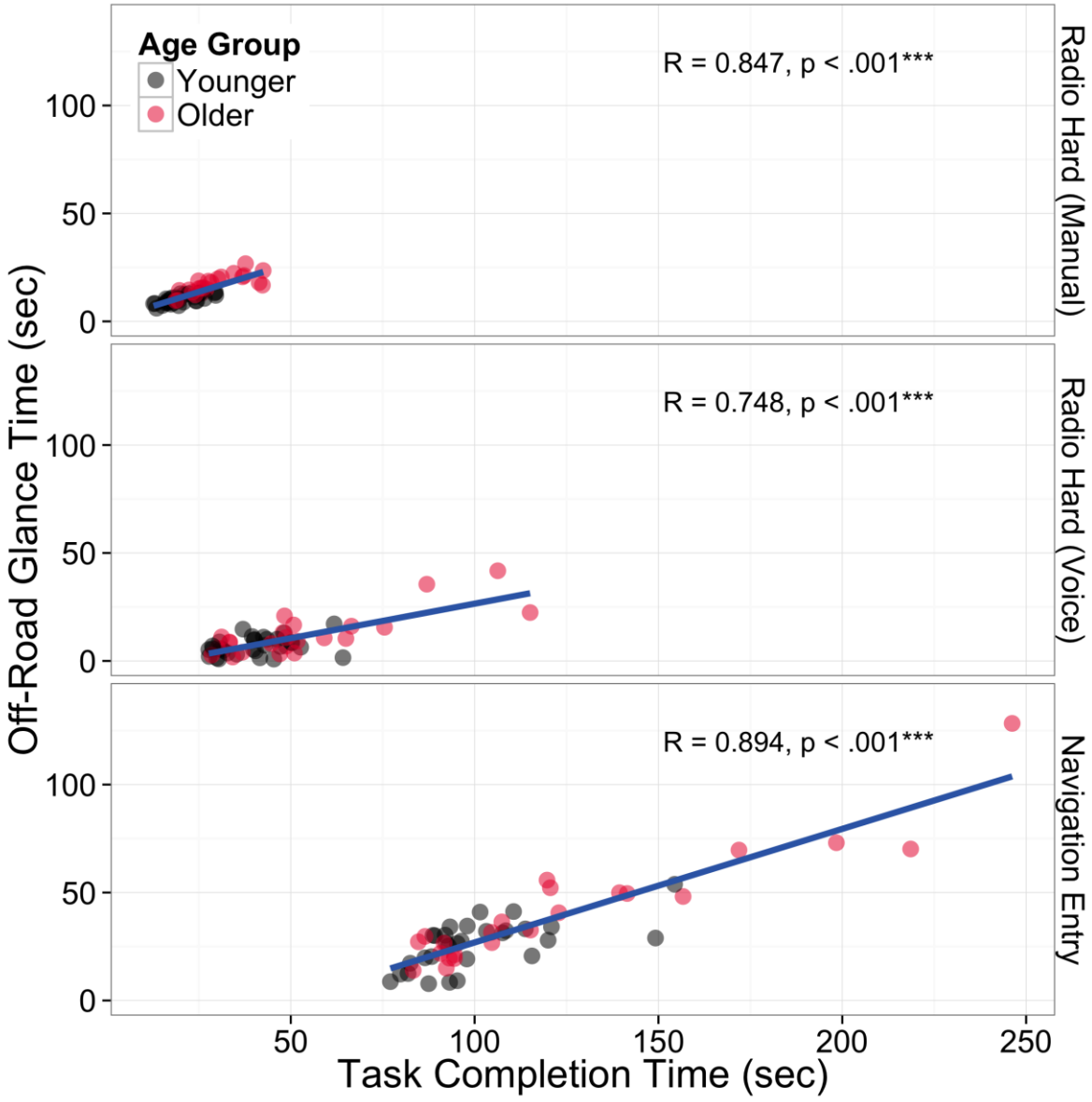


Figure 62: Off-Road glance time plotted against total task completion time.

Discussion

This technical report presents the analysis of data collected during the first phase of a larger project examining how drivers interact with production level voice-command interfaces during actual highway driving. The dataset created as part of the project is extremely rich, and we anticipate that additional analyses of specific issues may be done in the future. Due to the rapid proliferation of voice-command systems in production vehicles, NHTSA's current work on the development of guidelines for voice interfaces, and the findings of high visual-demand associated with address entry by voice in the navigation task, we felt it was important to present the basic findings for consideration by the broader development, research, and governmental regulatory community.

Key Observations

A number of comments and observations are provided throughout the presentation of data in the results section. In addition, we would like to highlight several aspects of those findings and suggest some conclusions:

1. Comparison of the visual-manual and voice-command interfaces for the complex radio tuning (Radio Hard) task in the vehicle under test - In line with the general design goals for providing a voice-command interface option, use of the voice interface could be considered to have provided equivalent or improved functionality across a number of measures.

    a. In specific, compared to the manual method of completing the Radio Hard tuning task, the voice option was given a lower self-reported workload rating, was associated with lower heart rate and SCL levels, was not statistically distinguishable in terms of a number of driving performance metrics (mean velocity, SD velocity, SD steering wheel angle, minor steering wheel reversals), and was associated with a lower rate of acceleration events and major steering wheel reversals.

    b. In terms of glance metrics, voice control of the Hard Radio tuning task was found to have a lower mean glance duration, a lower percentage of glances longer than 2 seconds, and a lower total eyes-off-road-time (TEORT).

