U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**

NHTSA

DOT HS 812 813

January 2020

In-Vehicle Voice Control Interface Evaluation: Preliminary Driver Workload and Risk Analysis

DISCLAIMER

Suggested APA Format Citation:

Jenness, J. W., Boyle, L. N., Lee, J. D., Miller, E. E., Yahoodik, S., Huey, R,...Petraglia, E. (2020, January). *In-vehicle voice control interface evaluation: Preliminary driver workload and risk analysis* (Report No. DOT HS 812 813). Washington, DC: National Highway Traffic Safety Administration.

1. Report No. DOT HS 812 813	2. Government Accession No.	3. Recipient's Catalog No.	
4.Title and Subtitle In-Vehicle Voice Control Interface Evaluation: Preliminary Driver Workload and Risk Analysis		5.Report Date January 2020	
		6.Performing Organization Code	
7.Authors James W. Jenness,[1] Linda Ng Boyle,[2] John D. Lee,[3] Erika E. Miller,[2] Sarah Yahoodik,[1] Richard Huey,[1] Ja Young Lee,[3] Amy K. Benedick,[1] Elizabeth Petraglia[1]		8.Performing Organization Report No.	
9.Performing Organization Name and Address 1. Westat, Inc. 1600 Research Boulevard Rockville, MD 20850 2. University of Washington (Seattle, WA) 3. University of Wisconsin-Madison (Madison, WI)		10. Work Unit No. (TRAIS) 11.Contract or Grant No. DTNH22-11-D-00237/0013	
12. Sponsoring Agency Name and Address National Highway Traffic Safety Administration 1200 New Jersey Avenue SE Washington, DC 20590		13. Type of Report and Period Covered Final Report	
		14.Sponsoring Agency Code	
15.Supplementary Notes The NHTSA Contracting Office Representative was Thomas Fincannon.			

16.Abstract

This project evaluated distraction and relative risk associated with using voice control systems (VCS) while driving. The objective was to explore potential empirical measures and to use a modeling approach for evaluating risk with these voice-based systems. The project included three studies. Study 1 and Study 2 were designed to assess potential measures of the workload and demands on the driver imposed by voice-based and hybrid (audio plus visual) tasks. Participants interacted with a "Wizard of Oz" VCS while driving and a novel radio tuning benchmark task was used. Study 1 (n = 9) compared response times for the tactile detection response task (TDRT), which is a standardized measure of cognitive load, with a modified remote detection response task (RDRT) which was implemented in this study to provide a complementary measure of visual attention toward the forward roadway. Response time was sensitive to differences between VCS tasks, and there was a significant interaction between detection response task (DRT) type and VCS task. Study 2 (n = 9) included both on-road data collection and data collection with a driving simulator. In three different sessions, measurement protocols included TDRT, response time to the lead vehicle (LV) brake light, speed matching to a speed-varying LV, and off-road eye glance measures. Results indicated that task completion time, TDRT performance, and eye glance measures distinguished between VCS tasks and provided similar results in the on-road and simulated driving contexts. Study 3 used an analytical approach to develop relative risk estimates and crash severity estimates for the VCS tasks tested in this project. Counterfactual -- or "what if" -- simulations made use of a set of crash events and near crash events recorded in the second Strategic Highway Research Program (SHRP 2) naturalistic driving study. These events were then replayed as a set of simulations, where the eye glance data collected in Study 1 and Study 2 were substituted for the original eye glance data recorded from drivers in the SHRP 2 study. Risk estimates and crash severity estimates developed using this technique varied considerably by VCS task and by driver.

17.Key Words Driver distraction, in-vehicle technologies, voice control system, driver vehicle interface, counterfactual simulation, on-road, driving simulator, N-back task	18.Distribution Statement This document is available to the public through the National Technical Information Service, www.ntis.gov.		
19.Security Classif.(of this report) Unclassified	20.Security Classif.(of this page) Unclassified	21.No. of Pages 103	22.Price

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

TABLE OF CONTENTS

ABBREVIATIONS, ACRONYMS, AND SYMBOLS

ANOVA	analysis of variance
CDT	collision detection task
DRT	detection response task
EOR	eyes-off-road
HMI	human machine interface
HSD	honest significance difference
ICC	intercounty connector
ISO	International Standards Organization
LV	lead vehicle
LS	least squares
MGD	mean glance duration
NADS	National Advanced Driving Simulator
ND	neutral density
NDS	Naturalistic Driving Study
POI	point of interest
RDRT	Remote [Visual] Detection Response Task
SDLP	Standard Deviation of Lateral Position
SHRP 2	Second Strategic Highway Research Program
TDRT	Tactile Detection Response Task
TEORT	total eyes-off-road time
UW	University of Washington
UW-Madison	University of Wisconsin-Madison
VCS	voice control system

1 EXECUTIVE SUMMARY

Auditory-vocal user interfaces, also called voice control systems, are a common feature of in-vehicle human machine interfaces. The National Highway Traffic Safety Administration initiated a research program to understand the potential safety impact of using VCS-enabled features in vehicles. A goal of this research program is to relate measures of cognitive and visual distraction to crash risk for the evaluation of VCS. The project described in this report builds upon previous research regarding the evaluation of VCS for potential driver distraction. Specifically, this project seeks to evaluate how to use various detection response measures of cognitive load with measures of visual attention (e.g., eye glance measures) for evaluating potential driver distraction from performing VCS tasks. This project also links test data on VCS to provide risk estimates for drivers engaged in VCS tasks.

The project included three related studies.

- Study 1 (Dual DRT). This study examined a dual detection response tasks protocol for its potential to detect both cognitive and visual distraction caused by performing VCS tasks. Variants of the DRT have been standardized (ISO, 2015) to assess changes in the driver's workload caused by performing secondary (non-driving) tasks while driving. One variant of the DRT uses a tactile DRT stimulus, and another variant uses a remote visual DRT stimulus. The driver's response time to the onset of these stimuli is measured with tasks that are thought to be sensitive to the driver's cognitive load. The visual load imposed on the driver from performing secondary tasks is often assessed by recording and analyzing the driver's eye glances. In practice, eye glance analysis is time consuming and eye tracking can be difficult under natural lighting conditions. It would be useful to have a DRT that could assess visual distraction so that eye glance analysis would not be needed. The standard RDRT stimulus is not particularly well suited for this, because the visual stimulus often can be detected even during off-road glances using peripheral vision. For Study 1, researchers developed a modified RDRT stimulus that is much more difficult to detect using peripheral vision. Thus, performance with the new RDRT should be sensitive to glances the driver makes away from the forward roadway. In Study 1, researchers used a modified RDRT protocol and the TDRT protocol together with the same set of participants to assess the potential for both cognitive and visual distraction from VCS tasks. A second purpose of Study 1 was to examine a radio tuning task that uses voice interface inputs. Prior research used manual radio tuning with a knob as a baseline task for the NHTSA Phase 1 Driver Distraction Guidelines (NHTSA, 2013), so the current study supplements that research by exploring radio tuning tasks that use voice interfaces for future considerations.

- Study 2 (Testing context). The purpose of Study 2 was to compare laboratory data using a driving simulator to field data collected from a real vehicle on a limited access highway. This study compared several potential evaluation measures for VCS tasks to determine which were the most robust to changes in the testing context. Three test protocols were examined across the two testing contexts. These included: (a) TDRT to provide measures of cognitive load, (b) following a speed-varying lead vehicle, and (c) responding as quickly as possible to a LV brake light.

- Study 3 (Relative risk). The purpose of study 3 was to develop an analytical framework for estimating relative crash risk for VCS tasks. This study examined the VCS evaluation measures of visual attention (glance measures) to their safety relevance and potential crash risk by combining glance data from Study 1 and Study 2 with rear end crash and near crash scenarios from the second Strategic Highway Research Program (SHRP 2) naturalistic driving study. Statistical modeling, using counterfactual simulation, was used to explore a multitude of "what if" scenarios. Study participants' glance patterns for each VCS task were applied to the set of SHRP 2 scenarios, and from the outcome of these simulations (crash versus no crash), relative crash risk and crash severity were estimated.

In this project, participants performed an Easy Radio Tuning task as a baseline VCS task (see Section 3.2.3 for more details). Researchers selected this task to represent a less demanding set of typical interactions between the driver and infotainment system, which do not require manual input. As expected, participants completed this task more quickly and accurately than the other VCS tasks studied.

The driving task used in Study 1 was not very demanding (i.e., following a LV on a straight road). The two variants of DRT exercised in the study did not significantly differ from each other in terms of their influence on driving performance (i.e., standard deviation of lateral position [SDLP], SD speed) in the simulator.

Participants' responses on DRT were evaluated using response time and miss rate to determine which measures may be sensitive to differences in the demands imposed by different VCS tasks. The response time measure was sensitive to differences between VCS tasks, and overall the TDRT and RDRT response times were similar. However, there was a significant interaction between DRT type and VCS task. On most VCS tasks, response times for TDRT were shorter than, or similar to, response times for RDRT. However, for the Easy Hybrid Radio Tuning task, TDRT response times were longer than RDRT response times. We have no explanation for this difference.

For miss rates, there were no significant differences between TDRT and RDRT protocols. However, the miss rate was sensitive to differences between Audio Only, Hybrid (Audio and Visual display), and 1 – Back voice tasks.

Driving performance measures, including standard deviation of lane position and standard deviation of speed were sensitive to differences between VCS voice task types. Study 1 also found that Hybrid VCS tasks, which included a visual display, were completed more quickly and more accurately than similar Audio Only tasks. This may be explained by the fact that the Audio Only tasks required time for the full list of menu options to be presented in a serial fashion. The visual displays used for the Hybrid tasks presented the full set of options to the driver simultaneously. Another aspect of the Audio Only tasks was that the presentation and pacing of information was not under the control of the driver. When the visual display was present for Hybrid tasks, the driver could choose to get information more quickly by making glances at the display. Consistent with these results, participants reported the Audio Only tasks as being more demanding than the comparable Hybrid tasks.

Eye glances recorded (see Section 3.2.7.2 for details) in the study were analyzed with metrics suggested by NHTSA Driver Distraction Guidelines (78 FR 24817, 2013). Total eyes off road durations were greater for TDRT than for RDRT on all voice tasks except the most complex task (Navigation Hard). This suggests that the modified RDRT task used in this study may encourage participants to focus visual attention in the forward direction.

In Study 2, researchers compared laboratory data collected using a driving simulator to data collected from driving a real vehicle on a limited access highway. This study compared several potential evaluation measures for VCS tasks to determine which were the most robust to changes in the testing context. Three test protocols were examined across the two testing contexts. These included: (a) TDRT, (b) following a speed-varying LV, and (c) responding as quickly as possible to a LV brake light.

VCS tasks plus the N-Back task (1-back) were evaluated using three different driving scenarios (test protocols) in two data collection contexts (driving simulator versus on-road). Eye glances were recorded in every scenario. In the first scenario, the participants were asked to follow a LV traveling at a fixed speed. The participants performed the set of voice tasks while driving, and cognitive load was assessed with the TDRT. The TDRT response times did not vary significantly with context, but they were sensitive to differences between tasks and distinguished between voice tasks and baseline driving with no voice task. There was no statistically significant interaction between tasks and context. Together, these findings suggest that the TDRT may be a robust test protocol.

In the second test protocol, participants followed a LV that was constantly changing its speed. They were instructed to maintain a constant following distance. Speed correlation between the LV and the participants' vehicle was calculated and cross-correlated to determine the optimal delay (lag) in LV speed that would maximize the correlation. This measure was not very reliable across task attempts and across participants for the same task, or across contexts. As a potential measure of workload imposed by secondary tasks, it did not distinguish VCS tasks from the no-task baseline driving.

In the third test protocol, participants were asked to respond as quickly as possible to illumination of the LV brake lights by tapping their own brake pedal. This brake light response measure was analogous to the modified RDRT used in Study 1, but only a single brake light signal was initiated per task. The brake light response times did not vary significantly with context, but were sensitive to differences between tasks. There was no statistically significant interaction between tasks and context. However, this measure did not distinguish between voice tasks and baseline driving with no voice task. It is possible that other differences between tasks could have been detected with the increased statistical power of a larger sample.

The glance data collected in Study 1 and Study 2 were sensitive to differences between voice tasks and were reliable across contexts. With only nine participants, a data analysis of glance criteria cannot identify small effect sizes, but we did find clear overall differences between Hybrid tasks and Voice Only tasks. The details of passing or failing individual criteria per task varied between contexts. The total eyes-off-road time (TEORT) measure for the Hybrid tasks showed an interesting trend across the three driving scenarios such that the Fixed Speed scenario

produced slightly more off-road glance time than the Variable Speed scenario, and the Variable Speed scenario produced slightly more off-road glace time than the Brake Light scenario. We speculate that the demands of the driving-related tasks were lower in the Fixed Speed scenario as compared to the other scenarios, which allowed participants to allocate more of their attention to the in-vehicle screen used for the Hybrid Voice tasks. In Study 1, TEORT was also higher in the TDRT protocol as compared to the RDRT protocol, perhaps for a similar reason. The RDRT protocol may have encouraged more visual attention toward the forward roadway.

Study 3 used modeling and simulation techniques to develop a method of estimating relative crash risk between VCS tasks from test data collected in a driving simulator or on-road. The modeling approach incorporated a set of existing data from 34 crash events and 190 near-crash events that were recorded previously during the SHRP 2 naturalistic driving study. A model of the driver's braking response to forward looming cues was used to simulate a multitude of "what if" scenarios (called counterfactual simulation) regarding the driver's glance patterns away from the forward roadway and deceleration of a LV. Each simulation resulted in a crash or no crash. Glance data obtained in Study 1 and Study 2 were incorporated into the simulations of the SHRP 2 events and the outcome of these simulations was used to estimate crash risk for the various VCS tasks tested in this project.

Substantial differences in overall crash risk were noted across participants. Study 1 data led to a higher risk of crash overall than data collected in Study 2, which may be attributed to differences in glance duration distributions collected from the two sets of participants. The data collected in Study 2 also show greater differences in crash risk between tasks where the Audio Only tasks produced essentially no crashes.

Findings from Study 3 suggest that counterfactual simulation shows promise in linking glance patterns collected during on-road or simulator-based evaluation to potential risks of the device for drivers. However, this type of analysis also has several important caveats. Most importantly, the crash rates that result from the method depend on the sample of potential crash situations. The more severe these situations, the higher the rate of crashes. Consequently, the crash rates determined from this method should be considered as indicators of relative rather than absolute crash rates. The SHRP 2 data used in this study are predominantly low-speed events and therefore might underestimate risk compared to a larger sample that includes a more complete range of initial speeds.

2 INTRODUCTION

Auditory-vocal user interfaces, also called voice control systems are prevalent in the auto and consumer electronics industries. VCS range from simple command-and-control interactions (climate control, radio, contact name call, etc.) to more complex search/text tasks (full string address entry, points-of-interest (POI) search, web search, text dictation, location-based services, apps, infotainment, etc.). VCS are already a common feature of in-vehicle human-machine interfaces (HMI). The National Highway Traffic Safety Administration initiated a VCS research program to understand the breadth and impact of using VCS-enabled features while driving. This report details the work conducted for Voice 2, the second of two projects in this program.

2.1 PREVIOUS WORK

The aim of the first VCS project (Voice 1) was to determine the range of user experiences and common interaction problems that occur with current generation VCS and included on-road contextual interviews with voice control users (Jenness et al., 2015). The purpose of the Voice 1 project was to conduct empirical research about the use of VCS by drivers and potential measures that could be used for evaluating possible distraction from using these systems while driving. Laboratory studies were also conducted to explore the sensitivity of eye glance protocols in the NHTSA Visual-Manual Driver Distraction Guidelines (NHTSA, 2013), and the International Standards Organization Tactile Detection Response Task, to a decrease in VCS recognition accuracy and an increase in system delay (ISO, 2015; Jenness et al., 2015).

An on-road, contextual interview study was conducted in Rockville, Maryland, and Seattle, Washington, to identify drivers' existing patterns of use and interaction errors encountered with VCS while driving. Differences were observed between those who used original equipment VCS and those who used portable smart devices that were paired to the vehicle. The research team noted 22 themes that characterized participants' interactions with VCS (Jenness et al., 2015). Most notably, drivers often had trouble using their VCS but did not necessarily blame the system for the errors or the lengthy system interactions that they experience. Instead, the participants sometimes blamed themselves or they merely tolerated the suboptimal interactions with the system. Interactions frequently included errors including speech recognition errors. These results suggest that an evaluation protocol based solely on error free trials would not be representative of many VCS interactions commonly experienced by users while driving.

Two other studies were conducted under the Voice 1 project. Both were controlled laboratory experiments in which participants interacted with a "Wizard of Oz" VCS while performing a surrogate driving task. In the first study, the surrogate driving task was operating a driving simulator and in the second study, the surrogate driving task was a computer-based collision detection task. As previously mentioned, cognitive load was measured by performance on the ISO TDRT. Eye glance measures, based on NHTSA's Visual-Manual Driver Distraction Guidelines for In-Vehicle Electronic Devices (NHTSA, 2013) were also used. Results indicated that both TDRT performance and eye glance measures may be appropriate for evaluation of VCS and that the CDT protocol yielded similar results to the driving simulator protocol. Another driving simulation study, "Bridge Study" was conducted after the completion of the Voice 1 project and was documented in an unpublished memo to NHTSA (Jenness, Boyle, Guo, Lee, &

Chang, 2015). The "Bridge Study" showed that these measures (eye glances and TDRT) are sensitive to different display modes (visual, auditory, both) and to two different difficulty levels of the VCS task.

Both the contextual interviews and laboratory studies indicated that drivers were more likely to look away from the forward roadway when engaged in voice-based tasks. The initial studies suggest that some combination of TDRT measures and glance measures may be useful for the evaluation of VCS. The motivation of the current program (Voice 2) was to examine how these measures can be used together to relate the demands of VCS tasks to on-road safety consequences and then to determine threshold values for the measures that would be indicative of acceptable performance. This may entail using one or more baseline tasks to gauge relative performance for VCS tasks as compared to other secondary tasks that are deemed acceptable to perform while driving. This approach was used in developing the NHTSA Phase 1 Driver Distraction Guidelines (NHTSA, 2013) where a visual-manual radio-tuning task was used as a baseline acceptable task.

Voice 1 used TDRT and recorded eye glances as measures of distraction. Voice 2 sought to investigate other metrics to determine their sensitivity to distinguish between VCS tasks. The same standard that established the TDRT (ISO, 2015) offered a visual alternative, the remote visual detection response task (RDRT). While the RDRT is recommended by ISO for auditory-only tasks if measuring cognitive load, it may be possible for RDRT to be used for visual tasks to measure visual demand. Miss rates of the RDRT have been shown to be sensitive to visual demands in addition to cognitive demands (ISO, 2015). However, in the past, researchers have anecdotally noted that it is possible to detect the onset of the red light from the RDRT, even when looking away at a secondary screen. For the current study, researchers modified the red RDRT light (LED light source) so that it was presented within the context of a small white, steadily illuminated background. This made the RDRT harder to see using one's peripheral vision. Using the modified RDRT, unless a participant is looking in the direction of the RDRT light (towards the forward roadway) it is difficult to detect when it is turned on. Researchers hypothesized that this modified RDRT would be a sensitive tool for detecting visual attention away from the forward roadway. The modified version (see Section 3.2.4) of the RDRT was used in Study 1.

2.2 OBJECTIVES

The current project, Voice 2, builds upon previous research regarding the evaluation of VCS for potential driver distraction. Specifically, this project seeks to evaluate how TDRT measures of cognitive load may be used, in conjunction with glance measures, for evaluating potential driver distraction from performing VCS tasks. An ultimate goal is relating measures of cognitive and visual distraction to crash risk for the evaluation of VCS. The project was structured as three related studies.

- Study 1 - Explored a variant of the ISO detection response task protocols that includes a new, modified RDRT protocol and the TDRT protocol. The proposed dual DRT protocol may have the potential for detecting both cognitive and visual distraction caused by performing VCS tasks. If these two measures are effective, data from this dual protocol

could reduce the need to perform a separate analysis of eye glance data when evaluating VCS.

- Study 2 - Compared several potential VCS evaluation measures of cognitive load (TDRT) and visual attention (glance measures) collected in the laboratory with a driving simulator to the same measures collected with the same set of participants in a real vehicle on-road.

- Study 3 - Connected potential VCS evaluation measures of visual attention (glance measures) to their safety relevance and potential crash risk by combining new empirical data from Study 1 and Study 2 with previous crash data taken from the second SHRP 2 naturalistic driving study. Statistical modeling, using counterfactual simulation, explored a multitude of "what if" scenarios and from these simulations, a measure of relative crash risk was developed.

2.3 RESEARCH QUESTIONS

Each study addressed specific research questions regarding their respective objectives, as follows:

Study 1

- Are the variant of the ISO RDRT and the ISO TDRT protocols sensitive to the demands imposed by VCS tasks?
- Does the dual DRT capture both the cognitive load and visual distraction measured by TDRT and eye glance measures from NHTSA's Phase 1 Driver Distraction Guidelines (NHTSA, 2013)?
- What are the relative amounts of cognitive load and visual load (off-road glances) that are measured when participants perform candidate baseline tasks such as N-back and a new auditory-vocal radio-tuning task as compared to previously developed radio-tuning, navigation, and calendar tasks? Would the new auditory-vocal radio-tuning task be a good baseline task to be used for evaluating VCS?

Study 2

- Do similar driving simulator and on-road protocols for measuring the distraction potential of VCS tasks yield similar results?
- How does performing VCS tasks in the driving simulator and on-road influence braking response time for a surrogate hazard?
- How does performing VCS tasks in the driving simulator and on-road influence drivers' eye glance behavior? Are the eye glances similar in the two contexts?

Study 3

- Primary overarching question: How do glance measures and TDRT measures collected in a laboratory (with driving simulation) and on-road relate to crash probability and crash severity?

- Can a set of data containing crashes and near crashes (SHRP 2 data) be used to support counterfactual simulation that estimates relative crash risk for various VCS tasks evaluated in a driving simulator or on-road?

An overview of the research plan is shown in Figure 1. The project consisted of a combination of laboratory and on-road studies. In addition, statistical modeling and simulation was also used to determine crash probability and severity for the VCS tasks used in the project. The team at the University of Washington led Study 1, which is described on the upper left side of Figure 1. Westat led Study 2, which is described in the center and right side of Figure 1. The team at the University of Wisconsin led Study 3, the statistical modeling effort, which is described at the bottom of Figure 1. The same set of VCS tasks were used in Study 1 and throughout Study 2.

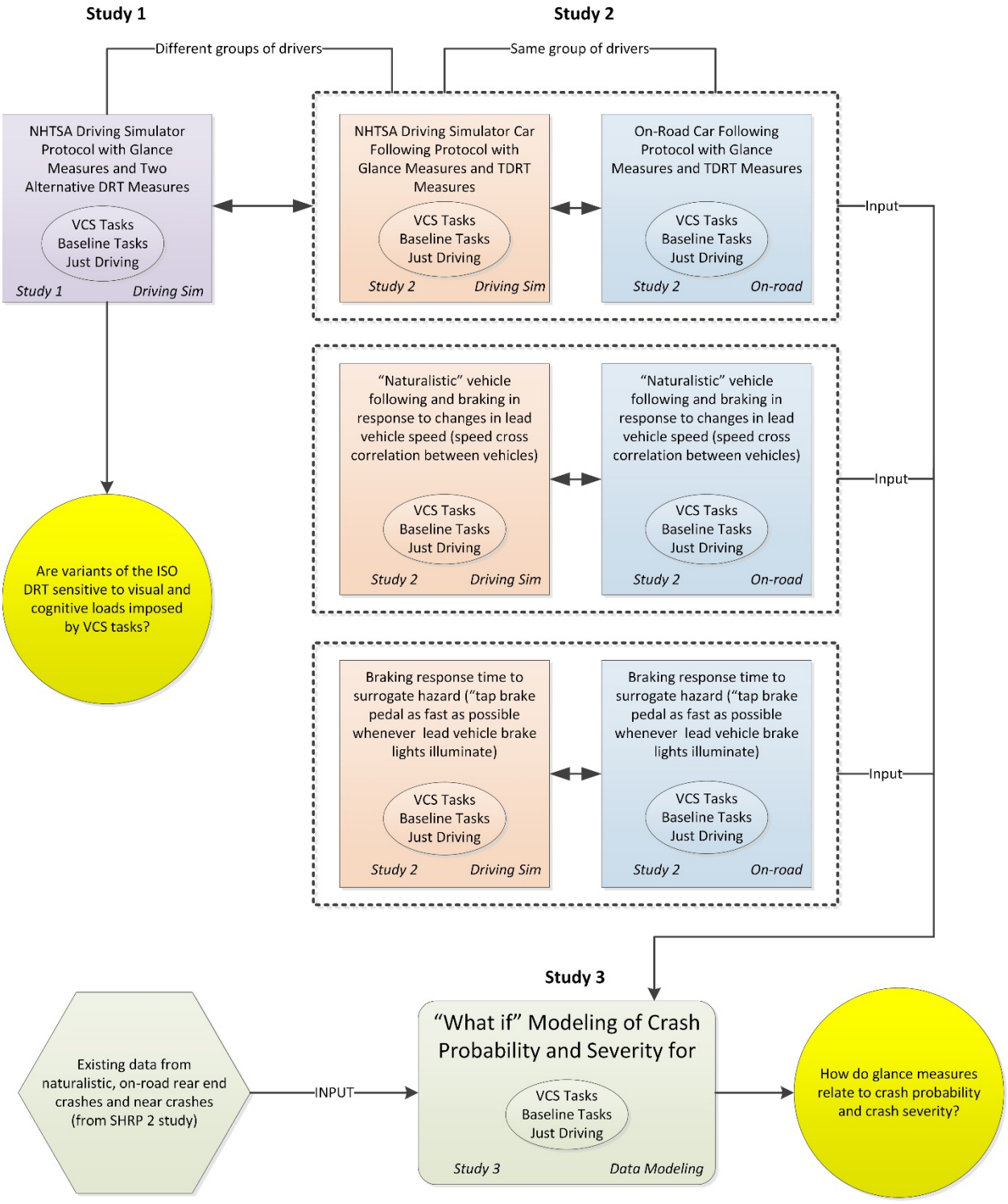


Figure 1. Overview of Research Plan

3 STUDY 1 – TWO DETECTION RESPONSE TASK PROTOCOLS FOR EVALUATING VOICE CONTROL SYSTEMS

3.1 PURPOSE

The main goal of Study 1 was to explore whether a variant of the ISO DRT may be more sensitive than TDRT alone to both visual and cognitive demands of VCS tasks. A second goal was to develop a verbal/auditory radio-tuning task for evaluating other VCS tasks.

3.2 METHOD

A driving simulator study was conducted at the University of Washington to get a better understanding of how ISO DRT measures (reaction time, miss rate) relate to the cognitive demands of interacting with VCS while driving (purple box on the left side of Figure 1).

3.2.1 Participants

There were nine participants recruited for this study. This sample size is sufficient to detect large to moderate effect sizes and support the development of radio-tuning tasks. All the participants were 25 to 54 years old, had valid U.S. driver licenses, drove at least 3,000 miles per year, and were native English speakers. Participants were each compensated $50 for their time for participating in the study. The study was reviewed and approved by the University of Washington's IRB for the protection of human research participants.

3.2.2 Study Design

The study was a 2 x 2 x 2 x 2 (VCS display mode: Audio Only, Audio + Visual Hybrid x Task Type: Navigation, Radio x Task Difficulty: Easy, Hard x Gender: Male, Female) mixed factorial design. VCS display mode, task type, and task difficulty were within subject variables, where every participant experienced all of these conditions, but in a random order, for a total of four drives.

There were eight unique combinations for the order of the study tasks, thus only two participants experienced the same task sequence. The two TDRT and two RDRT drives were blocked together for each participant, such that four participants experienced the two TDRT drives first and five participants experienced the two RDRT drives first. However, within these drives, the order of display mode was randomized, thus making possible eight combinations. Within each drive, the order for the VCS task type (i.e., Navigation and Radio) and task difficulty (i.e., Easy and Hard) were randomly generated by the computer. For each drive, participants completed 3 Radio Easy, 3 Radio Hard, 3 Navigation Easy, and 3 Navigation Hard tasks. All tasks within the same DRT mode were different, but tasks between DRT modes were the same, such that each participant completed each task once with TDRT and once with RDRT. A detailed diagram of this study design is provided in Figure 2.

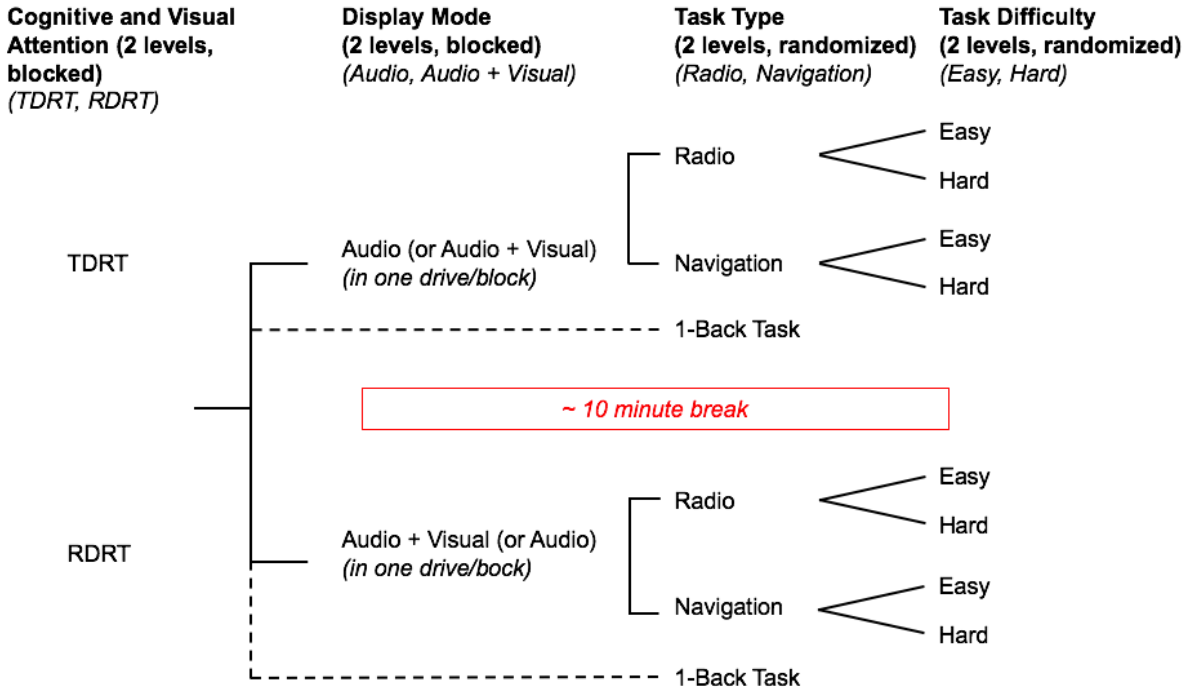


Figure 2. VCS Task Design: 2 (DRT) x 2 (display mode) x 2 (task type) x 2 (task difficulty) x 3 replication blocked factorial design (or 48 VCS trials + 2 1-Back trials)

3.2.3 Driving Simulator

The study was conducted using a fixed-base National Advanced Driving Simulator miniSim. A photograph of this setup is provided in Figure 3. A 7-inch monitor was mounted to the right of the steering wheel to provide a visual display for the VCS tasks with a visual component; for the VCS tasks that were Audio Only, this 7-inch monitor displayed a solid black screen.

Figure 3. Driving Simulator Setup in Seattle location, with VCS display circled

Participants were first trained on the study procedures by completing a ten-minute practice drive. During this practice, participants first started out by just driving the simulator, then after

approximately two minutes the VCS tasks were added. After completing four VCS tasks, a one-minute practice of DRT and driving began, followed by the addition of two more VCS practice tasks while still driving and engaging with the DRT. Finally, while still driving and doing the DRT, the participants practiced the 1-Back task. During this practice drive, participants only practiced one DRT type (i.e., TDRT or RDRT). Participants were given a separate one-minute practice with the second DRT between the second and third main study drive.

After the participant was comfortable with the study procedures, each participant completed four drives that were approximately 15 minutes each, with approximately a 10-minute break between each drive. Each drive interval included 12 VCS tasks: 3 Radio Easy, 3 Radio Hard, 3 Navigation Easy, and 3 Navigation Hard. These tasks were randomized within blocks, where each drive block contained one VCS display mode (i.e., Audio Only or Audio + visual) and one DRT type (i.e., TDRT or RDRT). A computer program randomly ordered the 12 VCS tasks for each drive. A 1-Back task was performed as the last task of drives 1 and 3, such that each participant completed the 1-Back task while doing each type of DRT.

Drivers were instructed to maintain a safe following distance from the lead vehicle and stay in the right lane during the entire drive. The driving scenario was an undivided four-lane (2 lanes in each direction) straight, flat roadway with a posted speed limit of 50 mph. The LV speed was generated by a function consisting of the sum of three different sine waves, in which the speed variations were smooth and unpredictable, having an average speed of approximately 50 mph.

3.2.4 Voice Control System Tasks

During all four drives, participants interacted with a "Wizard of Oz" style VCS (i.e., a method of simulation where participants are told that they are interacting with a VCS but actually interact with a human confederate), designed to simulate an in-vehicle VCS. This VCS was previously developed for Voice Part 1 project (Jenness et al., 2015) but was adapted for this project.

There were two types of VCS tasks, Radio tasks and Navigation tasks, and within each of these task types there were Easy and Hard tasks. Furthermore, there were two types of VCS display modes: Audio Only and Hybrid (Audio + Visual).

For the Radio tasks, participants were instructed to tune to the radio station that closest matched the genre stated by the system. For example, the system would say "Please tune to an International news station." Participants were then given a list of 12 options (11 radio stations and 1 "none of these") to choose from. For the Audio Only mode, the system simply said the list of 12 options; for the Hybrid mode, the system read the list of 12 options and displayed these options on the monitor, for which there were two pages of radio stations with 6 options on each page. Participants could say "next" and "previous" to switch between pages. More specifically, for the Radio Easy tasks, the name of the correct radio station always contained at least one of the key words listed in the genre described by the system, and was within the first 6 radio station options (i.e., on the first page for the Hybrid display mode). Figure 4 below shows an example of a Radio Easy task for the Hybrid display mode.

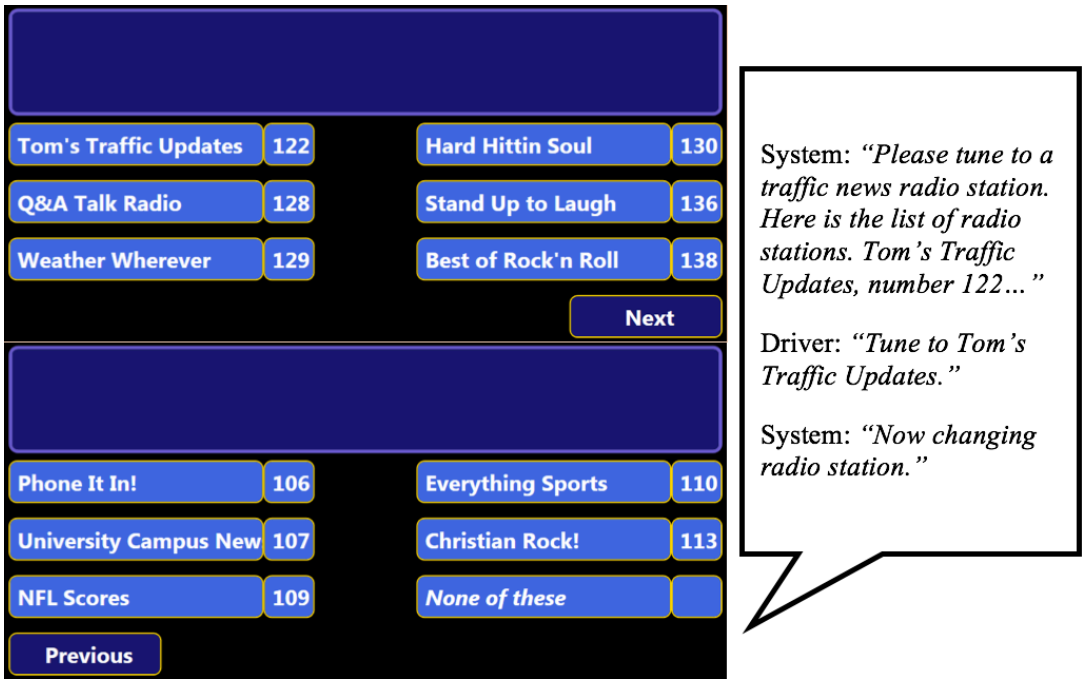


Figure 4. Radio Easy Hybrid Example

For the Radio Hard tasks, the exact wording used in the genre description was not necessarily contained within the correct radio station name (i.e., a synonym was used instead), and the correct answer was always within the last half of the radio station options (i.e., on the second page for the Hybrid display mode). Figure 5 shows an example of a Radio Hard Hybrid task.

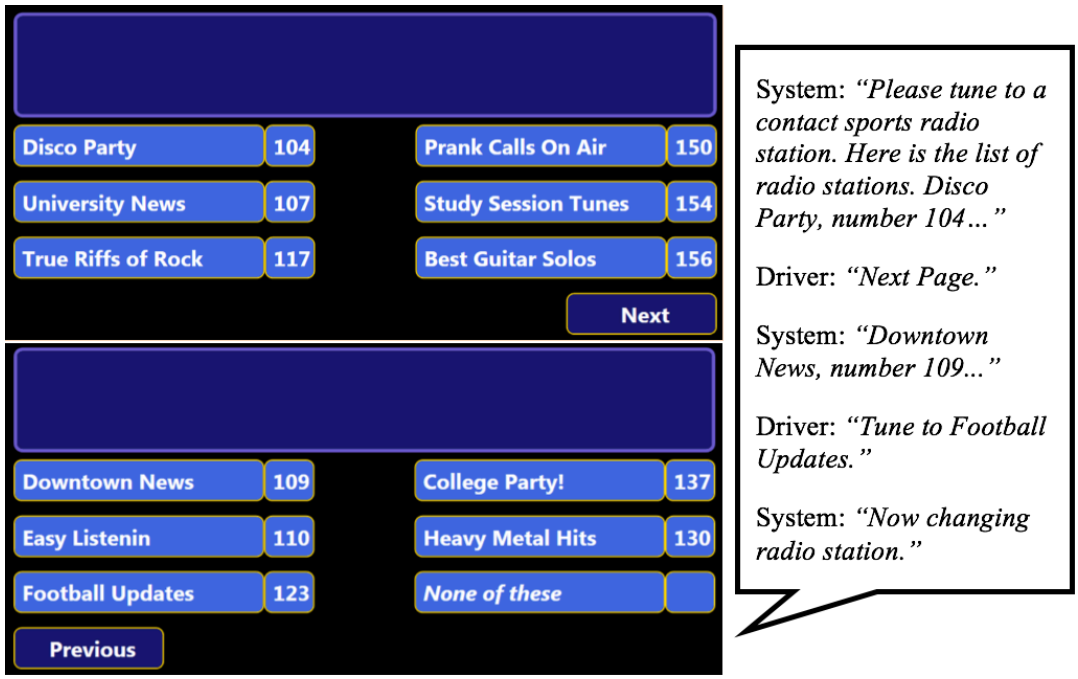


Figure 5. Radio Hard Hybrid Example

For the Navigation tasks, participants were instructed to always navigate to the optimal American restaurant based on a set of criteria; the criteria were to find the American restaurant with the highest star rating, closest distance, and cheapest price. For this task, the system would say, "What would you like to do" and the participants were trained to say, "Find American restaurant." The system would then return a list of five restaurants, including the restaurant name, cuisine type, number of reviews, star rating, price, and distance. For the Audio Only tasks, this list would be read and for the Hybrid tasks the information was additionally displayed as a list on the monitor. For the Navigation Easy tasks, the list of five restaurants would contain two American options and three non-American cuisines; of the two American restaurants, only one would be better on all three of the criteria, see Figure 6 for an example of a Hybrid display Navigation Easy task.