2. Comparison of the workload / distraction associated with using the voice-command method of entering a full-address into the navigation system relative to a complex

manual radio tuning reference task - The results observed when participants interacted with the navigation interface in particular indicate that the visual feedback and menu structure for a voice activated system must be carefully designed for use while driving just as it should be for visual-manual systems where guidelines have already been established.

a. For the majority of the voice interface tasks in this study, the total eyes-off-road-time (TEORT) as defined by NHTSA was less than or equal to the TEORT found during two minutes of driving without a secondary task (baseline driving).

b. For navigation destination entry by voice, TEORT was more than 2 times of that observed during 2 minutes of baseline driving and for the duration of manual Radio Hard tuning task, and exceeded the criterion for visual-manual tasks as set by the NHTSA visual-manual guidelines.

c. Thus, interfaces designers should avoid assuming that providing a speech recognition engine is sufficient to minimize increasing the risk of a crash because the driver no longer needs to manipulate physical controls.

3. Age – Thirteen of the twenty variables assessed for age effects showed statistically significant main effects across task periods. The pattern of findings (summarized below) might be interpreted as indicating that older drivers experienced higher levels of demand from the tasks, at least partially compensated for the overall demand by driving slower, and did not show an appreciable vehicle control issues relative to the younger participants in dealing with the combined challenges of driving and engaging with the secondary tasks. The extent to which the older drivers' spare capacity might have been impacted is largely unknown, except to note that they show no overt decrement in the driving performance metrics compared to the younger participants.

a. Older participants had higher self-reported workload scores, longer task completion times, had higher percentage increases in SCL in response to tasks, had higher mean glance durations, percentages of long duration glances, total glance time, and number of glances. These glance metric differences appeared using both the eyes-off-the-forward-roadway and the glance-to-device methods of assessing glance behavior.

b. Older participants drove slower and showed less acceleration events.

c. The younger and older age groups did not differ significantly in standard deviation of steering wheel angle, major or minor steering wheel reversals, standard deviation of velocity, or heart rate.

4. Gender - Only two out of twenty of the variables assessed for gender effects were statistically significant. Given that one out of 20 tests might be expected to show a significance finding at the 0.05 level by chance alone, it is apparent that gender does not appear as a major overall factor in the data analysis.

    a. Perhaps the most meaningful observation for gender was the finding that, using the glance-to-device metric, males showed a higher main effect of percentage of glances greater than 2 seconds. Men also showed a non-significant trend toward longer mean glance-to-device at a p value of 0.099. These findings are consistent with previous work we have carried out looking at an in-vehicle display in a simulation study in which males showed higher mean duration of glances and a higher number of glances greater than 1.5 seconds (Reimer, Mehler, Matteson, Levantovsky, et al., 2012). Note - this gender effect was not statistically significant when considering the eyes-off-the-forward-roadway method of quantifying glance behavior.

    b. The other main effect for gender appeared for SCL, where women showed a higher overall value. Since most of these analyses were carried out considering SCL change scores, this might suggest a somewhat higher reactivity in the female sample. We have not seen this difference in previous unpublished work with the n-back task that considered young adults, so this finding should be interpreted cautiously. In that work, we found males to typically have higher baseline SCL values but saw males and females showing similar reactivity profiles.

5. Scaling of the in-vehicle interface control tasks against the n-back reference task - Using physiological measures as indirect indicators of cognitive demand, the results indicate that the cognitive workload associated with the vehicle interface tasks (both voice and visual-manual interfaces) performed in this study did not exceed that imposed by the moderate level of the 1-back task. This finding was somewhat in contrary to our expectation that the heart rate and SCL indicators of workload for some of the interface tasks might scale between the 1-back and the 2-back level.

a. A primary reason why the auditory presentation / verbal response delayed digit recall task (n-back) employed in this study works well for inducing scaled levels of workload is that there is relatively little that the driver can do to compensate for the increased level of load that comes from increasing the number of digits that have to be held in memory with each level of the task. In contrast, for the interface tasks evaluated, drivers had available to them the option to manage total demand to some degree by pacing their response to task steps and by reduce driving demands somewhat by slowing their speed of travel (and thus likely increase their headway distances). As detailed in the results, the mean reduction in velocity for most tasks was around 5km/h.

b. It should be noted that the extent to which attentional absorption, as opposed to cognitive processing load, might be a distraction issue with any of the tasks, was not directly assessed in this research design.

6. Orienting Response - Many drivers exhibit what might be termed an Orienting Response (OR) – with glances to the visual feedback and /or perceived microphone location – when using a voice interface.

   a. In general, the longer the interaction with the voice system, the more prominent the OR response, likely contributing to longer TEORT.

   b. This capturing of the driver's visual attention, even when the driver is speaking a response to the system, was more prominent in older participants.