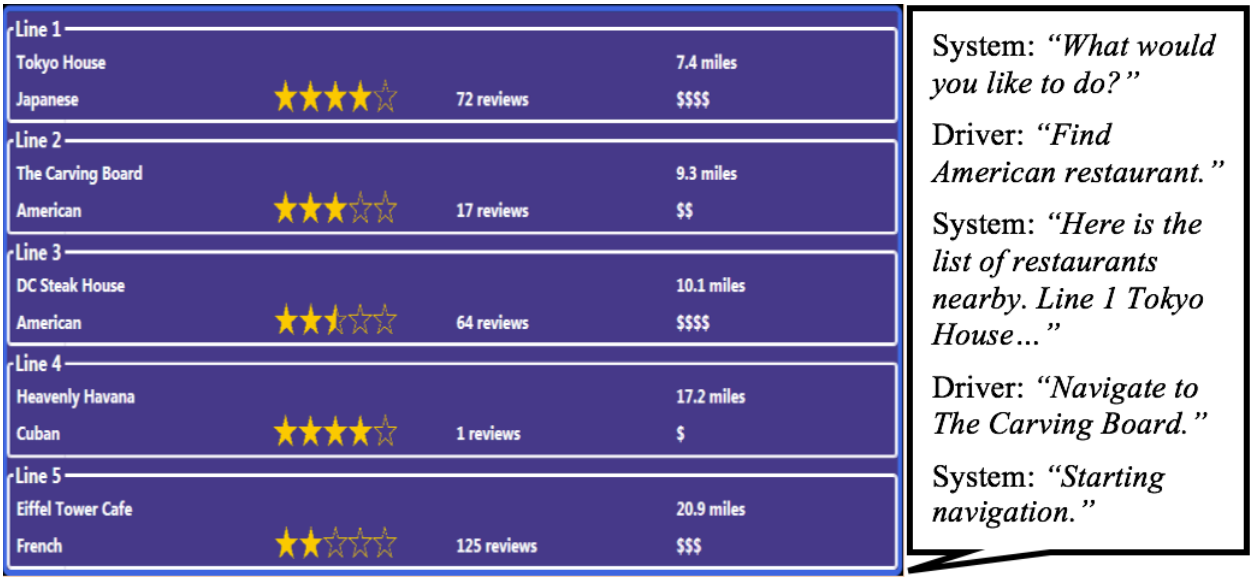


Figure 6. Navigation Easy Hybrid Example

In contrast, the Navigation Hard tasks contained four American restaurant options and only one non-American, for which only one American restaurant was better on two of the three criteria. An example of the Navigation Hard Hybrid task is provided in Figure 7.

Figure 7. Navigation Hard Hybrid Example

Additionally, an N-back (1-Back) task was administered to each participant during two of the drives, once for each DRT type. For this task, an automated voice said a series of 14 numbers, and participants were instructed to repeat back the number previous to the one they just heard. For both 1-Back tasks, the VCS mode was always Audio Only, such that there was no display on the 7-inch monitor for this task.

3.2.5 Detection Response Task

This study used TDRT and a new variant of the RDRT to examine measures of cognitive and visual load. The modified RDRT stimulus consisted of a constantly illuminated white disk (illuminated by a flashlight beam) and a smaller, red target light that appeared in the middle of the disk. This visual stimulus was designed to be more difficult to detect than the standard red LED alone. The white illuminated background helped to mask the light from the red LED and made the onset of the red LED quite difficult to detect using peripheral vision when the driver looked toward the infotainment system display, but still easy to detect when the driver looked toward the forward roadway. It was thought that this design for the RDRT would provide a purer measure of attention toward the forward roadway than the conventional ISO RDRT. The improved RDRT apparatus is shown in Figure 8.

Figure 8. Modified RDRT apparatus positioned in the driving simulator (left) and the RDRT apparatus turned on, displaying the light signal from the red LED shining through the white background (right).

The white background fixture was constructed as a sandwich of two outer layers of white plastic and two inner layers of white paper. The front surface has a matte finish to reflect light diffusely. The back surface has a small piece of 1.0 neutral density (ND) gelatin optical filter taped to it. This filter reduces the light intensity from the LED to 10 percent of the normal light output. With the 1.0 ND filter in place, the RDRT red channel LED was operated at 100 percentintensity.

The TDRT stimulus (vibrating motor) was taped to the base of the participant's neck where it intersects with the shoulder as described in ISO standard 17488, and the RDRT stimulus (red LED) was placed in the direction of the forward roadway. The RDRT LED was positioned such that it was easily visible when looking at the forward roadway, but not visible in the peripheral when looking at the VCS monitor. On separate drives, the TDRT or RDRT vibrated (TDRT) or illuminated (RDRT) according to the temporal parameters defined in the ISO standard (ISO 17488). The basic DRT hardware and controller unit were purchased from Red Scientific (Salt Lake City, UT).

The TDRT was used in two drives, one for the Audio Only display and one for the Audio + visual display; the RDRT was also used in two drives, also for one Audio Only display and one Audio + visual display. The order of drives with TDRT and RDRT was randomized across participants.

Baseline performance for TDRT and RDRT was measured for each participant by recording their response (see Section 3.2.7 for details) while just driving and not engaging in any of the secondary tasks. Cognitive load and visual attention was measured using response time and miss rate recorded from the TDRT and improved RDRT protocols.

3.2.6 Eye Glance Behavior

Visual attention was additionally measured in the study by evaluating eye glance behavior (i.e., eyes off road). Eye glances were recorded using a video camera at 720p and 30 frames/sec. The video data was manually reduced using Tech Smith Morae Manager.

3.2.7 Dependent Measures

3.2.7.1 DRT Measures of Workload

DRT measures for reaction time and miss rate were used to evaluate workload. As per ISO standards, a valid response was considered to a reaction within the threshold of 100 to 2500 ms from stimulus onset.

A successful hit was defined as a valid response to the DRT (i.e., TDRT or RDRT) stimulus. The miss rate was calculated as the number of misses divided by the total number of stimulus events (see Section 3.2.4), for the given time span.

3.2.7.2 Eye Glance Criterion

The three criteria for eye glance behavior outlined in the NHTSA Visual-Manual Driver Distraction Guidelines (NHTSA, 2013) were applied for analysis on the Hybrid VCS tasks. For each Hybrid VCS task, eyes-off-road glance durations were evaluated in terms of long (i.e., ≥ 2.0 seconds) eyes-off-road glances (Criterion 1), mean glance duration (Criterion 2), and total eyes-off-road time (Criterion 3). The NHTSA Visual-Manual Driver Distraction Guidelines call for conformance of at least 21 out of 24 test participants; since only 9 participants were considered for this study, these metrics and thresholds were adapted to provide within subject comparisons across various VCS tasks.

Criterion 1: Percentage of Long Eyes-Off-Road Glances: The threshold of 15 percent should be used for the upper limit for the total number of eye glances away from the forward roadway greater than or equal to 2.0 seconds, within a trial.

$$\%\ Long\ EOR_{ij} = \frac{No.of\ Long\ EOR_{ij}}{Total\ No.EOR_{ij}} \times 100$$

where, Long EOR = EOR glance ≥ 2.0 seconds
i = participant
j = trial

Each participant completed each task three times within a given experimental condition, for example a participant completed Radio Easy tasks under each Audio Only mode with RDRT, Audio Only mode with TDRT, Hybrid mode with RDRT, and Hybrid mode with TDRT. Thus, there were 3 trials for each task; % Long EOR for a task was computed as the mean and maximum value across the three trials within a participant.

Criterion 2: Mean Glance Duration: The threshold of 2.0 seconds was applied, for which the mean duration of all eye glances away from the forward roadway should be less than or equal to 2, within a trial.

$$MGD_{ij} = \frac{\sum_1^n EOR\ Duration_{ij}}{n_{ij}}$$

where, i = participant
j = trial

Similar to the rational from Criterion 1, the mean and maximum values of MGD for a given task within a participant was computed across the three trials and compared.

Criterion 3: Total Eyes-Off-Road Time: Within a trial, the sum of the durations of eyes-off-road glances should not exceed 12.0 seconds.

$$TEORT_{ij} = \sum_1^n EOR\ Duration_{ij}$$

where, i = participant
j = trial

Mean and maximum values for TEORT across a given task for the three trials were computed for each participant.

As discussed within each criterion, mean and maximum values for each measurement were computed for each participant. The inclusion of the maximum value of a trial within a task for a participant provides sensitivity to outliers, and thus does not overlook a trial in which a participant was exhibiting unsafe behaviors.

3.2.8 Statistical Methods

Analysis of variance was used to evaluate the outcome variables standard deviation of lateral position (SDLP), standard deviation of vehicle speed, and DRT response time. For each of these three continuous dependent variables, two ANOVAs were performed: *(1)* 4 (VCS Mode: drive, 1-Back, Hybrid, Audio) x 2 (DRT: tactile, visual) and *(2)* 6 (VCS Task Type: drive, 1-Back, Radio Easy, Radio Hard, Navigation Easy, Navigation Hard) x 2 (DRT: tactile, visual). In order to adhere to the normality assumption of the ANOVA, SDLP was log transformed, SD speed was square root transformed, and DRT response time was log transformed. These transformations allowed the dependent variables to be normally distributed.

A negative binomial model was used for the outcome variable DRT miss count, as ANOVA could not be applied due to the distribution of the dependent variable (i.e., right skewed and bound at 0). Similar to the three continuous dependent variables, two negative binomial models

were fit: *(1)* full interaction of VCS Mode x DRT and *(2)* full interaction of VCS Task Type x DRT.

ANOVA was also used to evaluate the outcome variable of mean eye glance duration away from roadway. This ANOVA was only performed on the Hybrid VCS tasks, as the Audio tasks did not have a visual component to measure eyes-off-road time. The ANOVA was a 4 (VCS Task Type: Radio Easy, Radio Hard, Navigation Easy, Navigation Hard) x 2 (DRT: tactile, visual). This ANOVA was performed twice, once on the mean of the three trials per task type, and once on the maximum value of the three trials per task type.

In all the above models (ANOVAs and negative binomial), a random intercept on the participant was fit to account for subject specific variance. Each model accounted for the fixed effects of the independent variables and the random effect of the participant. Thus, the model accounted for the correlation of observations within participants. For each model, a Likelihood Ratio Test was performed to test whether the variance was different from 0 (i.e., if the random intercept was warranted). In all cases, the Likelihood Ratio Test suggested that the participant level random intercept was significant.

3.3 RESULTS

This study sample consisted of three males and six females; their mean age was 31.3 (SD = 5.5, range = 26 to 43). Three participants had a 2-year college degree, three had a 4-year college degree, and three had a graduate degree. None of the participants reported having hearing impairments.

Participants reporting driving a mean of 96.3 miles (SD = 109.5 miles, range = 0 to 320 miles) in the previous week. Six participants had not been involved in a collision as a driver with in the past 3 years, three had been in one crash as the driver within the past 3 years. Seven participants had received no moving violation tickets within the past 3 years and two participants had received one moving violation ticket within the past 3 years.

There were no statistically significant gender differences observed in driving performance or cognitive workload, thus each model was run without gender included.

3.3.1 Task Performance

For each task type, participants completed most Hybrid trials faster than the equivalent Audio Only display mode. The distributions of these task completion times by task type and display mode are shown in Figure 9, which includes duration for each trial for every participant.

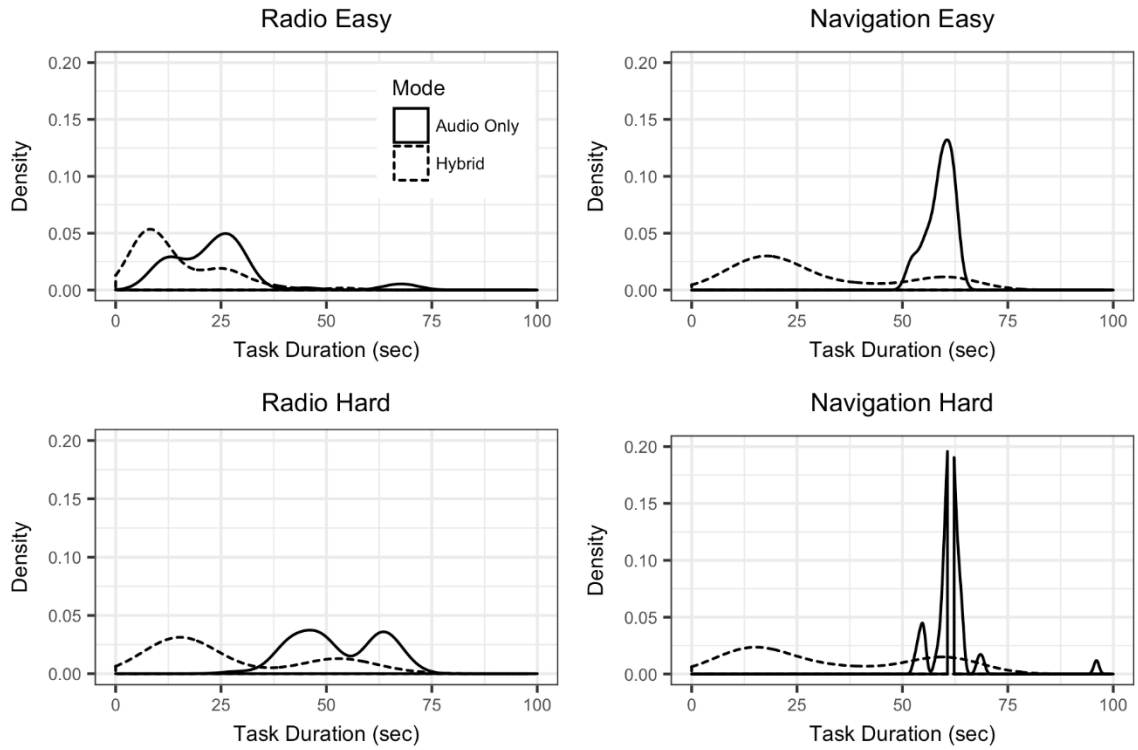


Figure 9. Task duration distributions by task type and display mode

The order of tasks by average task duration within VCS mode was the same across Audio Only and Hybrid display modes; Radio Easy was completed most quickly on average, followed by Radio Hard, followed by Navigation Easy, and lastly Navigation Hard. Table 1 provides mean durations for task completion accuracy percentage by task type (i.e., for the number of tasks completed correctly by each participant), which was then averaged across all participants. Radio Easy tasks were completed with similar accuracy between VCS mode, however Hybrid tasks were completed more accurately on average than Audio Only tasks for Radio Hard, Navigation Easy, and Navigation Hard tasks.

Table 1. Task Mean Duration and Accuracy by Task and Mode

Task Type	**VCS Mode**	**Duration (sec)**	**Accuracy (%)**
Radio Easy	Audio	24.4	94.5
	Hybrid	14.8	94.4
Radio Hard	Audio	52.6	90.8
	Hybrid	27.1	96.3
Navigation Easy	Audio	58.9	92.6
	Hybrid	29.3	96.3
Navigation Hard	Audio	61.6	83.4
	Hybrid	32.4	92.6

3.3.2 Driving Performance

Changes in speed and lateral position were used to evaluate the effects of VCS tasks on driving performance. Figure 10 illustrates this information, where speed (left) and lateral deviation (right) are plotted across baseline driving (Drive) Audio Only, Hybrid, and 1-Back tasks. Recall this data was sampled at 60Hz, therefore an average of the 60 observations within each one second interval was computed to provide a single data point per one second. This one-second average value was plotted to minimize repetitive detail. Among the three VCS modes, the largest spread in speed and lateral deviation was in Hybrid tasks (i.e., wider distribution and more outliers as shown by the dots), followed by the Audio Only tasks, and lastly with the 1-Back task.

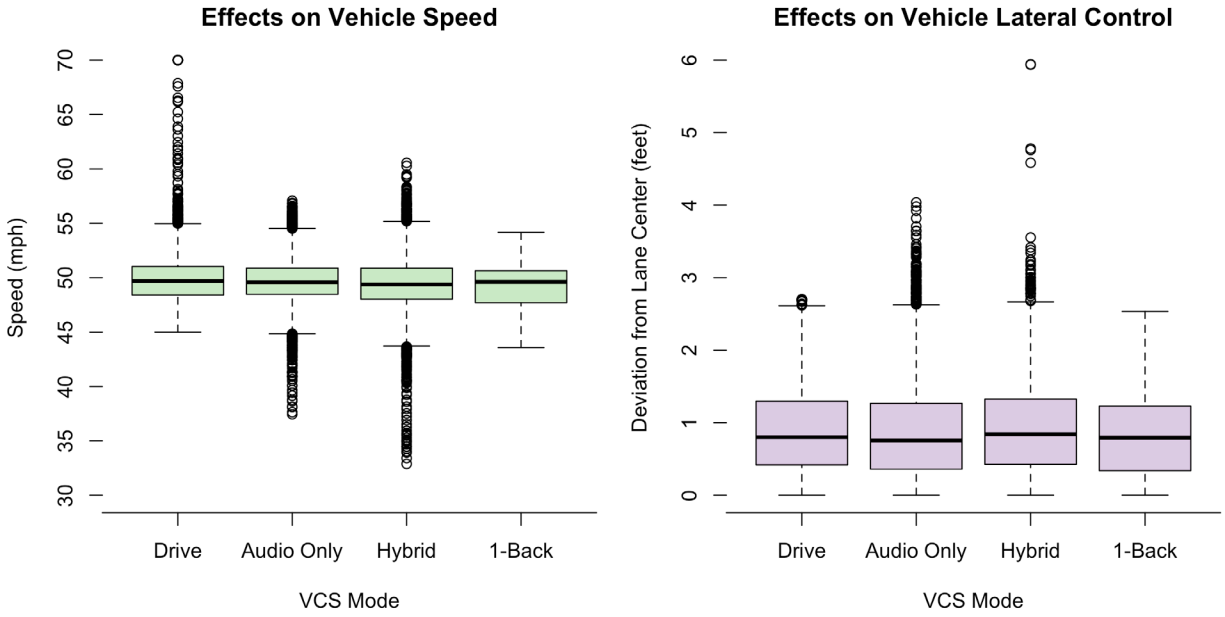


Figure 10. Driving Performance by VCS Mode

3.3.2.1 Standard Deviation of Lateral Position

A further analysis on VCS mode, task, and DRT type was performed on the SDLP. For this analysis, SDLP was measured in inches and log transformed for normality. A random intercept was fit for each participant to account for within subject differences. Two ANOVAs were performed, one evaluating VCS Mode (Table 2) and a second for VCS Task (Table 3). These were evaluated separately due to overlap between mode and task as drive and 1-Back . There were no differences in mean SDLP values by DRT type or the interaction with DRT type. However, there were significant differences in mean SDLP values for the main effects of VCS mode ($p < 0.01$) and VCS task type ($p < 0.01$).

Table 2. ANOVA on log(SDLP) for VCS Mode

Variable	DF	SS	F-value	p-value
DRT (Visual, Tactile)	1	0.040	0.292	0.590 (ns)
Mode (Drive, 1-Back, Audio, Hybrid)	3	6.189	15.199	< 0.001
DRT x Mode	3	0.147	0.048	0.782 (ns)

Table 3. ANOVA on log(SDLP) for VCS Task Type

Variable	DF	SS	F-value	p-value
DRT (Visual, Tactile)	1	0.040	0.306	0.581 (ns)
Task (Drive, 1-Back, Radio Easy, Radio Hard, Nav Easy, Nav Hard)	5	7.673	11.823	< 0.001
DRT x Task	5	0.200	0.308	0.908 (ns)

We conducted a Tukey Honest Significance Difference test on VCS mode. This analysis showed that each contrast involving the 1-Back task had significantly different mean values of SDLP ($p < 0.5$). The mean SDLP for the 1-Back task was small for each case (i.e., 1-Back – Audio, 1-Back – Hybrid, and 1-Back – Drive).

The Tukey HSD test on VCS Task Type also suggested that every contrast involving the 1-Back had significantly different mean values of SDLP ($p < 0.5$), and that in every case the 1-Back task had smaller mean SDLP values. Other contrasts that had significantly different mean SDLP values were:

- Radio Easy (Mean = 6.49 in, SD = 2.07 in) and Drive (Mean = 8.47 in, SD = 2.58 in)
- Radio Easy (Mean = 6.49 in, SD = 2.07 in) and Navigation Hard (Mean = 8.85 in, SD = 4.09 in)
- Radio Hard (Mean = 8.05 in, SD = 2.48 in) and Radio Easy (Mean = 6.49 in, SD = 2.07 in)

It is important to note that the driving task included portions in between each VCS task, thus it is possible that lane position correcting due to VCS distractions occurred after the task was completed during the driving only segments.

3.3.2.2 Standard Deviation of Vehicle Speed

A similar analysis as above was performed on VCS mode, task, and DRT type on standard deviation of vehicle speed. SD Speed was measured in miles per hour and square root

transformed for normality and a random intercept on participant was fit to account for within subject differences.

There was no significant difference between DRT types or the interaction of DRT with Mode or with Task. The main effects of each Mode ($p < 0.001$) and Task ($p < 0.001$) were significant. These results are further detailed in Table 4 and Table 5.

Table 4. ANOVA on (SD Speed)^0.5 for VCS Mode

Variable	DF	SS	F-value	p-value
DRT (Visual, Tactile)	1	0.003	0.030	0.862 (ns)
Mode (Drive, 1-Back, Audio, Hybrid)	3	2.864	9.035	< 0.001
DRT x Mode	3	0.037	0.115	0.951 (ns)

Table 5. ANOVA on (SD Speed)^0.5 for VCS Task Type

Variable	DF	SS	F-value	p-value
DRT (Visual, Tactile)	1	0.003	0.030	0.863 (ns)
Task (Drive, 1-Back, Radio Easy, Radio Hard, Nav Easy, Nav Hard)	5	2.690	4.948	< 0.001
DRT x Task	5	0.117	0.216	0.956 (ns)

The Tukey post hoc test on the factor VCS mode indicated that the following pairwise contrasts had significantly different ($p < .05$) mean values of SD speed:

- Drive (Mean = 2.26, SD = 1.53) and Audio (Mean = 1.59, SD = 0.72)
- 1-Back (Mean = 0.98, SD = 0.53) and Audio (Mean = 1.59, SD = 0.72)
- 1-Back (Mean = 0.98, SD = 0.53) and Drive (Mean = 2.26, SD = 1.53)
- 1-Back (Mean = 0.98, SD = 0.53) and Hybrid (Mean = 1.87, SD = 0.95)

The results of the Tukey HSD test on VCS Task Type showed that each pairwise contrast involving the 1-Back task had significantly different mean SD speed values, smaller mean SD speeds. Additionally, mean SD speeds for Radio Easy (Mean = 1.63, SD = 1.03) were significantly lower than Drive (Mean = 2.26, SD = 1.53).

3.3.3 Cognitive and Visual Attention

Response time and miss rate for TDRT and RDRT were used to evaluate cognitive and visual attention while engaging in VCS distracting tasks.

3.3.3.1 Response Time

A comparison of response times across VCS tasks by DRT type for all trials is plotted in Figure 11. For every task except for Radio Hybrid, response time was larger for the RDRT measures as compared to the TDRT measures. However, the relative trends across VCS tasks between the two DRT types are similar. For the Easy Navigation Audio Only and Hard Radio Audio Only tasks, the mean reaction times were shorter than for baseline driving, for the respective DRT type. All other tasks showed an increase in mean response time above baseline driving relative to the respective DRT type.

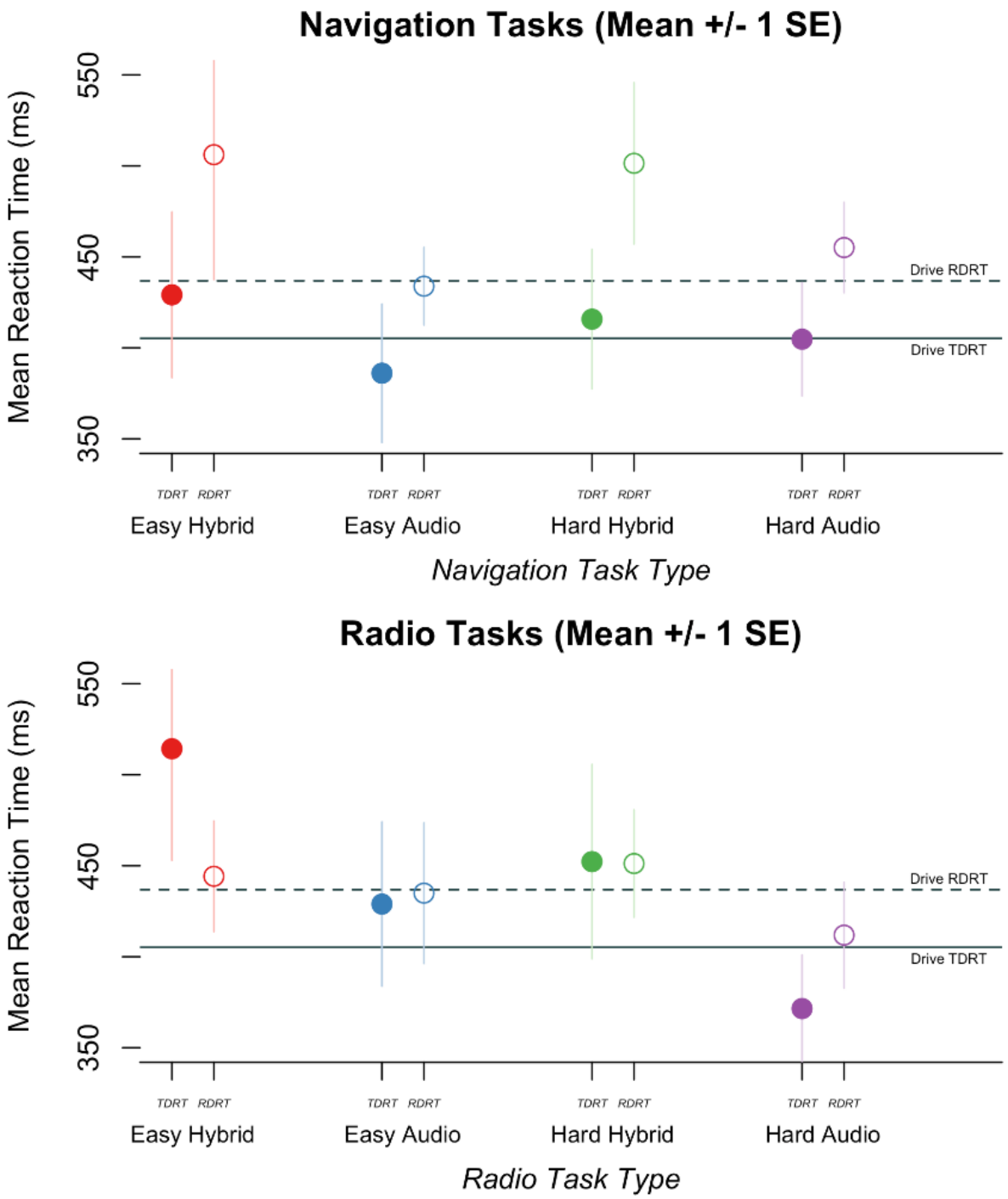


Figure 11. Response Time by VCS Task and DRT Type Compared to Baseline Driving Mean Response Times for RDRT and TDRT

To normalize the distribution of DRT Response Time data, we performed a log transformation to conduct analyses with two ANOVAs.. As with the ANOVA on SDLP and SD speed, DRT type did not have a significant effect on mean response time. However, VCS display mode ($p < 0.001$) and the interaction of DRT with display mode ($p = 0.009$) were significant, as were VCS task type ($p < 0.001$) and the interaction of DRT with task type ($p = 0.009$).

Table 6. ANOVA on log(Response Time) for VCS Mode

Variable	DF	SS	F-value	p-value
DRT (Visual, Tactile)	1	0.104	2.347	0.127 (ns)
Mode (Drive, 1-Back, Audio, Hybrid)	3	2.862	21.515	< 0.001
DRT x Mode	3	0.527	3.957	0.009

The Tukey HSD test was performed on the factor VCS Mode (Table 6). All pairwise contrasts including the 1-Back task, as well as the contrast between Hybrid and Audio had significantly different DRT reaction times. These differences ($p < .05$) were as follows:

- Hybrid (Mean = 464 ms, SD = 141.8 ms) and Audio (Mean = 416 ms, SD = 97.6 ms)
- 1-Back (Mean = 702 ms, SD = 363.9 ms) and Audio (Mean = 416 ms, SD = 97.6 ms)
- 1-Back (Mean = 702 ms, SD = 363.9 ms) and Drive (Mean = 412 ms, SD = 81.5 ms)
- 1-Back (Mean = 702 ms, SD = 363.9 ms) and Hybrid (Mean = 464 ms, SD = 141.8 ms)

Table 7. ANOVA on log(Response Time) for VCS Task

Variable	DF	SS	F-value	p-value
DRT (Visual, Tactile)	1	0.104	2.280	0.133 (ns)
Task (Drive, 1-Back, Radio Easy, Radio Hard, Nav Easy, Nav Hard)	5	0.525	11.501	< 0.001
DRT x Task	5	0.146	3.207	0.009

The Tukey HSD test on VCS Task (Table 7) also showed in all cases, the 1-Back task had significantly higher mean DRT reaction times as compared to Drive, Radio Easy, Radio Hard Navigation Easy, and Navigation Hard.

3.3.3.2 Miss Rate

Figure 12 shows the aggregation of mean miss rate by VCS task and DRT type for all trials. For five of the eight tasks, RDRT had higher mean miss rates than the equivalent TDRT task. However, for Navigation Hard Audio, Radio Hard Hybrid , and Radio Hard Audio tasks, TDRT had higher mean miss rates.

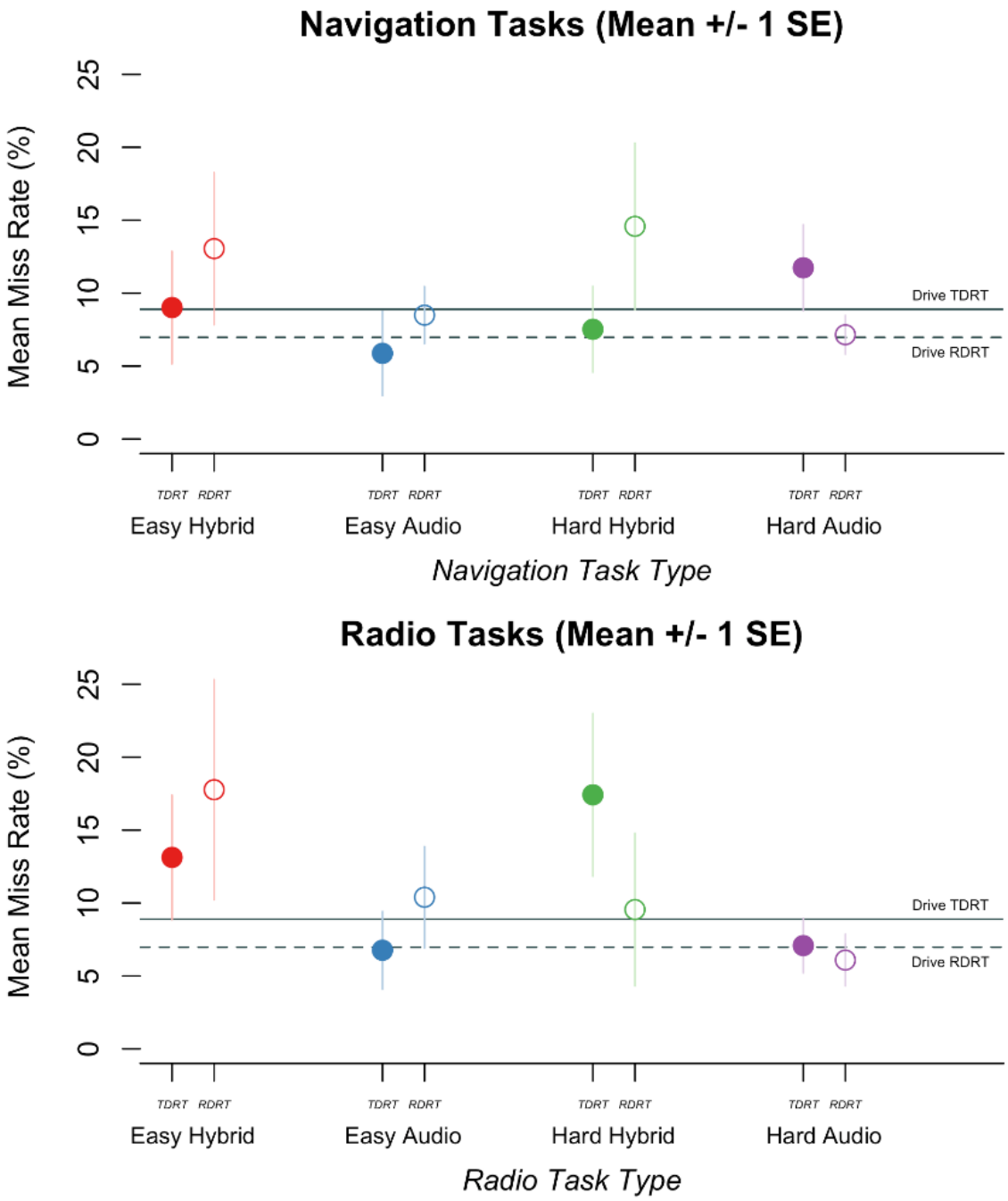


Figure 12. Miss Rate by VCS Task and DRT Type

Two negative binomial models on miss count per total stimulus events (i.e., miss rate) were fit, both with a random intercept on participant. The first model evaluated the fixed effects of VCS Mode and DRT. The second model evaluated the fixed effects of VCS Task Type and DRT.

Similar to the results from the ANOVAs, the effect of DRT type was not significantly different on miss count. In terms of VCS Mode, Hybrid tasks and 1-Back tasks were each associated with higher miss counts as compared to Audio Only tasks. These results are further provided in Table 8.

Table 8. Negative Binomial on Miss Count for VCS Mode

Variable	Estimate	SE	z-value	p-value
Intercept	-2.61	0.221	-11.804	< 0.001
DRT Visual	-0.15	0.118	-1.290	0.197 (ns)
Mode (reference Audio)				
Drive	0.04	0.167	0.242	0.809 (ns)
Hybrid	0.31	0.139	2.231	0.026
1-Back	1.17	0.228	5.140	< 0.001

The negative binomial on miss count for VCS Task Type is detailed in Table 9. The effect of DRT was not significant. Only the 1-back task had a significant effect above driving on miss count.

Table 9. Negative Binomial on Miss Count for VCS Task

Variable	Estimate	SE	z-value	p-value
Intercept	-2.58	0.248	-10.404	< 0.001
DRT Visual	-0.15	0.120	-1.242	0.214 (ns)
Task Type (reference Drive)				
Radio Easy	0.28	0.217	1.302	0.193 (ns)
Radio Hard	0.01	0.200	0.055	0.956 (ns)
Navigation Easy	-0.03	0.197	-0.177	0.860 (ns)
Navigation Hard	0.14	0.190	0.712	0.476 (ns)
1-Back	1.14	0.256	4.436	< 0.001

3.3.4 Eye Glance Behavior for Hybrid Tasks

Eye glances for the Hybrid VCS display mode tasks were evaluated to compare eyes-off-road time between these different task types (e.g., task difficulty). Audio Only tasks were not included in this analysis because the monitor was blacked out for the Audio Only display tasks.

3.3.4.1 Criterion 1: Percentage of Long Eyes-Off-Road (EOR) Glances

The distributions of all eyes-off-road glances for each participant is plotted in Figure 13. The percentage in the top right of each plot shows the percentage of the glances that were long eyes-off-road glances (≥ 2.0 seconds) for that participant. The "N" represents the total number of glances. Five of the nine participants exceeded the 15 percent threshold; one participant made 47 glances with 53 percent exceeding the 2.0 second threshold.

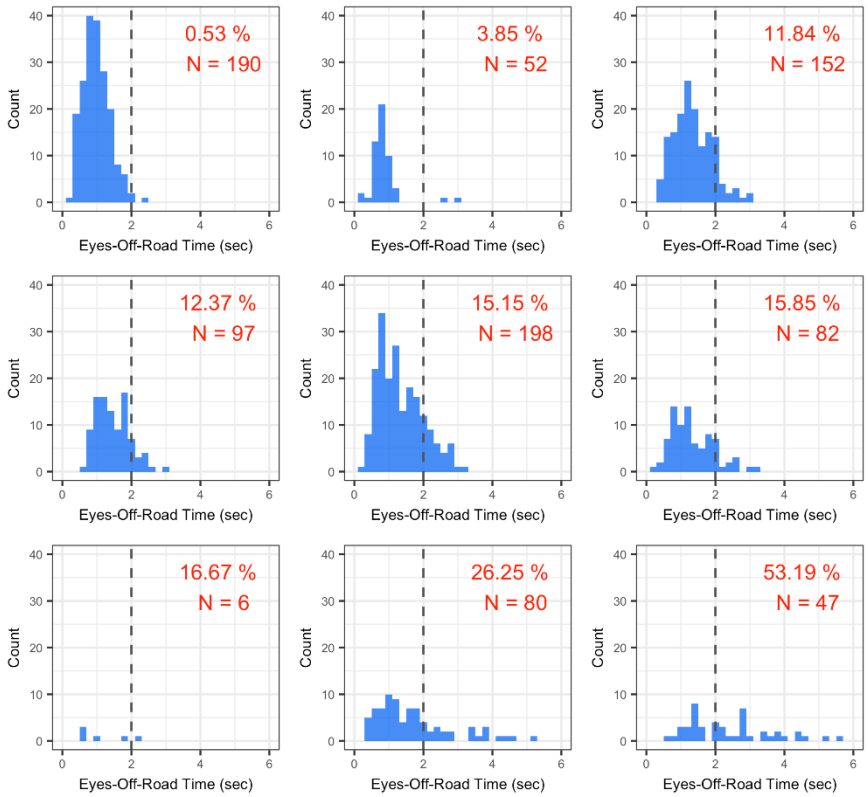


Figure 13. Distribution of All Eyes-Off-Road Glances by Participant

A similar aggregation of all eyes-of-road glances by task type is provided in Figure 14, with the percentage of long eyes-off-road glances and total number of glances provided in the top right of each plot. As shown in this figure, Radio Easy had a smaller percentage of long glances as compared to Radio Hard; Navigation tasks had a larger percentage of long glances as compared to radio tasks; Navigation Easy had the largest percentage of long glances, but Radio Hard had the most total number of glances.