7. Speech Recognition Quality - The system's speech recognition engine performed very well. Only 2 out of more than 90 participants who were introduced to the voice system under parking lot conditions had to be excluded due to the system not recognizing the voice.

Eye Tracking & Eye Glance Metrics

As discussed earlier in this report, the reliability of portions of the automated eye tracking data were compromised due to variable lighting conditions encountered on-road, inherent limitations of the eye tracking system, and other considerations. Thus the automated eye glance data was deemed not suitable for analysis. Consequently, manually coded eye glance behavior (frame-by-frame review of video) was used to produce the eye glance metrics in this report.

Initially, a single coder approach was used, but to insure a high degree of reliability, we eventually moved to double coding with mediation for the entire dataset.

In order to provide a useful perspective for understanding/comparison of the visual demand associated with voice interfaces, we have compared our glance measure findings to the thresholds identified in the Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices (NHTSA, 2013). This comparison is not intended to indicate any violation of the guideline as, in reference to voice interface systems, it is explicitly stated that they "are currently not applicable to the auditory-vocal portions of human-machine interfaces of electronic devices."

Given that NHTSA has explicitly emphasized that the guidelines are voluntary recommendations, and that there is significant need for ongoing research to determine if the guidelines should change as the science in this area evolves, we considered a number of alternative analyses of the data to assess the impact on the overall pattern of findings. These included looking at both NHTSA's eyes off-the-forward-roadway metric and the glance-to-device metric originally adopted by the Alliance, consideration of NHTSA's 12 second total glance time criterion and the Alliance's 20 second criterion (2.1 A); use of the mean value across two repetitions of each task during the on-road drive as well as the first and second trials alone, and only error-free trials; and consideration of values for the entire sample as well as younger and older drivers as distinct groups. While some of the alternate analyses influence whether certain tasks fall immediately above or below a particular cut-point in our sample, the overall pattern is quite consistent. Moreover, no matter what analysis is employed, full address entry into the navigation system does not meet the total glance time criterion points for both NHTSA's visual manual guidelines and The Alliance 20 second threshold. This strongly suggests the need to consider the visual aspects of demand in voice-interfaces since they can clearly be multi-modal in nature (i.e. include meaningful levels of visual demand). If visual demand is formally assessed as part of the design process, the data collected here make it apparent that the thresholds used in assessing classical visual-manual interfaces are likely to be problematic for some existing voice-command systems. It would thus appear that future multi-modal system designs will need to be modified accordingly to meet classic visual demand standards and/or the TEORT variable will need to be reexamined in light of the multi-step and extended task time characteristics of such multi-modal interfaces.

Limitations / Generalizability

It is unknown if, or to what extent, exceedance of the total glance time criteria in either the Alliance or NHTSA visual-manual distraction guidelines represent a safety risk. Comparison to these established metrics is seen as informative, not necessarily predictive. It is also not clear to what extent a given level of visual demand distributed across longer task periods often characteristic of voice interactions are associated with defined long duration glance risk measures (Klauer, et al., 2006) or current concepts around "inopportune glances" (Victor & Dozza, 2011). What is apparent is that use of the voice-command interface assessed here, running under the default settings for entering a full address into the navigation system, had a very high level of visual demand.

There is no a priori reason to assume that the issues with visual demand observed here are unique to the specific voice-command interface tested in this study. Other systems employing similar design characteristics may demonstrate similar issues. It is also unknown to what extent, if any, these findings are indicative of a safety relevant increase in crash risk. Future naturalistic and/or epidemiological research will be necessary to gauge the degree to which interaction with systems such as those studied here present any significant elevation in actual risk. No crash or near-crashes were observed during data collection, nor were there aberrant vehicle kinematics (e.g., accelerations > 0.5g) recorded.

An important aspect of the NHTSA visual-manual guidelines is the need to consider a representative sample of drivers in assessing an HMI design, particularly in terms of age and gender. We were well aware of the potential significance of age and gender and, in developing the study design prior to the release of the draft guidelines, specifically choose to study participants in their 20's and 60's so as to consider younger but experienced drivers likely to be relatively more comfortable with emerging technologies and an equal number of drivers old enough that they would likely be less experienced with daily interaction with new electronic technologies but not so old that cognitive constraints would necessarily significantly impact their ability to learn how to use a new interface if appropriately introduced to it. While we believe that the age groups selected serve this purpose, they do not represent a complete match with the age distribution called for in the NHTSA visual-manual guidelines which call for an equal sampling of participants across the age groupings of 18-24, 25-39, 40-54, and 55+ years. Thus, definitive statements about whether an HMI meets NHTSA's guidelines must, by definition, follow this age distribution.

Another variation appears in the form of the Radio Hard task that we employed. We followed the approach taken by the CAMP Driver Workload Metrics project (Angell, et al., 2006) in which the task included first turning the radio on. This makes our version of the manual radio tuning task slightly more challenging as a reference task than the NHTSA specification that calls for the radio to be on at the start of the task. It is thus possible, that in those instances in our data where the manual radio turning reference task appears to be at or above the criterion point for selected variables, it might fall below those thresholds when employing the NHTSA age groupings and the task without the requirement to first turn on the radio.