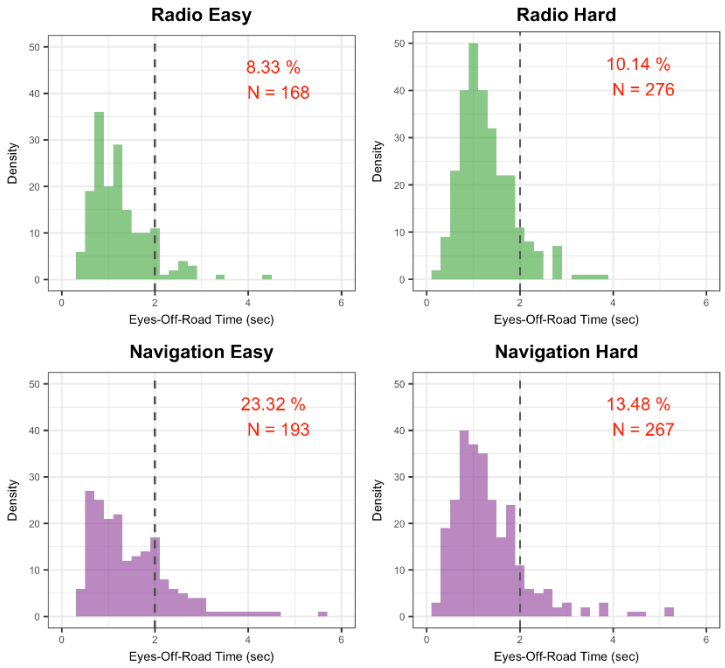


Figure 14. Distribution of All Eyes-Off-Road Glances by Task Type

The NHTSA threshold that eyes-off-road glances that exceed 2 seconds should not exceed 15 percent of glances was applied to understand how many participants did not meet this criterion. Long eyes-off-road glances were aggregated by participant for each task type and DRT for both the mean of the three repeated trials and the maximum (e.g., worst-case scenario) of the three trials. These results are tabulated in Table 10.

Table 10. Number of Participants that Do Not Conform with Percent Long EOR Glances

Task	**Difficulty**	**DRT**	**Number of Participants (out of 9) and % that Do Not Conform**			
			Mean of Trials		**Max Trial**	
Radio	Easy	Tactile	2	(22%)	3	(33%)
		Visual	2	(22%)	3	(33%)
	Hard	Tactile	2	(22%)	4	(44%)
		Visual	3	(33%)	7	(78%)
Navigation	Easy	Tactile	5	(56%)	7	(78%)
		Visual	5	(56%)	5	(56%)
	Hard	Tactile	3	(33%)	6	(67%)
		Visual	6	(67%)	6	(67%)

3.3.4.2 Criterion 2: Mean Glance Duration (MGD)

Figure 15 shows the boxplot for the distribution of mean eyes-off-road glance durations aggregated by participant for each of the VCS tasks and DRT types. Recall that each participant had three trials of each VCS task type, thus the top boxplot shows the MGD for each participant across all three trials. The lower boxplot shows the MGD for the single of the three trials that had

the largest MGD. Note that the trends are similar between DRT type across VCS task types. The boxplots indicate that the Navigation tasks had overall higher MGDs as compared to the Radio Easy tasks. The median bars within the boxplots for Navigation Easy tasks suggest that at least half of the MGDs exceeded the 2.0 second threshold.

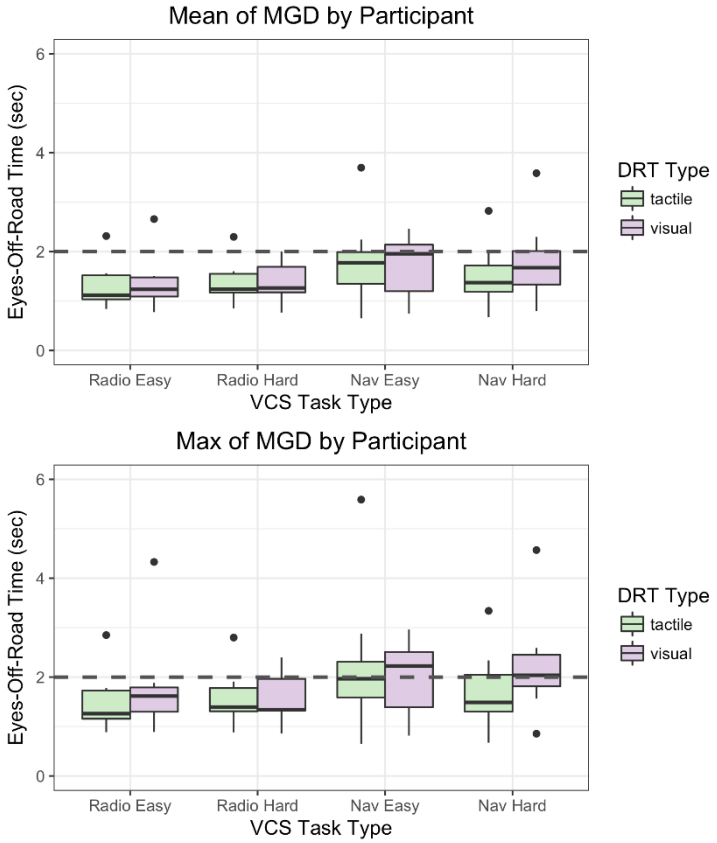


Figure 15. Boxplot of Mean Eyes-Off-Road Time by Task Type for Mean of All Trials (top) and Maximum Trial (bottom)

Two ANOVAs were performed on the MGD, one on the mean of all trials by participant (Table 11) and one of the trial with the maximum MGD by participant (Table 12). When aggregated across all trials by participant, there were no significant differences between DRT type on MGD, but task type ($p = 0.002$) was significant. When aggregated across the trial with the maximum MGD by participant, gender was not significant, but task type was significant ($p = 0.028$).

Table 11. ANOVA on log(MGD), Mean Across Trials by Participant

Variable	DF	SS	F-value	p-value
DRT (Visual, Tactile)	1	0.078	1.659	0.204 (ns)
Task (Radio Easy, Radio Hard, Nav Easy, Nav Hard)	3	0.816	5.808	0.002
DRT x Task	3	0.102	0.726	0.542 (ns)

Table 12. ANOVA on log(MGD), Max of Trials by Participant

Variable	DF	SS	F-value	p-value
DRT (Visual, Tactile)	1	0.245	3.089	0.085 (ns)
Task (Radio Easy, Radio Hard, Nav Easy, Nav Hard)	3	0.785	3.302	0.028
DRT x Task	3	0.235	0.989	0.406 (ns)

The NHTSA 2.0-second MGD for visual-manual tasks was used to assess the VCS tasks for conformance. The number of participants that did not conform are aggregated in Table 13 by task type and DRT.

Table 13. Number of Participants that Do Not Conform with MGD

Task	Difficulty	DRT	Number of Participants (out of 9) and % that Do Not Conform			
			Mean of Trials		Max Trial	
Radio	Easy	Tactile	2	(22%)	3	(33%)
		Visual	2	(22%)	3	(33%)
	Hard	Tactile	2	(22%)	4	(44%)
		Visual	3	(33%)	7	(78%)
Navigation	Easy	Tactile	5	(56%)	7	(78%)
		Visual	5	(56%)	5	(56%)
	Hard	Tactile	3	(33%)	6	(67%)
		Visual	6	(67%)	6	(67%)

3.3.4.3 Criterion 3: Total Eyes-Off-Road Time

The boxplot of TEORT by VCS task, parsed by DRT type, is provided in Figure 16. The upper boxplot represents the mean of all three trials for each participant, while the lower boxplot represents the trial with the maximum TEORT for each participant. Although the median (i.e., horizontal line within a box) across all tasks is below the recommended 12.0 second recommendation for visual-manual tasks, it is important to note that this is a voice control task and there are still several outliers (i.e., points) of participants exceeding 12.0 second threshold (Note: The outliers were not the same participants across tasks).

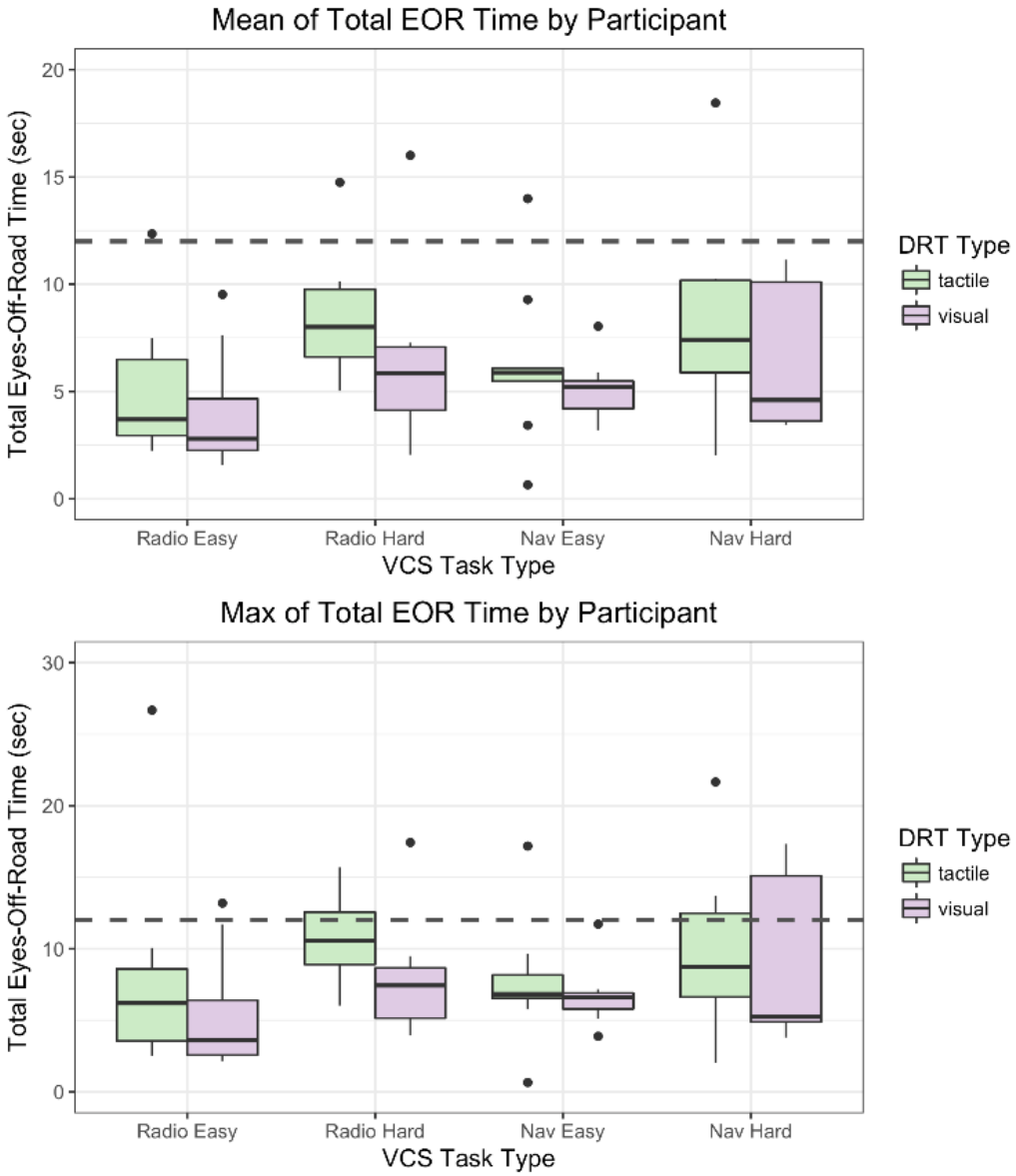


Figure 16. Total Eyes Off Road Time by Task Type for Mean of All Trials (top) and Maximum Trial (bottom)

The number of participants that did not comply with the NHTSA guideline of TEORT less than 12.0 seconds was aggregated by task type and DRT (see Table 14).

Table 14. Number of Participants that Do Not Comply with Total EOR Time

Task	Difficulty	DRT	Number of Participants (out of 9) and % that Do Not Comply			
			Mean of Trials		Max Trial	
Radio	Easy	Tactile	1	(11%)	1	(11%)
		Visual	0	(0%)	1	(11%)
	Hard	Tactile	1	(11%)	2	(22%)
		Visual	1	(11%)	1	(11%)
Navigation	Easy	Tactile	1	(11%)	1	(11%)
		Visual	0	(0%)	0	(0%)
	Hard	Tactile	1	(11%)	3	(33%)
		Visual	0	(0%)	3	(33%)

3.3.5 Subjective Measures of Performance

A survey was administered at the very end of each study session to the participant to evaluate subjective measures of driving performance and task difficulty (see Figure 17). All the participants rated their driving performance as somewhat safe or very safe while not driving distracted. The two Hybrid tasks, as compared to the three remaining tasks, had the lowest ratings of safety. In terms of task difficulty, more participants rated the Navigation Audio tasks as being harder than the other tasks. Counterintuitively, the Radio Hybrid had one of the lowest ratings of safety while also having one of the easiest ratings across all tasks.

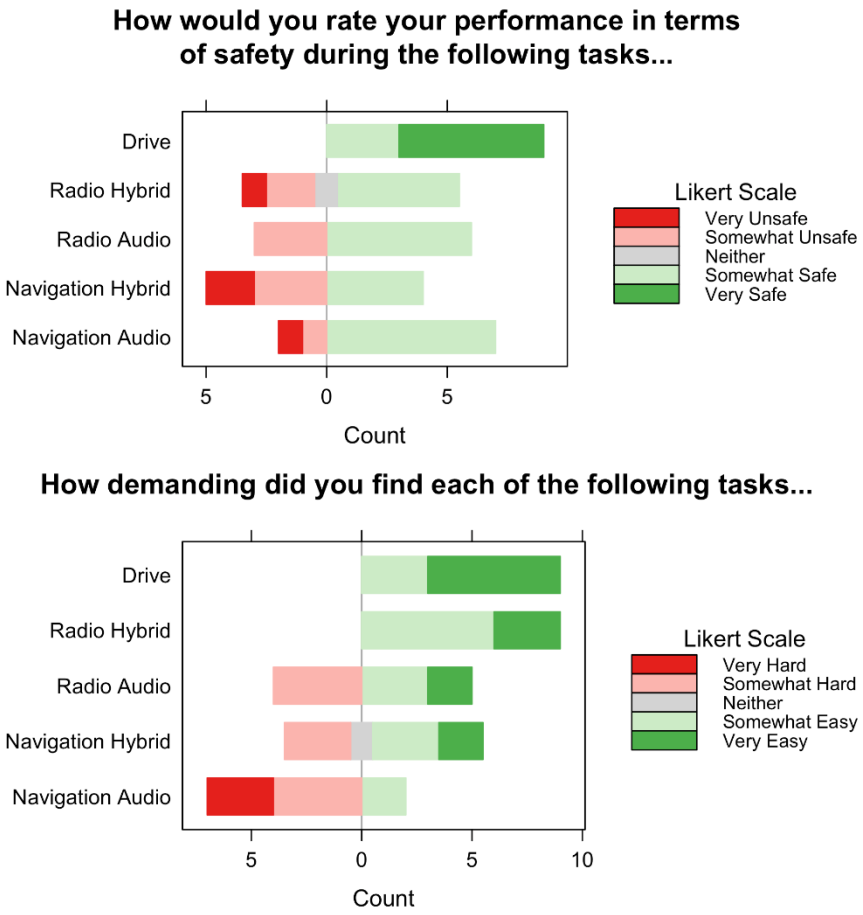


Figure 17. Subjective Measures of Driving Performance *(top)* and Task Difficulty *(bottom)*

3.4 DISCUSSION

This study explored a variant of the ISO DRT protocol using a new visual RDRT. The variant RDRT was evaluated for use with in-vehicle VCS by measuring the cognitive workload induced by various task types and the associated effects of workload on driving performance and eye glance behavior. This study also evaluated a new baseline task for assessing VCS workload.

Participants in this study completed 24 unique VCS tasks with workload measured using TDRT and the same 24 tasks with RDRT. There were four different tasks used in this study: Radio Easy, Radio Hard, Navigation Easy, and Navigation Hard. The Navigation tasks were the same as used in the previous Voice Part 1 study. The Radio tasks were designed to convey less information to the driver than the Navigation tasks. The Radio tasks, as compared to the Navigation tasks, were also designed to require less comparison between options and less memory load to make the correct decision. In addition to the different VCS task types, there were two different types of VCS display mode evaluated: Audio Only and Audio + Visual (Hybrid). Thus, the intent was to design the Radio Easy task as the least complex and the Navigation Hard task as the most complex.

Tasks with the visual display were completed quicker and more accurately than the equivalent Audio Only task, except for the Radio Easy (baseline) task. For this Radio Easy task, the Hybrid and Audio tasks were both completed with a mean accuracy of 94 percent. Within each respective display mode (Audio, Hybrid), Radio tasks were completed quicker than Navigation tasks. For audio tasks, all Easy tasks were completed quicker and with higher accuracy than their comparable Hard task (i.e., Radio Easy versus Radio Hard). These findings suggest that tasks of different complexity and modality have different completion times and accuracies. This also suggests that the Easy Radio task could be a good baseline candidate task as it was completed quickest and most accurately..

Participants self-reported Audio Only tasks as more demanding as compared to the equivalent Hybrid task type. However, participants on average rated themselves as safer while completing Audio Only tasks as compared to the equivalent Hybrid tasks.

The DRT type (i.e., TDRT versus RDRT) did not have a statistically significant effect on driving performance (i.e., SDLP or SD speed), cognitive workload (i.e., response time or miss rate), or eye glance behavior (MGD). However, the DRT type did impact MGD; tasks with the TDRT had noticeably smaller values of mean off road glances as compared to tasks with the RDRT. When evaluating eye glance behavior, trends in plotted means and conformance with NHTSA Driver Distraction Guidelines (NHTSA, 2013) also suggested that differences may exist between DRTs. Specifically, total eyes off road durations were greater for TDRT on all tasks except Navigation Hard (most complex task). However, for percent of long glances (threshold 15%) and MGD (threshold 2.0 seconds), the RDRT tasks had a lower percentage of participants who met thresholds. Further evaluation with a larger sample size could confirm and help to explain the potential significance of these findings. Overall, these findings suggest that both types of DRT capture similar measures of cognitive workload and that neither interfere with driving performance in a simulator. However, the results support the notion that the modified RDRT

may be biased toward measuring visual attention, while TDRT may be a purer measure of cognitive attention.

The different VCS task types had a significant effect on driving performance (variations in speed and lateral position), cognitive workload (reaction time), and eye glance behavior (MGD). Similarly, the display modes had a significant effect on driving performance (variations in speed and lateral position) and cognitive workload (reaction time and miss count). On average, Navigation tasks had more long eyes-off-road glances and longer mean eye glance durations. This suggests that the complexity of information and modality of providing the information affects both driving performance and workload.