NHTSA visual-manual guidelines define a specific simulation assessment protocol with a car following situation. The degree to which field data collected under moderate density, free flowing highway conditions can be considered appropriately comparable to this driving scenario has not, to our knowledge, been addressed. Thus, while the data collected represent a real-world look at how a participant interacts with the system under study under actual driving conditions, the extent to which the data collected compare to what would be measured under simulation is open to question. Previous research suggests that the overall processing demands of actual driving a real vehicle are higher than driving a simulator (Reimer & Mehler, 2011), while the allocation of visual attention to secondary activities is more constant (Wang, et al., 2010).

As in most experimental settings, participants were requested in this study to perform a set of activities of interest to the research project. In this instance, particular care was taken in selecting a set of activities representative of tasks drivers often wish to undertake while underway and to "pace" the activates with appropriate rest intervals between tasks. It is unknown to what extent drivers would actually engage in the activities of interest if these interface systems and options were available to them under normal driving conditions.

Finally, the level of skill with an interface often increases with time. The data generated as part of this experimental assessment provides limited indication of the degree to which learned behavior may adapt over time or by which individuals may identify more intuitive ways of operating the system. We are currently engaged in steps to address some of these and other considerations as described below.

Next Steps

While we intend to delve deeper into the existing dataset and develop academic publications for peer review and critique, we feel strongly that no single study can provide a definitive

evaluation of a particular technology or a full understanding of how humans interact with a general class of technology. A systematically developed body of research is needed to even approach these ideal goals. It is our hope that this report will stimulate productive work by other researchers, and we are also fortunate to have support to continue work in this area. A follow-on study is currently intended to address several possible critiques of the original design. These include the recruitment of a sample that directly corresponds to NHTSA's recommendations for age distribution (e.g. equal distribution of participants across the age groupings of 18-24, 25-39, 40-54, and 55+ years) and a partially revised design that employs a slightly less demanding version of the manual radio tuning reference task (beginning with the radio in the "on" position, per NHTSA visual-manual guidelines). Results from this second study should be available shortly. In addition, funding was recently obtained from the Santos Family Foundation to add an arm to the study that explores the impact of using the "expert" modes of the voice system reduce the amount of auditory prompting and the number of confirmatory responses required of the driver. It is conceivable that this optional mode of operating the system might result in a reduction in task time and some of the visual orienting to the display screen. Once the second study is complete, we will further explore the question of generalizability by expanding our study of production voice system implementations to include at least two other vehicle brands.

It is hoped that the findings from this overall project will be useful in informing the development of voice system implementations that optimize the potential of this compelling interface concept.

Version Notes

The initial version of this technical report (2013-17) dated November 4, 2013 was given limited release for background briefings on this work. The current version (2012-17A) includes a refined description of the voice systems considered in this study based on feedback from representatives of the vehicle manufacturer along with minor typographical and style corrections. In addition, hyperlinks have been added to videos demonstrating each of the tasks under the actual driving conditions employed in the study.

Acknowledgements

Acknowledgement is extended to The Santos Family Foundation and US DOT's Region I New England University Transportation Center at MIT for providing the support for the initiation of this project. This funding provided support for the project's conceptual development, instrumentation of the research vehicle, as well as initially planned data collection and

preliminary reporting on the project (Reimer, et al., 2013). The vehicle itself was purchased through funding from Ford Motor Company for an earlier project assessing the Ford Active Park Assist™ feature (Reimer, Mehler, & Coughlin, 2010). The current project was conducted without consultation or involvement of Ford Motor Company.

This work would not have been possible without the support of AgeLab staff and visiting scholars including: Hale McAnulty, Daniel Munger, Alea Mehler, Erin McKissick, Enrique Abdon Garcia Perez, Adrian Rumpold, Thomas Manhardt, Yutao Ba, Yan Yang, Ying Wang, Brahmi Pugh, Martin Lavalliere and Brendan Drischler in the development of the protocol, collection of data, and exhaustive reduction and coding of eye glance and other data. In addition, we are grateful for the valuable, constructive comments of James Foley and Kazutoshi Ebe of CSRC provided during the development of the study.

The interpretive aspects of this report reflect the views of the authors, who are also responsible for the factualness and accuracy of the information presented herein. This document is disseminated under the sponsorship noted above.

References

Angell, L., Auflick, J., Austria, P. A., Kochhar, D., Tijerina, L., Biever, W., et al. (2006). Driver Workload Metrics Task 2 Final Report. Washington, DC: U.S. Department of Transportation National Highway Traffic Safety Administration.

Barón, A., & Green, P. (2006). Safety and usability of speech interfaces for in-vehicle tasks while driving: a brief literature review. Ann Arbor, MI: The University of Michigan Transportation Research Institute (UMTRI).

Brookhuis, K. A., & de Waard, D. (1993). The use of psychophysiology to assess driver status. *Ergonomics, 36(9),* 1099-1110.

Burns, P., Harbluuk, J., Foley, J., & Angell, L. (2010). The importance of task duration and related measures in assessing the distraction potential of in-vehicle tasks. Proceedings of the Second International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2010), November 11-12, 2010, Pittsburgh, PA, USA.

Carter, C., & Graham, R. (2000). Experimental comparison of manual and voice controls for the operation of in-vehicle systems. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 3-286-283-289.

Chiang, D. P., Brooks, A. M., & Weir, D. H. (2005). Comparison of visual-manual and voice interaction with contemporary navigation system HMIs. SAE Technical Paper 2005-01-0433.

Dalton, P., Agarwal, P., Freankel, N., Baichoo, J., & Masry, A. (2013). Driving with navigational instructions: Investigating user behaviour and performance. *Accident Analysis & Prevention, 50,* 298-303.

Dopart, C., Häggman, A., Thornberry, C., Mehler, B., Dobres, J., & Reimer, B. (2013 in press). A driving simulation study examining destination entry with iPhone iOS 5 Google Maps and a Garmin portable GPS system. Proceedings of the 57th Annual Meeting of the Human Factors and Ergonomics Society, San Diego, CA.

Driver Focus-Telematics Working Group. (2006). Statement of principles, criteria and verification procedures on driver interactions with advanced in-vehicle information and communication systems, Version 2.0: Alliance of Automotive Manufacturers.

Dybkjær, L., Bernsen, N. O., & Minker, W. (2004). Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication, 43(1-2),* 33-54.

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F, 8(2),* 97-120.

Fitch, G. A. Soccolich, S.A., Guo, F., McClafferty, J., Fang, Y., Olson, R. L., Perez, M. A., et al. (2013). The impact of hand-held and hands-free cell phone use on driving performance and safety-critical event risk (Report No. DOT HS 811 757). Washington, DC: National Highway Traffic Safety Administration (NHTSA).

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini Mental State: A practical method for grading the cognitive state for the clinician. *Journal of Psychiatry Research, 12(3),* 189-198.

Forlines, C., Schmidt-Nielsen, B., Raj, B., Wittenburg, K., & Wolf, P. (2005). A comparison between spoken queries and menu-based interfaces for in-car digital music selection. Proceedings of the Human-Computer Interaction-INTERACT 2005, 536-549.

Garay-Vega, L., Pradhan, A. K., Weinberg, G., Schmidt-Nielsen, B., Harsham, B., Shen, Y., et al. (2010). Evaluation of different speech and touch interfaces to in-vehicle music retrieval systems. *Accident Analysis & Prevention, 42(3),* 913-920.

Gärtner, U., König, W., & Wittig, T. (2001). Evaluation of manual vs. speech input when using a driver information system in real traffic. Proceedings of the International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, 7-13.

Gellatly, A. W., & Dinges, T. A. (1998). Speech recognition and automotive applications: using speech to perform in-vehicle tasks. Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting, 1247-1251.

Geutner, P., Steffens, F., & Manstetten, D. (2002). Design of the VICO Spoken dialogue system: evaluation of user expectations by wizard-of-oz experiments. Proceedings of the LREC.

Graham, R., & Carter, C. (2001). Voice dialling can reduce the interference between concurrent tasks of driving and phoning. *International Journal of Vehicle Design, 26(1),* 30-47.

Greenberg, J., Tijerina, L., Curry, R., Artz, B., Cathey, L., Kochhar, D., et al. (2003). Driver distraction: evaluation with event detection paradigm. Transportation Research Reccord, 1843, 1-9.

Grothkopp, D., Krautter, W., Grothkopp, B., Steffens, F., & Geutner, F. (2001). Using a driving simulator to perform a Wizard-of-Oz experiment on speech-controlled driver information systems. Proceedings of the 1st Human-Centered Transportation Simulation Conference.

Harbluk, J., Burns, P. C., Lochner, M., & Trbovich, P. L. (2007). Using the lane-change test (LCT) to assess distraction: Tests of visual-manual and speech-based operation of navigation system interfaces. Proceedings of the 4th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Stevenson, WA.

Harbluk, J., & Lalande, S. (2005). Performing e-mail tasks while driving: The impact of speech-based tasks on visual detection. Proceedings of the 3rd International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Rockport, ME.

Hart, S. G., & Staveland, L. E. (1988). Development of the NASA-TLX (Task Load Index): results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), Human mental workload. Advances in Psychology (pp. 139-183). Oxford England: North-Holland.

Horberry, T., Anderson, J., Regan, M. A., Triggs, T. J., & Brown, J. (2006). Driver distraction: the effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. *Accident Analysis and Prevention, 38(1),* 185-191.

Horrey, W. J., & Wickens, C. D. (2006). Examining the impact of cell phone conversations on driving using meta-analytic techniques. *Human Factors, 48(1),* 196-205.

Hu, J., Winterboer, A., Nass, C. I., Moore, J. D., & Illowsky, r. (2007). Context & usability testing: user-modeled information presentation in easy and difficult driving conditions. Proceedings of the CHI, San Jose, CA., 1343-1346.