4 STUDY 2 – COMPARING POTENTIAL VCS EVALUATION MEASURES IN DRIVING SIMULATION AND ON-ROAD CONTEXTS

4.1 PURPOSE

The goal of Study 2 was to compare potential VCS evaluation measures, such as measures of cognitive load (TDRT) and visual attention (glance measures) collected in the laboratory with a driving simulator to the same measures collected on-road in a real vehicle. A second goal was to determine if these measures are sensitive to differences between VCS tasks. Data collected in Study 2 were also used for analysis of relative crash risk in Study 3.

4.2 METHOD

4.2.1 Participants

Nine people participated in the Study 2 including five women and four men. This sample size is sufficient to detect large to moderate effect sizes in VCS evaluation measures. All were 30 to 56 years old, and the average age of participants was 42. All participants had valid U.S. driver licenses, drove at least 3,000 miles per year, were native English speakers, and had previous exposure to VCS. By self-report, they were not currently taking any medication nor did they have any medical condition that would influence their driving performance. In addition, potential participants were screened by Westat's corporate background screening unit to ensure that their driving records were in good standing.

Participants were recruited via a posting on Westat's intranet page, and by word-of-mouth. No employees in Westat's Center for Transportation, Technology and Safety Research could participate. Participants were each compensated $50 the first session, $75 the second and third session, and $100 the fourth session, for a total possible compensation of $300. The study was reviewed and approved by Westat's Institutional Review Board for the protection of human research participants.

4.2.2 Study Design

Study 2 was a within-subjects design, with participants coming to Westat on four separate occasions for various types of data collection. The first session functioned as a practice session, which introduced the participants to the types of tasks they would be doing in the simulator and on the road.

During each of the three data collection sessions, the participant experienced a slightly different driving scenario. However, all scenarios centered on the concept of following a LV at a constant distance. Participants experienced each driving scenario both in the driving simulator and on the road.

The within-subjects study design included: 2 contexts (on-road versus driving simulator) x 3 driving scenarios x 8 VCS tasks (including a no-task baseline, driving-only condition) x 2 attempts (replication of tasks). The various VCS tasks and driving scenarios are described in the following sections.

4.2.3 Voice Control System (VCS) Tasks

The same participant-facing VCS "Wizard of Oz" style interface used in Study 1 was used for Study 2. As in Study 1, participants completed two voice control tasks: navigating to a restaurant and selecting a radio station. Unlike in Study 1, participants were only given one page of options for the Radio-Tuning tasks in Study 2. There were two modalities for the Navigation and Radio tasks: Audio Only and Hybrid, where the options were both read aloud and displayed on a screen next to the participant. There were both "Easy" and "Hard" Navigation tasks. In addition to these tasks, participants also completed a 1-Back task. Two different predefined task sequences were used in the study; participants with even participant ID numbers performed task order "A", and participants with odd ID numbers performed task order "B."

The seven voice tasks were:

1. Auditory Radio-Tuning task (Auditory Only),
2. Hybrid Radio-Tuning task (Visual and Auditory),
3. Easy Auditory Restaurant Navigation (Auditory Only, Clear Correct Restaurant),
4. Hard Auditory Restaurant Navigation (Auditory Only, Harder-to-Identify Correct Restaurant),
5. Easy Hybrid Restaurant Navigation (Auditory and Visual, Clear Correct Restaurant),
6. Hard Hybrid Restaurant Navigation (Auditory Only, Harder-to-Identify Correct Restaurant), and
7. 1-Back verbal task.

Each participant performed the seven VCS tasks two times. An eight baseline task (i.e., designated as "None") consisted of a period of driving without performing any voice tasks. The periods defined for "None" tasks were one-minute intervals that followed execution of the fourth and ninth task attempts within each driving scenario. The analyses included all tasks.

A researcher trained each participant on all of the VCS tasks during an initial practice session. Then, at the beginning of each subsequent data collection session, the researcher quickly reviewed each task to ensure the participant was familiar with all of them before data collection began.

4.2.4 Study Scenarios

Participants engaged in three different test protocols, or study scenarios (one per day). They completed each of these scenarios both in the driving simulator and on the road, usually on the same day. For scenario A, the participant followed the LV (driving at a fixed speed) while completing the VCS tasks and engaging in the TDRT. In scenario B only, the LV varied its speed slightly throughout the drive and the participant was supposed to follow at a constant distance and perform VCS tasks. In scenario C, the brake lights (shown in Figure 18) illuminated occasionally (although the vehicle did not actually slow down). Participants were instructed to perform VCS tasks and to tap their own brake pedal as quickly as they could as soon as they saw the brake lights illuminate on the red LV. Table 15 illustrates the details and key measures collected for each study scenario. The order of the study scenarios was counterbalanced for the nine participants. Additionally, the order that the participant completed the on-road portion and the simulator portion of the session was also counterbalanced. Each data collection session took approximately two hours to complete.

Table 15. Study Scenarios

Scenario	Lead Vehicle (LV) Characteristics	Participant's Driving Task	Key Measures
(A) LV – fixed speed	LV cruise control engaged at fixed speed of 60 mph	Follow the LV at a safe, constant distance	• TDRT (RT, hit rate) • Eye glances
(B) LV – variable speed	LV speeds up or slows down based on a predefined function	Follow the LV at a safe, constant distance	• Speed correlation (match between LV speed and Experimental vehicle) • Eye glances
(C) LV – brake light illuminated	LV cruise control engaged at 60 mph, brake lights illuminate occasionally (but LV does not change speed)	Follow the LV at a safe, constant distance, and tap brake pedal as quickly as possible whenever the LV brake lights illuminate	• Brake (tap) response time • Eye glances

Figure 18. Illuminated Brake lights used for Brake Light scenario on-road (left) and in the driving simulator (right)

4.2.5 Study Procedures

Recruitment and Introduction Session

A small number of participants were required for this study, so advertisements were placed only on Westat's intranet site. Friends and family of Westat employees were eligible to participate. Employees who worked in Westat's Center for Transportation, Technology and Safety Research or who were familiar with the study were ineligible. One Westat employee from a different part of the company participated. People interested in participating were screened over the phone by a researcher. Driving records were checked to ensure the participant had no infractions within the past five years. Sessions were completed on weekdays at 10 a.m. and 1 p.m. to avoid rush hour traffic. On-road sessions were not conducted when it was raining.

The first session functioned as a practice session, where the participant was introduced to the VCS tasks and completed the VCS tasks both while driving on the road and in the simulator. A researcher walked the participant through each task step by step and answered any questions. Participants first practiced the VCS tasks without driving. After they felt comfortable with the tasks, they practiced the VCS tasks first in the simulator (with the LV traveling at a fixed speed) and then while driving the instrumented experimental vehicle on local roads.

Participants were instructed to drive safely and treat all driving scenarios (both in the simulator and on the road) as if they were driving their own vehicle. After the introductory session, the researcher decided whether the participant's driving behavior was safe enough to continue through data collection sessions. While no participant's driving behavior was deemed unsafe during Study 2, researchers were instructed to cancel future appointments and inform the participant they were ineligible if they determined the participant as being incapable of completing the data collection sessions.

4.2.5.1 *Driving Simulator*

The driving simulator portion of the study was conducted using a fixed-base NADS miniSim (version 2.2). The driving environment was displayed on three 60" Vizio model M602i-B3 monitors positioned 2 m from the participant.

Once the participant was in the driver's seat, they were instructed to fasten their seatbelts. The researcher gave instructions specific to the day's session scenario (see Appendix B). Before starting data collection, participants were given 2 to 3 minutes to practice driving the session's specific simulator scenario. Participants were told to maintain a 2-second gap between themselves and the red LV. The LV appeared on the road in front of the participants several seconds after they start driving the simulator. Participants were instructed to drive at approximately 60 mph, but it was emphasized that it was more important to maintain a 2-second gap between the participant and the LV than it was to drive at a certain speed. If the simulator portion came before the on-road portion that day, the researcher reviewed the VCS tasks.

When the participant felt comfortable with the tasks and the driving scenario, the researcher started the recorded trial. The researcher would select the appropriate VCS series, depending on (a) the driving scenario and (b) the task order. About 60 seconds after the LV appeared on the road, the researcher initiated the first VCS task. The computer station was positioned such that the participant could not see that the researcher was controlling the voice control interface. The researcher initiated the next task 60 seconds after the participant had finished the previous task. Once all 14 tasks were completed, the researcher stopped the simulator. If the participant still needed to complete the on-road portion of the session, the researcher would escort them outside. If the simulator portion was the last of the day, the researcher gave the participant their incentive, had the participant sign their receipt and confirm the date and time of their next appointment.

On-Road

Participants drove an instrumented Subaru Outback research vehicle along approximately 35 miles of Maryland Route 200, also called the Intercounty Connector (ICC). The ICC (shown in Figure 19) is a divided, limited-access toll road in Montgomery and Prince George's counties in Maryland. The participants began VCS tasks at predefined, fixed locations along the route, first

driving east from I-370 and then exiting and turning around at Konterra Drive, and then returning on the ICC westbound. The locations of VCS tasks are shown in Figure 19. Circles indicate tasks performed while traveling east and squares indicate locations of tasks performed while traveling west. The cross-hatched areas of the route indicate approximately where baseline driving data, (i.e. Task = "None") were collected. The "None" tasks followed completion of the fourth task (eastbound) and the ninth task (westbound). To the extent possible, all tasks were performed on relatively straight or gently curving segments of the roadway.

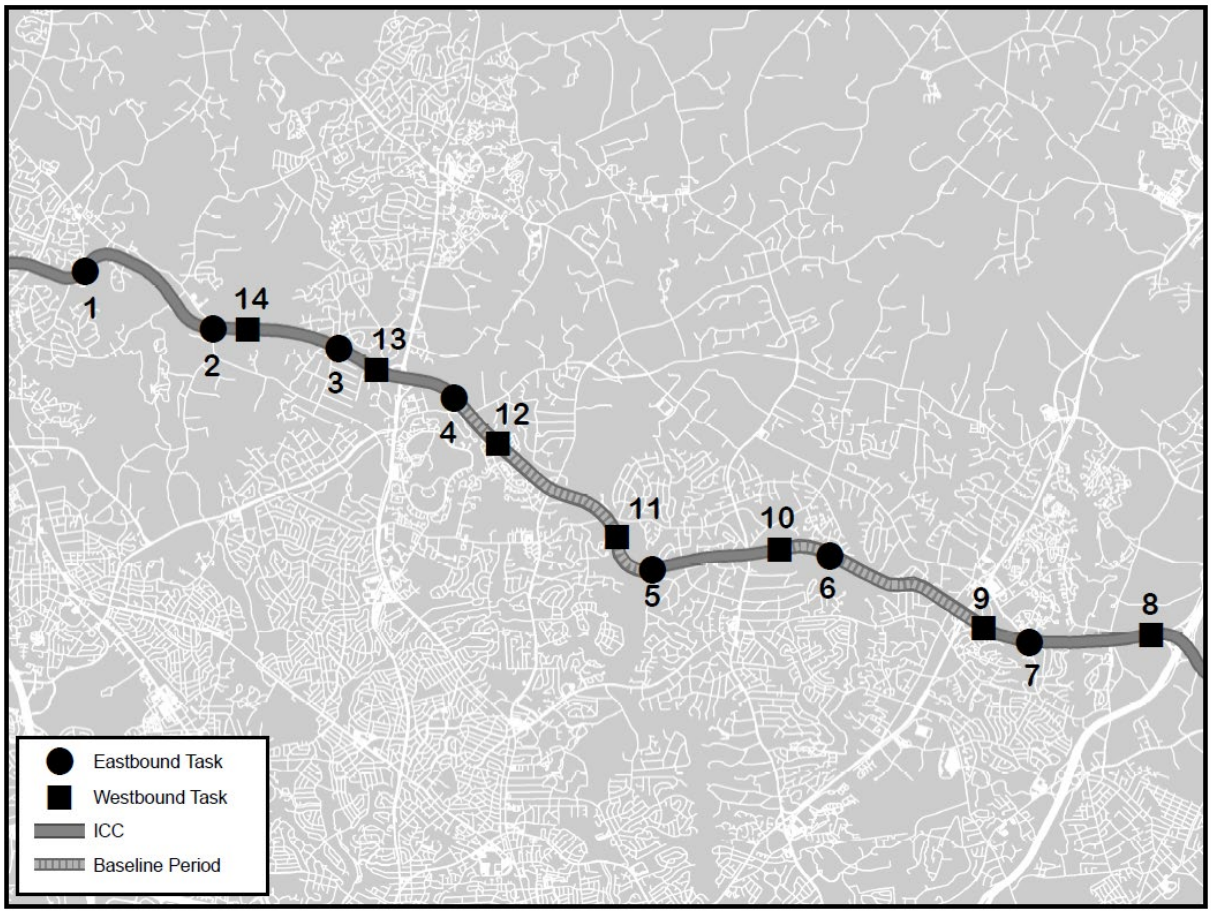


Figure 19. Section of the ICC (MD 200) used for data collection

The accompanying researcher instructed the participant to adjust their mirrors and seat, fasten their seatbelt, and drive safely. The researcher gave instructions specific to the day's scenario (see Appendix B). Participants were told to maintain a 2-second gap between themselves and the LV. If the on-road portion of data collection came before the data collection with the driving simulator that day, the researcher reviewed the VCS tasks with the instrumented vehicle parked. When the participant felt comfortable with the tasks, the researcher radioed the other members of the vehicle platoon that the participant was ready to start data collection. The participant followed the LV out of the Westat parking lot, eventually entering onto the ICC. The participant was instructed to follow the LV, but the accompanying researcher also provided appropriate directions to help the participant drive safely. The trailing vehicle followed the participant to prevent outside traffic from getting too close to the participant's vehicle.

The researcher selected the appropriate VCS series, depending on (a) the driving scenario and (b) the task order. VCS tasks were initiated according to ICC mile markers. The researcher in the participant's vehicle triggered the VCS tasks when they passed certain mile marker signs on the ICC, if the traffic situation was safe. When the participant finished a task, the accompanying researcher clicked a "Task Complete" button on the VCS laptop. This sent an audio message to the other two cars in the platoon to signify that the participant was no longer engaged in a task. Lane changes or maneuvers were never initiated when the participant was engaged in a task.

Figure 20. Researcher setting up an on-road data collection session

After the first seven VCS tasks, the platoon exited the ICC to turn around and drive back to complete the remaining tasks. Once the platoon arrived back at Westat, the researcher escorted the participant inside if they still needed to complete the driving simulator portion of the session. After the participant completed data collection for that day, the researcher gave the participant their incentive payment and confirmed the date and time of their next appointment.

4.2.6 Vehicle Platoon

To ensure safety on the road, participants completed on-road sessions in a vehicle platoon, shown in Figure 21. The platoon consisted of three vehicles: a LV (always a 2017 Chrysler 300), an instrumented vehicle (i.e., a 2011 Subaru Outback that was driven by the participant), and a trailing vehicle (i.e., vehicle type varied). The participant was tasked with following the LV at a safe distance. The trailing vehicle acted as a buffer for the participant to prevent outside traffic from interfering with the study. The platoon was outfitted with two-way communication between all researchers. For each on-road session, four researcher roles were needed to ensure a safe session:

- *Lead vehicle driver:* responsible for setting the pace of the platoon. For Fixed Speed (TDRT) and Brake Light scenarios, the LV driver drove at a steady speed using cruise control (usually 60 mph). For the Variable Speed scenario, the driver listened to the speed outputs coming from the LV laptop. The driver notched the cruise control speed up or down approximately every five seconds, driving between 58 and 62 mph.

- *Accompanying researcher*: responsible for instructing the participant where to drive, providing feedback to participants on driving behavior and initiating all VCS tasks. This researcher sat directly behind the participant's seat.
- *Trailing vehicle driver:* responsible for driving the trailing vehicle. The trailing vehicle driver prevented other cars on the road from driving behind the participant's vehicle, and provided a space buffer for other traffic in case the participant slowed down suddenly.
- *Safety observer*: located in the trailing vehicle and responsible for identifying and avoiding unsafe driving situations. The safety observer also initiated lane changes by radioing instructions to the LV driver and the accompanying researcher. As the platoon approached the designated mile marker where a task was supposed to start, the safety observer examined roadway conditions and surrounding traffic and informed the accompanying researcher over the radio when it was safe to start the task.

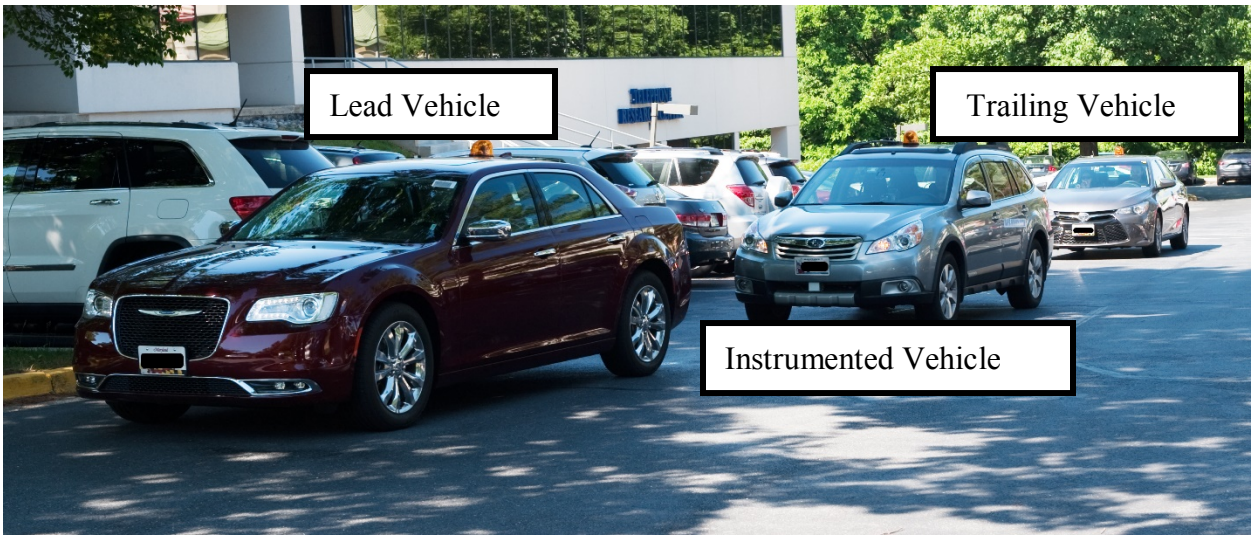


Figure 21. The three-vehicle platoon used for on-road data collection. Participants drove the instrumented vehicle.

4.2.7 On-Road Sessions Safety Initiatives

Protocols were followed to minimize and manage potentially unsafe situations on the road. Sessions were only run in the late morning and early afternoon to avoid rush hour traffic. All researchers were instructed that safety was their top priority. The LV driver was instructed to keep a safe distance from outside traffic and to slow down in the event they got too close to the car in front of them, until a lane change could be initiated safely.

Tasks were not initiated if an outside vehicle cut into the platoon. If an outside vehicle cut in behind the LV or the participant's vehicle, the safety observer instructed the LV to drop their speed slightly below the speed limit to encourage the intervening vehicle to switch lanes and pass the platoon. If the intervening vehicle stayed in the platoon and the safety observer did not observe any upcoming exits (the platoon was in the right lane for most of the time), the safety observer initiated a lane change to bring the platoon back together.

Researchers instructed participants that driving safely was their top priority in the study. This instruction was repeated throughout the study. The participant's Mobileye safety device (see Section 4.1.8 for description) was kept on for all sessions. If the participant was too close to the car in front of them or if they deviated from their lane without using their turn signal, the Mobileye emitted an auditory alert, serving as a warning to participants. If the accompanying researcher thought that the participant was not driving safely (i.e., driving too close to the LV), the researcher instructed the participant to modify their behavior. While no sessions needed to be cut short during this project, accompanying researchers were instructed to cease the on-road session if, at any time, they felt unsafe with the participant driving the instrumented vehicle.

4.2.8 Vehicle Instrumentation

The study used a 2011 Subaru Outback fitted with driver and vehicle monitoring equipment including a Mobileye C2-270 system. During the drive, video was captured from four viewpoints; two views of the driver's face to maximize glance coding accuracy, one view out the front window of the vehicle toward the LV, and one view over the participant's shoulder toward the VCS interface display monitor.

Before each session, technicians set up the three platoon vehicles with the necessary equipment. Laptops were used to capture vehicle data from the experimental vehicle and to run the VCS program. The VCS interface display monitor was a 7" Xenarc monitor. Mounted on each of the three vehicles were XBee Pro S3B antennae (Figure 22), ensuring that all researchers in the platoon were aware of what was happening in the experimental vehicle. GPS antennae were also mounted on the roof of the LV and the participant's instrumented vehicle.

Figure 22. Instrumentation on roof of LV

4.3 DATA REDUCTION

All driving simulator data and on-road vehicle data were reduced using SAS 9.4. On-road speed data for the LV and the instrumented vehicle were synchronized with time values based on GPS

signals. For the variable speed scenario, speed data from both vehicles were interpolated and down-sampled to 10 Hz to create a regularly-sampled and synchronized dataset.

To quantify participants' eye movements, researchers manually coded all driver face videos using Morae Manager software (Tech Smith). Two video coders reduced the eye glance data in accordance with SAE Standard J2396 for eye-glance reduction (SAE, 2000). The coders watched videos of the participant's face and inserted markers into the video stream whenever the eyes looked away from the forward view. For analysis purposes, only glances toward the system display were counted as off-road glances. Glances toward the vehicle's mirrors and speedometer were considered part of the driving task. They were infrequent, brief, and were not included in analyses of eyes-off-road time.

4.4 ANALYSIS AND RESULTS

4.4.1 Task Completion Times

Figures 23 shows the mean task completion time across the nine participants for each task in each scenario and context. Error bars indicate +/- 1 SEM. Task completion times across the on-road and driving simulator contexts followed a similar pattern, with the Hybrid Radio generally having the shortest task completion times and the Audio Navigation tasks having the longest tasks completion times. Note that all N-back tasks were the same length by design (30 seconds). These have been included in Figure 23 for reference. Tasks durations were similar across the three types of data collection sessions, which may indicate that participants did not deviate markedly in the way that they allocated their attention between the driving task and the VCS tasks in the different driving scenarios or two contexts.

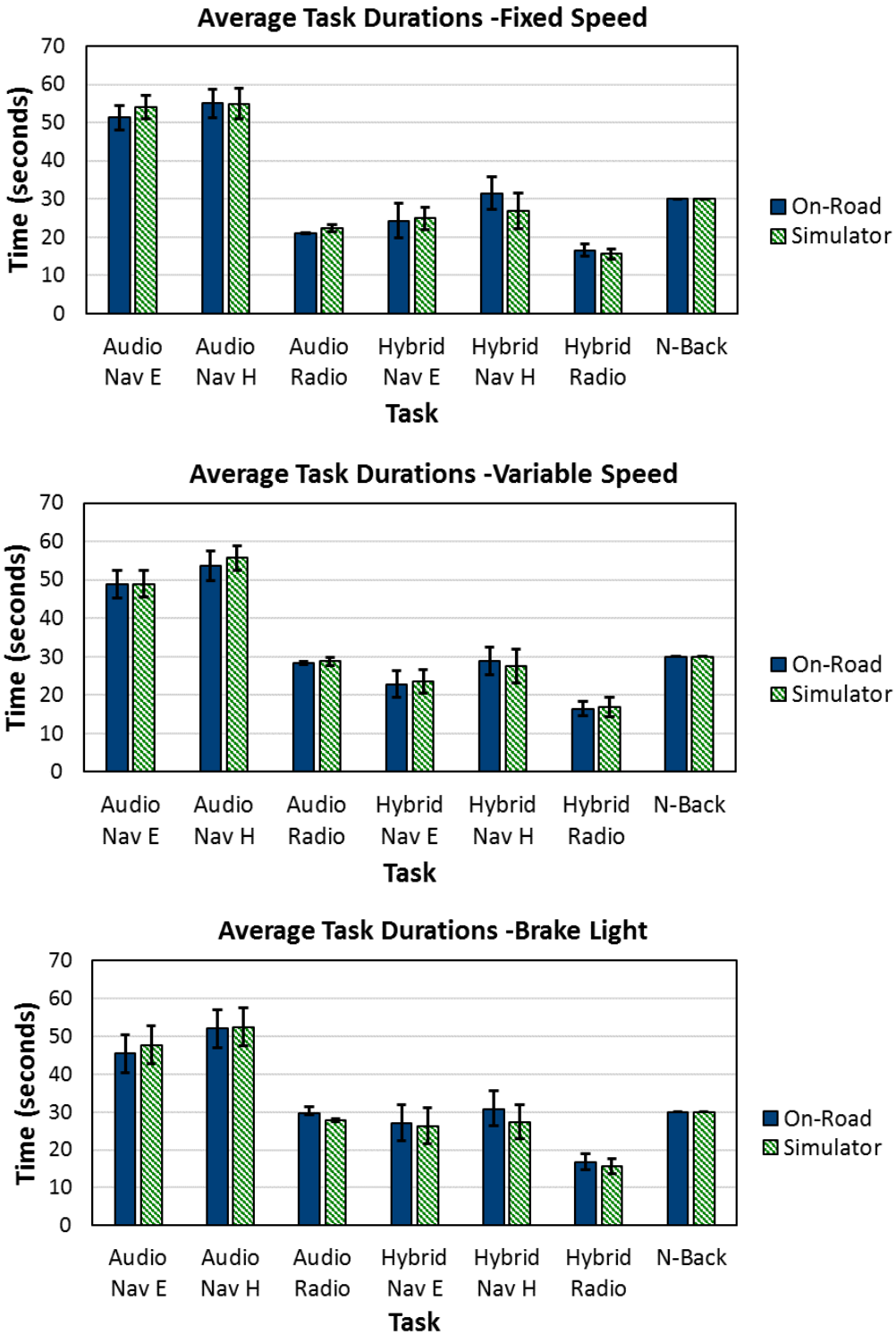


Figure 23. Average task durations, by task and by context for (from top) Fixed Speed, Variable Speed, and Brake Light scenarios

4.4.2 Tactile Detection Response Task – Fixed Speed Scenario

Throughout the Fixed Speed driving scenarios, participants were asked to tap the brake pedal as quickly as possible to a vibrating tactor as part of the TDRT. Longer response times on the TDRT generally indicated more cognitively demanding secondary tasks. TDRT response time data obtained during the Fixed Speed scenario were reduced by including only those responses that occurred during performance of VCS tasks, plus the baseline tasks periods. Responses that occurred more than 2500 ms after onset of the tactor vibration were excluded as misses in accordance with ISO Standard 17488. All valid response times were averaged within each task attempt for each participant. Then the mean of the two attempts was calculated for each participant per task. Finally, the mean and SEM across the nine participants for each task were calculated. These results are shown in Figure 24.

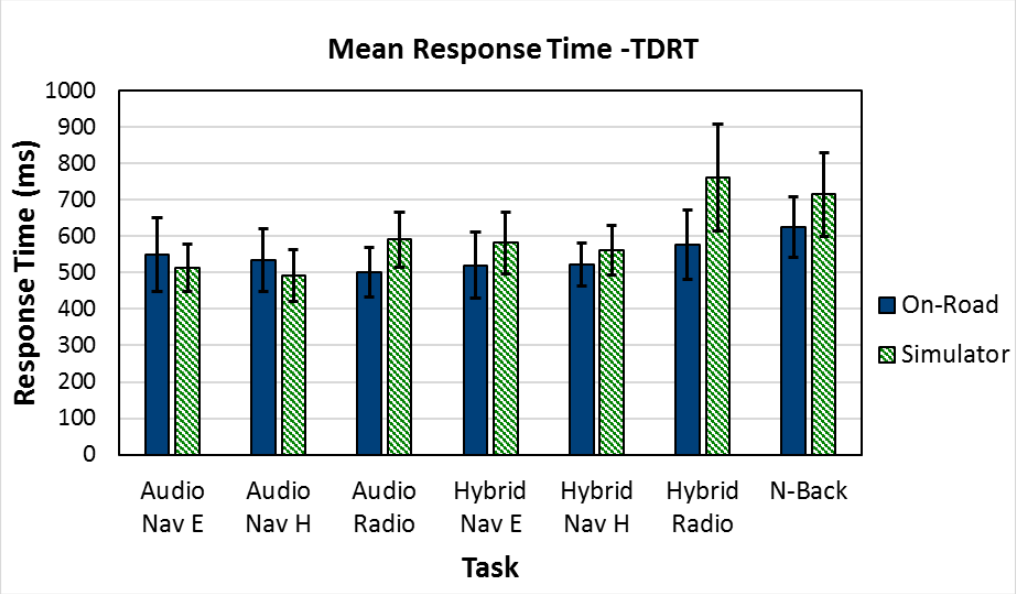


Figure 24. Comparison of mean TDRT response times by task and mode.

To examine differences in TDRT response times between data collection contexts and between tasks, all of the individual valid TDRT response times (n = 2153) were modeled using SAS 9.4 Proc Mixed. The model included the following fixed effects and two-way interactions.

- Context (2 levels, on-road or driving simulator)
- Task (8 levels, including 7 voice tasks plus baseline driving trials with no voice task)
- Attempt (2 levels, first or second time performing the task during the drive)
- Context*Task
- Task*Attempt
- Context*Attempt

A random effect, "Participant," was also included in the model to account for data clustered by participant.

There was a significant main effect of Task, $F(7, 2120) = 23.59$, $p < .0001$, but no significant main effect of Context, $F(1, 2120) = 0.01$, $p = .91$, or Attempt, $F(1,2120) = 0.61$, $p = .43$. There

was no significant interaction between Context and Task, F(7, 2120) = 1.65, p=.12 or between Task and Attempt, F(7,2120) = 1.69, p=.11. However, there was a significant interaction between Context and Attempt F(1, 2120) = 7.06, p = .008 which can be seen in Figure 25. While the differences are small (see Figure 25), they were useful for modeling in Study 3. When the context of the session was on-road, participants responded to the tactor faster on the first attempt of each task than on the second attempt of the task. However, the opposite was true during simulator sessions where participants responded to the tactor faster during the second attempt of a task. One possible explanation for this inconsistency could be that the on-road sessions took about twice as long as the simulator sessions, factoring in the time it took to leave Westat, drive to the ICC, complete the tasks, and drive back again. With the additional drive time, participants may have become fatigued such that their responses slowed by the time they completed the task the second time.

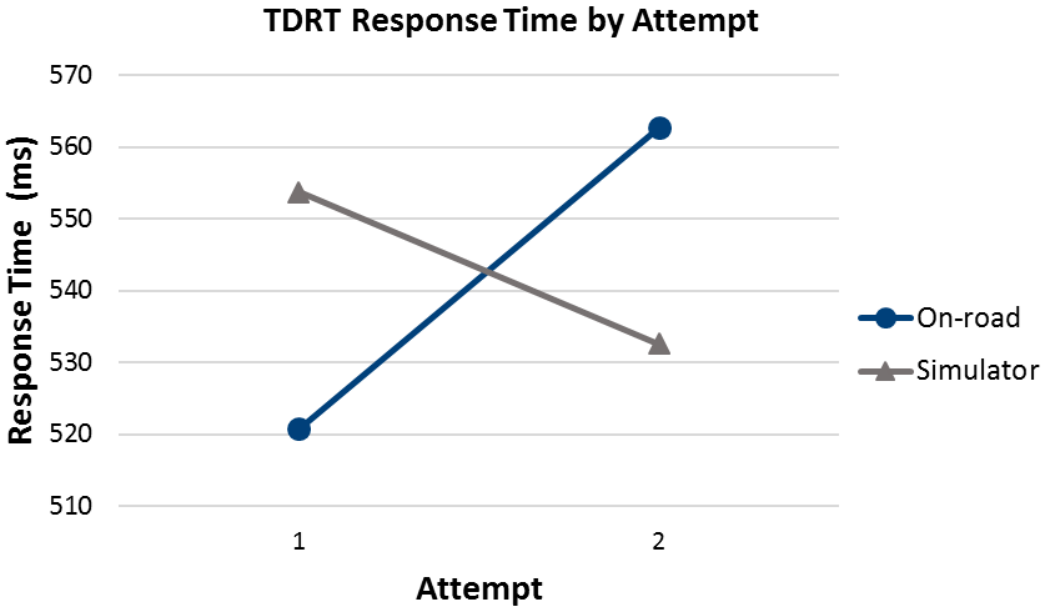


Figure 25. TDRT response time by attempt

Paired comparisons of the least squares (LS) means from the fit model were performed to examine differences between tasks. These LS means represent data collected in both contexts.

Note that in the analyses for Study 2 the p-values presented are **not** adjusted for multiple comparisons. Adjustments for multiple comparisons are not strictly necessary, especially when sample sizes are relatively small, the study is a pilot study intended to inform future research, and/or a finite number of preplanned comparisons are of interest (Gelman, Hill, & Yajima, 2012).

All of these are the case here. Since the power of this study is relatively low, not implementing a multiple comparisons adjustment makes it easier to identify, based on p-values, which differences are much larger than other differences and potentially of interest. However, this also means that it is critical to interpret all p-values with the usual caution about statistical testing using p-values: a threshold of 0.05 means that, on average, one would expect to see a false

positive (observe a p-value that is less than 0.05 in the data for a difference that is not actually statistically significant) in about 1 out of 20 tests. All p-values should be interpreted in conjunction with the other results presented in this report, such as descriptive statistics and plots. The p-values should primarily be treated as a useful guide for which results are worth investigating further rather than the final word on the importance of a given effect; this is consistent with the American Statistical Association's stance on the use of p-values in general (Wasserstein & Lazar, 2016).

The following list summarizes statistically significant differences ($p < .05$) between tasks for TDRT response times.

- "None" had a shorter response time than all other tasks Audio Navigation Easy, Audio Navigation Hard, Audio Radio, Hybrid Navigation Easy, Hybrid Navigation Hard, Hybrid Radio and N-Back)
- Audio Navigation Easy had a shorter response time than Hybrid Radio and N-Back
- Audio Navigation Hard had a shorter response time than Hybrid Radio and N-Back
- Audio Radio had a shorter response time than Hybrid Radio and N-Back
- Hybrid Navigation Easy had a shorter response time than N-Back
- Hybrid Navigation Hard had a shorter response time than Hybrid Radio and N-Back
- Hybrid Radio had a shorter response time than Hybrid Radio and N-Back

Mean response times for the periods of baseline driving with no VCS task were faster than response times for all the other voice tasks. The N-back task was associated with longer response times than all VCS tasks except for the Hybrid Radio task. The Hybrid Radio task had longer response times than all the other VCS tasks except for the Hybrid Navigation Easy task.

In addition to response time, another commonly used measure of TDRT performance is the "hit rate" (or "miss rate"). A "hit" for the TDRT is a response within 2500 ms after the onset of the tactile stimulus. As shown in Figure 26, hit rates were quite close to 100 percent for most tasks. Hybrid Navigation Easy, Hybrid Radio, and the N-Back tasks in the simulator, were the exception, with mean hit rates of about 93.3 percent, 91.6 percent, and 93.5 percent respectively. It is not clear why hit rates for some Hybrid tasks were slightly lower in the driving simulator context. It is possible that some participants felt more comfortable engaging with the visual aspects of the Hybrid tasks in the driving simulator as compared to driving on the road, and this deeper level of engagement in the simulator context may have resulted in more misses on TDRT.

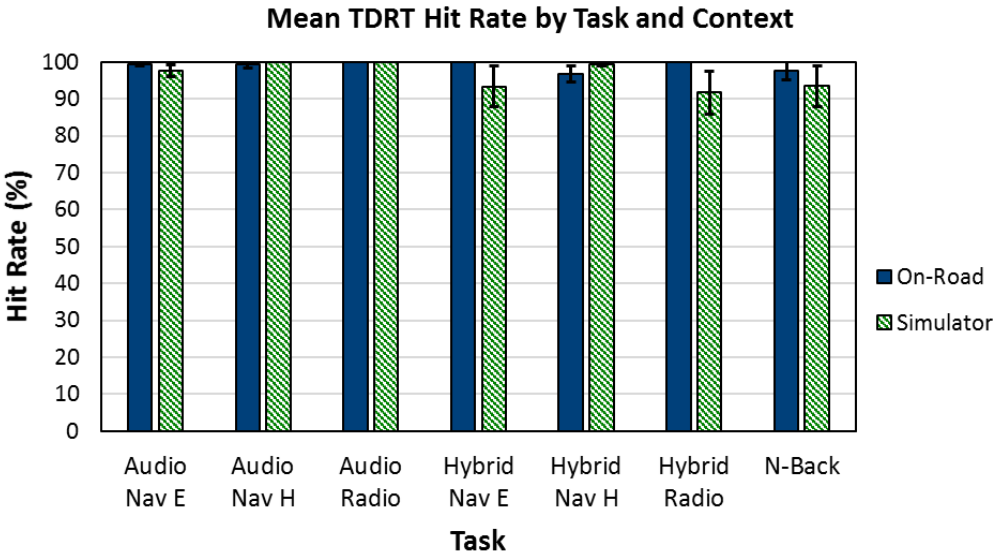


Figure 26. Comparison of TDRT hit rates by task and context.

4.4.3 Comparing the Participant's Speed to the Lead Vehicle Speed – Variable Speed Scenario

In the Variable Speed scenario, the participant had to try to maintain a constant following distance by responding to nearly continuous changes in the speed of the LV. This protocol has been used previously to assess potential driver distraction caused by performing secondary in-vehicle tasks (Ranney et al., 2011). We hypothesized that the participant's driving performance on this car following task would be affected by performing VCS tasks. We also hypothesized that VCS tasks that impose a higher workload on the driver would be associated with longer response times (time delay) for responding to speed changes of the LV. Thus, the correlation between the speed of the LV and speed of the participant's vehicle is one indicator of the quality of driving performance. Further, we hypothesized that a participant's delayed response to LV speed changes can be estimated by cross correlation of LV speed and participant's speed.

Each plot in Figure 27 shows the LV's speed in green, and the participant's speed in blue. The left-hand plot displays the raw data. In the right-hand plot, the participant's curve is shifted (advanced) by an optimal amount of time, so that the peaks and valleys of the two curves are aligned as much as possible. Similar alignment is possible by shifting the LV speed backward in time. The set of optimal time lags in LV speed obtained from this type of analysis can be compared between tasks.