Itoh, K., Miki, Y., Yoshitsugu, N., Kubo, N., & Mashimo, S. (2004). Evaluation of a voice-activated system using a driving simulator. SAE Technical Paper 2004-01-0232. doi: 10.4271/2004-01-0232

Jamson, A. H., Westerman, S. J., Hockey, G. R. J., & Carsten, O. M. J. (2004). Speech-based E-mail and driver behavior: Effects of an in-vehicle message system interface. *Human Factors, 46(4),* 625-639.

Jensen, B. S., Skov, M. B., & Thiruravichandran, N. (2010). Studying driver attention and behaviour for three configurations of GPS navigation in real traffic driving. Proceedings of the 28th international conference on Human factors in computing systems, 1271-1280.

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). Washington, DC: United States Department of Transportation, National Highway Traffic Safety Administration.

Kun, A., Paek, T., & Medenica, Z. (2007). The effect of speech interface accuracy on driving performance. In INTERSPEECH (pp. 1326-1329).

Lee, J. D., Caven, B., Haake, S., & Brown, T., L. (2000). Are conversations with your car distracting? understanding the promises and pitfalls of speech-based interfaces. SAE Technical Paper 2000-01-C012.

Lee, J. D., Caven, B., Haake, S., & Brown, T. L. (2001). Speech-based interaction with in-vehicle computers: the effect of speech-based e-mail on drivers' attention to the roadway. *Human Factors, 43(4),* 631.

Lenneman, J. K., & Backs, R. W. (2009). Cardiac autonomic control during simulated driving with a concurrent verbal working memory task. *Human Factors, 53(3),* 404-418.

Lerner, N., Singer, J., & Huey, R. (2008). Driver strategies for engaging in distracting tasks using in-vehicle technologies (Report No. HS DOT 810 919). Washington, DC: U.S. Department of Transportation National Highway Traffic Safety Administration (NHTSA)

Liu, Y., Schreiner, C. S., & Dinges, D. T. (1999). Development of human factors guidelines for advanced traveler information systems (ATIS) and commercial vehicle operations (CVO): Human Factors Evaluation of the Effectiveness of Multi-Modality Displays in Advanced Traveler Information Systems (FHWA-RD-96-150). McLean, VA: Office of Safety and Traffic Operations R&D Federal Highway Administration.

Lo, V.E-W., & Green, P.A. (2013). Development and evaluation of automotive speech interfaces: useful information from the human factors and the related literature. *International Journal of Vehicular Technology, 2013,* ID 924170. http://dx.doi.org/10.1155/2013/924170

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1(4),* 476-490. doi: 10.3758/BF03210951

Maciej, J., & Vollrath, M. (2009). Comparison of manual vs. speech-based interaction with in-vehicle information systems. *Accident Analysis & Prevention, 41(5),* 924-930.

Mazzae, E. N., Ranney, T., Watson, G. S., & Wightman, J. A. (2004). Hand-held or hands-free? the effects of wireless phone interface type on phone task performance and driver preference. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2218-2222.

McCallum, M. C., Campbell, J. L., Richman, J. B., Brown, J. L., & Wiese, E. (2004). Speech recognition and in-vehicle telematics devices: Potential reductions in driver distraction. *International Journal of Speech Technology, 7(1),* 25-33.

McCann Erickson. (2013). Find New Roads Retrieved May 29, 2013, from http://www.ispot.tv/ad/7oVl/chevrolet-sonic-with-siri-buttons

Mehler, B., & Reimer, B. (2013). An initial assessment of the significance of task pacing on self-report and physiological measures of workload while driving. Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Bolton Landing, New York, 170-176.

Mehler, B., Reimer, B., & Coughlin, J. F. (2012). Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: an on-road study across three age groups. *Human Factors, 54(3),* 396-412. doi: 10.1177/0018720812442086

Mehler, B., Reimer, B., Coughlin, J. F., & Dusek, J. A. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record(2138),* 6-12. doi: 10.3141/2138-02

Nasreddine, Z. S., Phillips, N. A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society, 53(4),* 695-699.

National Highway Traffic Safety Administration. (2013). Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices (Docket No. NHTSA-2010-0053). Washington, DC: U.S. Department of Transportation National Highway Traffic Safety Administration (NHTSA), .

Neurauter, M. L., Hankey, J. M., Schalk, T. B., & Wallace, G. (2009). Outbound texting: comparison of speech-based approach and handheld touch-screen equivalent. *Transportation Research Record: Journal of the Transportation Research Board, 2321,* 23-30.

Östlund, J., Nilsson, L., Carsten, O., Merat, M., Jamson, H., Jamson, S., et al. (2004). Human machine interface and the safety of traffic in europe (HASTE) project deliverables 2: HMI and safety-related driver performance.

Östlund, J., Peters, B., Thorslund, B., Engström, J., Markkula, G., Keinath, A., et al. (2005). Adaptive Integrated Driver-Vehicle Interface (AIDE): Driving performance assessment -

methods and metrics. (Report No. IST-1-507674-IP). Gothenburg, Sweden: Information Society Technologies (IST) Programme.

Owens, J. M., McLaughlin, S. B., & Sudweeks, J. (2010). On-road comparison of driving performance measures when using handheld and voice-control interfaces for mobile phones and portable music players. SAE International Journal of Passenger Cars – Mechanical Systems, 3(1), 734-743.

Owens, J. M., McLaughlin, S. B., & Sudweeks, J. (2011). Driver performance while text messaging using handheld and in-vehicle systems. *Accident Analysis & Prevention, 43(3),* 939-947.

R Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Ranney, T. A., Baldwin, G. H. S., Parmer, E., Domeyer, J., Martin, J., & Mazzae, E. N. (2011). Developing a test to measure distraction potential of in-vehicle information system tasks in production vehicles (Report No. DOT HS 811 463). Washington, DC: U.S. Department of Transportation National Highway Traffic Safety Administration (NHTSA).

Ranney, T. A., Harbluk, J. L., & Noy, Y. I. (2005). Effects of voice technology on test track driving performance: implications for driver distraction. *Human Factors, 47(2),* 439-454.

Ranney, T. A., Mazzae, E. N., Baldwin, G. H. S., & Salaani, M. K. (2007). Characteristics of voice-based interfaces for in-vehicle systems and their effects on driving performance (Report No. DOT-HS-810-867). Washington, DC: U.S. Department of Transportation National Highway Traffic Safety Administration (NHTSA).

Reimer, B. (2009). Cognitive task complexity and the impact on drivers' visual tunneling. Proceedings of the Transportation Research Board of The National Academies, Washington, DC, 13-19.

Reimer, B., & Mehler, B. (2011). The impact of cognitive workload on physiological arousal in young adult drivers: a field study and simulation validation. *Ergonomics, 54(10),* 932-942.

Reimer, B., Mehler, B., & Coughlin, J. F. (2010). An evaluation of driver reactions to new vehicle parking assist technologies developed to reduce driver stress (MIT AgeLab White Paper). Cambridge, MA: Massachusetts Institute of Technology.

Reimer, B., Mehler, B., Donmez, B., Pala, S., Wang, Y., Olson, K., et al. (2012). A driving simulator study examining phone dialing with an iphone vs. a button style flip- phone.

Proceedings of the 56th Annual Meeting of the Human Factors and Ergonomics Society, Boston, MA, 2191-2195.

Reimer, B., Mehler, B., Matteson, S., Levantovsky, V., Chahine, N., Gould, D., Wang, Y., Mehler, A., McAnulty, H., Mckissick, E., Greve, G. & Coughlin, J.F. (2012). An exploratory study on the impact of typeface design in a text rich user interface on off-road glance behavior. *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutoUI 2012),* Portsmouth, NH, October 17-19, 2012, pp. 25-32.

Reimer, B., Mehler, B., McAnulty, H., Munger, D., Mehler, A., Perez, E. A. G., et al. (2013). A preliminary assessment of perceived and objectively scaled workload of a voice-based driver interface. Proceedings of the Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Bolton Landing, NY, 537-543.

Reimer, B., Mehler, B., Wang, Y., & Coughlin, J. F. (2012). A field study on the impact of variations in short-term memory demands on drivers' visual attention and driving performance across three age groups. *Human Factors, 54(3),* 454-468. doi: 10.1177/0018720812437274

Roscoe, A. H. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration. *Biological Psychology, 34(2-3),* 259-287.

Shutko, J., Mayer, K., Laansoo, E., & Tijerina, L. (2009). Driver workload effects of cell phone, music player, and text messaging tasks with the Ford SYNC voice interface versus handheld visual-manual interfaces SAE Technical Paper 2009-01-0786. doi: 10.4271/2009-01-0786

Shutko, J., & Tijerina, L. (2011). Ford's Approach to Managing Driver Attention: SYNC and MyFord Touch. *Ergonomics in Design, 19(4),* 13-16.

Smith, D. L., Chang, J., Glassco, R., Foley, J., & Cohen, D. (2005). Methodology for capturing driver eye glance behavior during in-vehicle secondary tasks. *Transportation Research Record: Journal of the Transportation Research Board, 1937(1),* 61-65.

Son, J., Mehler, B., Lee, T., Park, Y., Coughlin, J. F., & Reimer, B. (2011). Impact of cognitive workload on physiological arousal and performance in younger and older drivers. Proceedings of the Sixth International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Lake Tahoe, CA, 87-94.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J., Medeiros-Ward, N., & Biondi, F. (2013). Measuring cognitive distraction in the Automobile. Washington, DC: AAA Foundation for Traffic Safety.

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

Victor, T., & Dozza, M. (2011). Timing matters: Visual behavior and crash risk in the 100-Car on-line data. Proceedings of the Driver Distraction and Inattention Conference (SAFER – Vehicle and Traffic Safety Centre at Chalmers).

Wang, Y., Mehler, B., Reimer, B., Lammers, V., D'Ambrosio, L., & Coughlin, J. (2010). The validity of driving simulation for assessing differences between in-vehicle informational interfaces: A comparison with field testing. *Ergonomics, 53(3),* 404-420. doi: 10.1080/00140130903464358

Wilson, G. F. (2002). Psychophysiological test methods and procedures. In S. G. Charlton & T. G. O'Brien (Eds.), Handbook of Human Factors Testing and Evaluation (pp. 127-156). Mahwah, NJ: Lawrence Erlbaum Associates.

Yager, C. (2013). An Evaluation of The effectivness of voice-to-text programs at reducing incidences of distracted driving. College Station, Texas: Texas A&M Transportation Institute.

Zhang, J., Borowsky, A., Schmidt-Nielsen, B., Harsham, B., Weinberg, G., Romoser, M. R., et al. (2012). Evaluation of two types of in-vehicle music retrieval and navigation systems. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 1992-1996.

ABOUT THE AUTHORS

Bryan Reimer, Ph.D.

Bryan Reimer is a Research Engineer in the Massachusetts Institute of Technology AgeLab and the Associate Director of the New England University Transportation Center. His research seeks to develop new models and methodologies to measure and understand human behavior in dynamic environments utilizing physiological signals, visual behavior monitoring, and overall performance measures. Dr. Reimer leads a multidisciplinary team of researchers and students focused on understanding how drivers respond to the increasing complexity of the operating environment and on finding solutions to the next generation of human factors challenges associated with distracted driving, automation and other in-vehicle technologies. He directs work focused on how drivers across the lifespan are affected by in-vehicle interfaces, safety systems, portable technologies, different types and levels of cognitive load. Dr. Reimer is an author on over 70 peer reviewed journal and conference papers in transportation. Dr. Reimer is a graduate of the University of Rhode Island with a Ph.D. in Industrial and Manufacturing Engineering.

reimer@mit.edu
(617) 452-2177
http://web.mit.edu/reimer/www/

Bruce Mehler, M.A.

Bruce Mehler is a Research Scientist in the Massachusetts Institute of Technology AgeLab and the New England University Transportation Center, and is the former Director of Applications & Development at NeuroDyne Medical Corporation. He has an extensive background in the development and application of non-invasive physiological monitoring technologies and research interests in workload assessment, individual differences in response to cognitive demand and stress in applied environments, and in how individuals adapt to new technologies. Mr. Mehler is an author of numerous peer reviewed journal and conference papers in the biobehavioral and transportation literature. He continues to maintain an interest in health status and behavior from his early work in behavioral medicine. He received an MA in Psychology from Boston University and a BS degree from the University of Washington.

bmehler@mit.edu
(617) 253-3534
http://agelab.mit.edu/bruce-mehler

Jonathan Dobres, Ph.D.

Jonathan Dobres is a postdoctoral Research Associate at the MIT AgeLab. Dr. Dobres's research interests include human-computer interaction, user experience design, visual attention, and visual learning. He received a BA, MA, and PhD in Psychology (Brain, Behavior, and Cognition) from Boston University. His research examined how visual perception changes with training. He has also worked for the Traumatic Brain Injury Model System at Spaulding Rehabilitation Hospital, part of a long-term national study on the effects of traumatic brain injuries. Dr. Dobres's current research primarily concerns the visual and cognitive demands of performing tasks while driving, as well as how the visual properties of in-vehicle interfaces affect usability and driver performance.

jdobres@mit.edu
(617) 253-7728

Joseph F. Coughlin, Ph.D.

Joseph F. Coughlin is founder and Director of the Massachusetts Institute of Technology AgeLab and Director of the US Department of Transportation's Region I New England University Transportation Center. He served as the Chair of the Organization for Economic Cooperation & Developments 21-nation Task Force on Technology and Transportation for Older Persons, is a member of the National Research Council's Transportation Research Board Advisory Committee on the Safe Mobility of Older Persons. He served as a Presidential appointee to the White House Conference on Aging and has consulted or served on technology and design boards for BMW, Daimler, Nissan, and Toyota. Prior to joining MIT, Dr. Coughlin led the transportation technical services consulting practice for EG&G a global Fortune 1000 science and technology firm.

coughlin@mit.edu
(617) 253-3534
http://www.josephcoughlin.com/

About the New England University Transportation Center & MIT Center for Transportation & Logistics

The New England University Transportation Center is a research, education and technology transfer program sponsored by the US Department of Transportation. Together the faculty, researchers and students sponsored by the New England Center conduct work in partnership with industry, state & local governments, foundations and other stakeholders to address the future transportation challenges of aging, new technologies and environmental change on the nation's transportation system. For more information about the New England University Transportation Center, visit utc.mit.edu. For more information about the US Department of Transportation's University Transportation Centers Program, please visit www.rita.dot.gov/utc/. The New England Center is based within MIT's Center for Transportation & Logistics, a world leader in supply chain management education and research. CTL has made significant contributions to transportation and supply chain logistics and helped numerous companies gain competitive advantage from its cutting edge research. For more information on CTL, visit ctl.mit.edu.

About the AgeLab

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