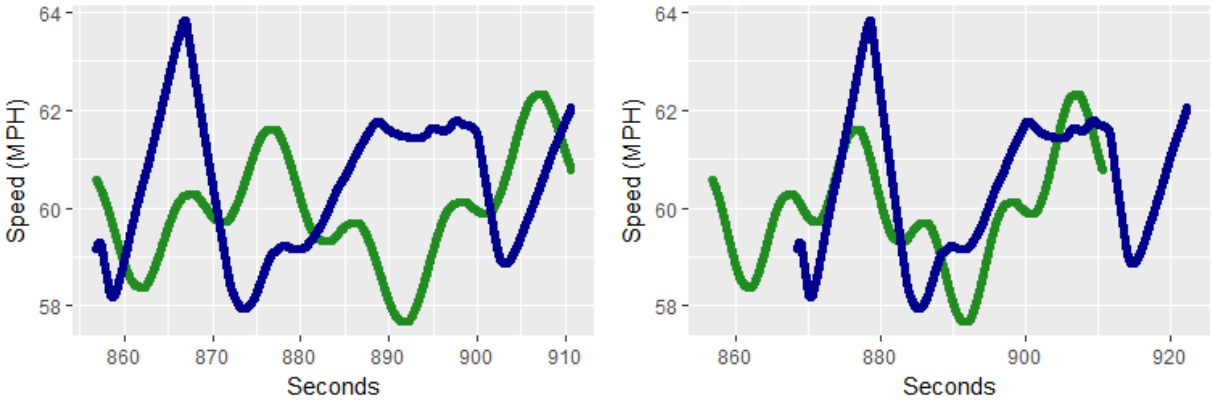
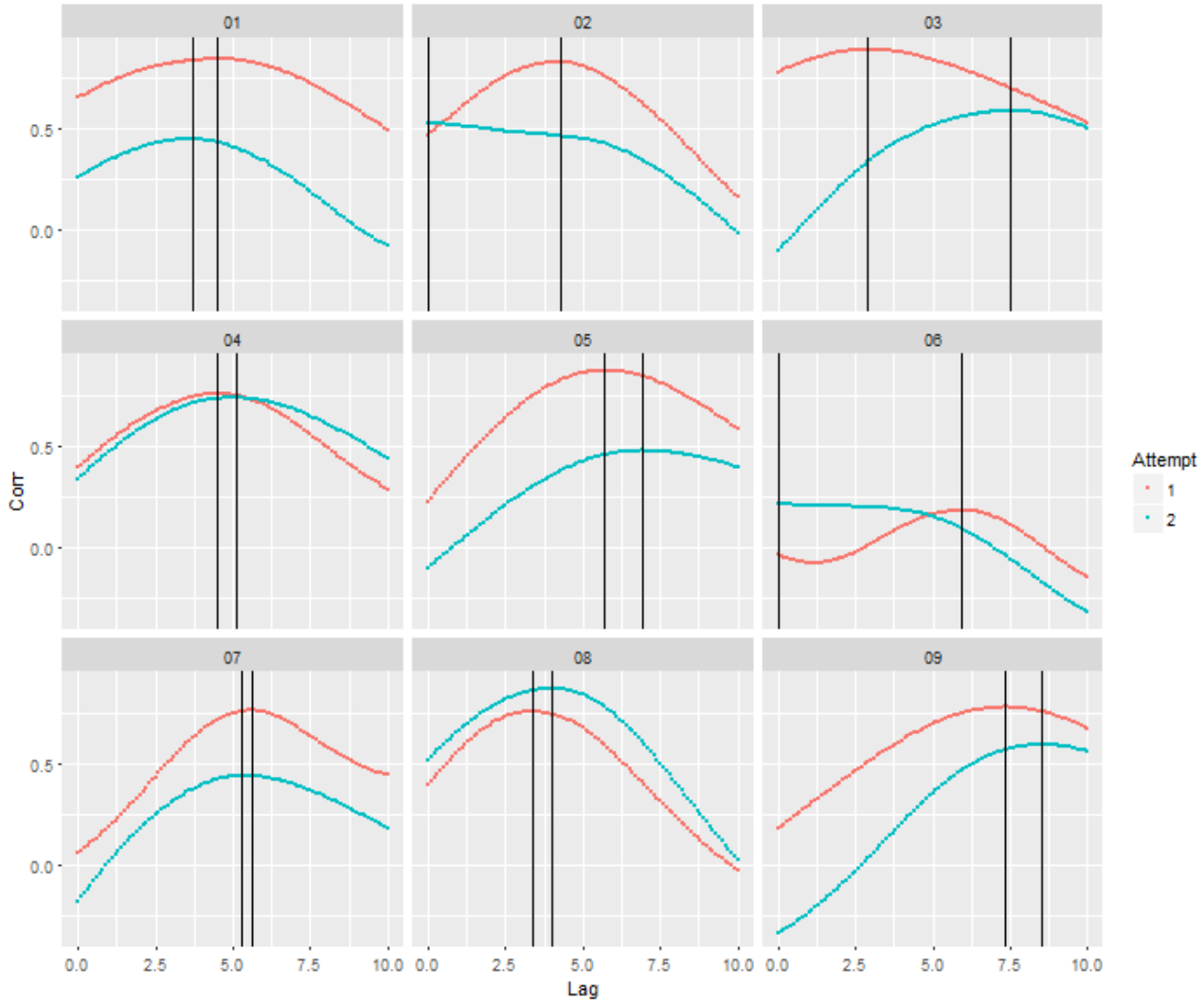


Figure 27. Plots of Participant 2, Auditory Navigation Hard task, Attempt 1. Raw (left) and shifted by optimal lag (right)

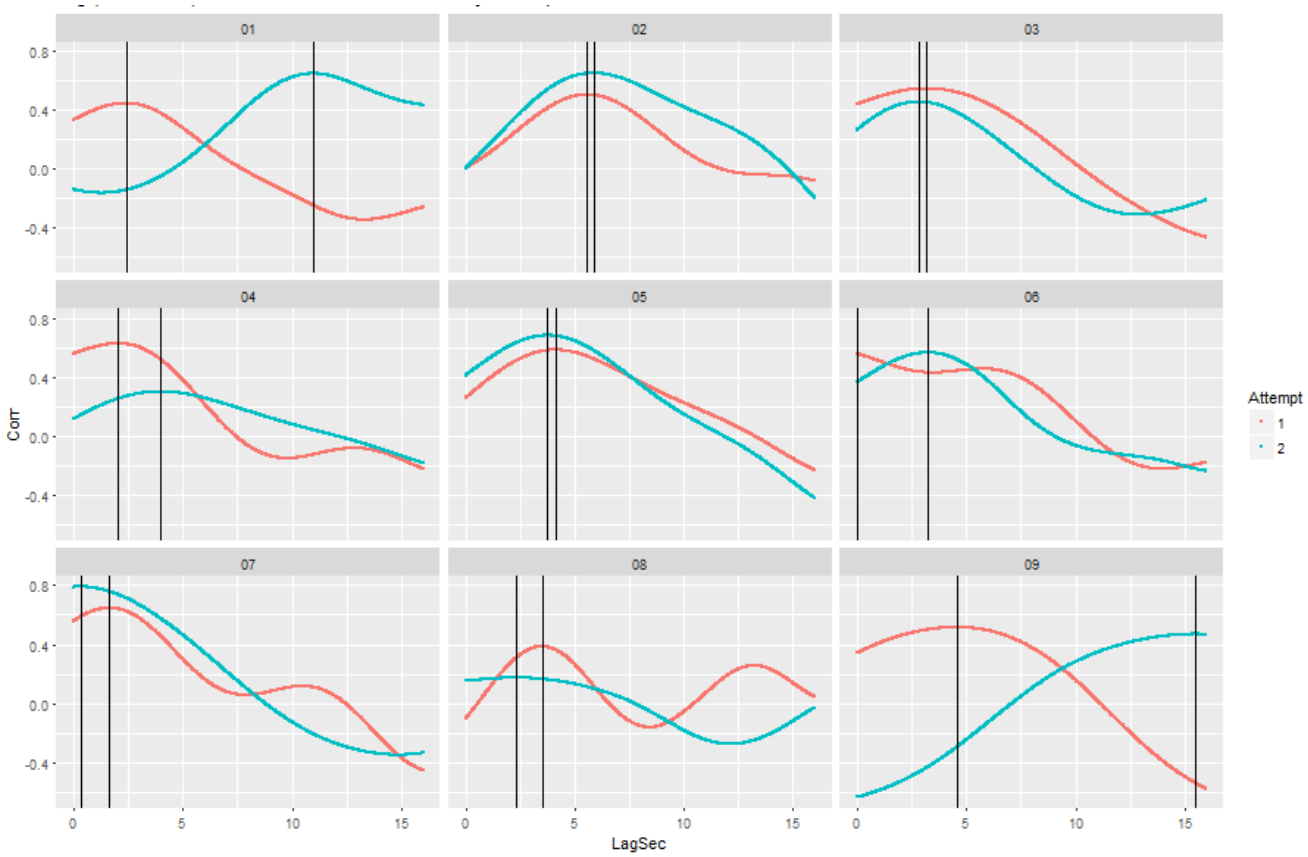
Optimal time lags in LV speed for each combination of participant, task, and attempt were computed separately for each context. The `ccf` function in R was used to compute the cross-correlation between the time series for the LV's speed and the participant's speed at all possible lags between 0 and 15 seconds. In the simulator context, this means testing 1,000 possible lags, since observations were recorded at a frequency of about one every 0.015 seconds; in the on-road context, there are 150 possible lags tested since speed measurements had been interpolated and synchronized to samples occurring every 0.1 seconds. The lag value with the maximum correlation was identified as the "optimal lag." This optimal lag value maximizes the cross-correlation between the vehicles' speed, meaning that at that lag the two-time series are optimally aligned. The final data files produced have one observation per combination of participant, task, and attempt, for a total of 9 x 8 x 2 = 144 observations per context. Seven simulator attempts were discarded because of issues with the LV's speed (either the speed was constant during the entire trial, or it shot up to a very high speed).

Figure 28 and Figure 29 illustrate how the correlation between LV speed and participant speed changes with various time lags in LV speed for each of the nine participants. These data were collected during periods of baseline driving when the participants were not performing any voice tasks. For each attempt, the thin black vertical line indicates the optimal lag where the correlation function reaches a maximum value. Note that there is substantial variability in these functions between participants and for some participants, a substantial difference between attempts. Similar variability across attempts was noted for speed data collected while participants performed voice tasks. Also note that for participant 6 (Figure 28) the maximum correlation never exceeds 0.25 at any lag examined. Other participants' data also exhibited very low maximum correlation values on some task attempts.



Note: Black vertical line indicates where the correlation function reaches a maximum value.

Figure 28. Lag (in seconds) versus correlation for 9 participants for baseline (none) task on-road.



Note: Black vertical line indicates where the correlation function reaches a maximum value.

Figure 29. Lag (in seconds) versus correlation for 9 participants for baseline (none) task in simulator

Figure 30 compares the optimal speed lags between tasks and between contexts. Generally, the median lag values for each task are similar for the on-road and simulator contexts and the pattern of median lag values across the tasks is similar for the two contexts. The box plots show that there is substantial variability in the lag measures within a task relative to the differences between tasks. Lags of approximately 5 seconds are consistent with the car following delays found by Ranney et al., 2011.

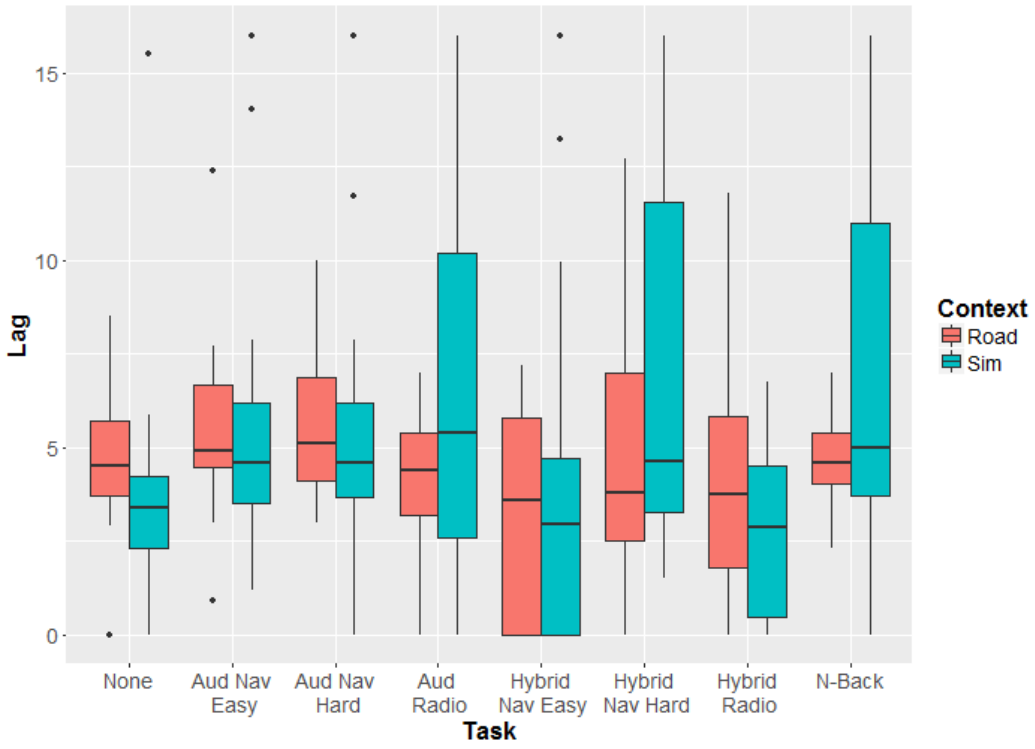


Figure 30. Median lag by task and context

Linear mixed models were fit to determine the relationship between optimal lag and task, controlling for within-participant effects. Models were first fit separately by context, of the form:

$$Lag = (\beta_0 + \beta_1 * Task) + (\gamma_0 + \gamma_1 * Task)$$

where γ_0 is a random intercept parameter and γ_1 is a random slope parameter, both dependent on within-participant effects. Both γ_0 and γ_1 are assumed to be normally distributed with mean 0 and unknown variances (to be estimated during modeling), and we further assume some nonzero correlation ρ between γ_0 and γ_1. The random intercept means that we expect some additive shift in lag that varies by participant, and the random slope term allows the effect of each task on lag to vary by participant. Similar models were also fit separately for each context using the maximum cross-correlation as the outcome.

All models were fit using the `lmer` function in R. Before model fitting, observations with a cross-correlation of less than 0.2 were dropped from the models with lag as the outcome. Visual

inspection of plots indicated that a maximum cross-correlation of 0.2 or less meant that the participant was not tracking the LV well, and there was no obvious optimal lag value—the two-time series would be misaligned at any lag. In such cases the maximum cross-correlation typically occurred at either 0 or 15, the endpoints of the allowable interval. Eight such observations were dropped from the simulator data, and four observations were dropped from the on-road data. Observations with a low cross-correlation were not dropped from the models using correlation as outcome, however, since low correlations are informative in these models.
An overall model, grouping observations from both contexts, was also fit:

$$Lag = (\beta_0 + \beta_1 * Task + \beta_2 * Context) + (\gamma_0 + \gamma_1 * Task)$$

This model adds a fixed main effect for context. Models including a fixed interaction term between task and context and allowing context to interact with the random effect for participant were also considered, but due to the limited number of participants, these more-complex models failed to converge properly. A parallel model with cross-correlation as the outcome was also fit, but even this simpler model failed to converge. The final overall model for correlation across contexts drops the random slope term $\gamma_1 * Task$.

After each model was fit, all pairwise differences between the LS means of tasks were calculated using the `difflsmeans` function from the `lmerTest` package in R. Satterthwaite's approximation was used to calculate degrees of freedom for each test.

Results for the driving simulator, using lag as the outcome, are presented in the table below. Each row corresponds to a difference between two tasks (e.g., the first row is for the estimate of the difference between None and the Audio Navigation Easy task). The estimate of -1.6 means that the lag for Audio Navigation Easy is slightly longer (and therefore this task is likely more difficult), on average, than the lag for None; however, the p-value of 0.25 indicates that this difference is not statistically significant at the 0.05 level. Note that the p-values presented are not adjusted for multiple comparisons. Also note that the statistical power of these comparisons appears to be quite low, such that differences in lags less than approximately 3 seconds did not reach statistical significance. For the on-road data, differences less than approximately 2 seconds usually did not reach statistical significance.

Table 16. Driving Simulator Context, Estimating Differences in Speed Lag

	Estimate	Std Err	t-value	Lower CI	Upper CI	p-value
None - Aud Nav Easy	-1.6	1.36	-1.18	-4.41	1.20	0.25
None - Aud Nav Hard	-1.6	1.34	-1.19	-4.33	1.14	0.24
None - Aud Radio	-3	1.63	-1.84	-6.56	0.57	0.09
None - Hybrid Nav Easy	0	1.38	0.02	-2.84	2.88	0.99
None - Hybrid Nav Hard	-3.6	1.47	-2.47	-6.78	-0.49	0.03
None - Hybrid Radio	1	1.51	0.67	-2.17	4.20	0.51
None - N-Back	-3.3	1.69	-1.98	-7.28	0.60	0.09
Aud Nav Easy - Aud Nav Hard	0	1.45	0.01	-3.10	3.12	0.99
Aud Nav Easy - Aud Radio	-1.4	1.74	-0.8	-5.31	2.52	0.44
Aud Nav Easy - Hybrid Nav Easy	1.6	1.59	1.03	-1.82	5.07	0.32
Aud Nav Easy - Hybrid Nav Hard	-2	1.70	-1.19	-5.86	1.80	0.26
Aud Nav Easy - Hybrid Radio	2.6	1.34	1.96	-0.08	5.32	0.06
Aud Nav Easy - N-Back	-1.7	1.70	-1.02	-5.67	2.20	0.34
Aud Nav Hard - Aud Radio	-1.4	1.39	-1.01	-4.31	1.50	0.32
Aud Nav Hard - Hybrid Nav Easy	1.6	1.41	1.14	-1.36	4.59	0.27
Aud Nav Hard - Hybrid Nav Hard	-2	1.34	-1.52	-4.78	0.71	0.14
Aud Nav Hard - Hybrid Radio	2.6	1.55	1.69	-0.69	5.90	0.11
Aud Nav Hard - N-Back	-1.7	1.47	-1.19	-4.95	1.46	0.26
Aud Radio - Hybrid Nav Easy	3	1.71	1.77	-0.87	6.92	0.11
Aud Radio - Hybrid Nav Hard	-0.6	1.44	-0.44	-3.70	2.44	0.67
Aud Radio - Hybrid Radio	4	1.75	2.29	0.15	7.88	0.04
Aud Radio - N-Back	-0.3	1.37	-0.25	-3.23	2.55	0.81
Hybrid Nav Easy - Hybrid Nav Hard	-3.7	1.39	-2.62	-6.55	-0.76	0.02
Hybrid Nav Easy - Hybrid Radio	1	1.79	0.55	-2.93	4.91	0.59
Hybrid Nav Easy - N-Back	-3.4	1.83	-1.84	-7.45	0.73	0.10
Hybrid Nav Hard - Hybrid Radio	4.6	1.84	2.53	0.57	8.73	0.03
Hybrid Nav Hard - N-Back	0.3	1.60	0.18	-3.16	3.75	0.86
Hybrid Radio - N-Back	-4.4	1.66	-2.62	-8.05	-0.66	0.03

Table 17. On-Road Driving Context, Estimating Differences in Speed Lag

	Estimate	Std Err	t-value	Lower CI	Upper CI	p-value
None - Aud Nav Easy	-1.20	0.79	-1.52	-2.81	0.39	0.14
None - Aud Nav Hard	-1.30	0.84	-1.49	-3.01	0.51	0.15
None - Aud Radio	-1.30	1.06	-1.20	-3.66	1.12	0.26
None - Hybrid Nav Easy	0.70	0.83	0.90	-0.96	2.46	0.38
None - Hybrid Nav Hard	-1.80	1.01	-1.78	-4.03	0.44	0.10
None - Hybrid Radio	0.60	0.92	0.65	-1.37	2.58	0.52
None - N-Back	-1.50	0.97	-1.54	-3.65	0.65	0.15
Aud Nav Easy - Aud Nav Hard	0.00	0.85	-0.05	-1.83	1.74	0.96
Aud Nav Easy - Aud Radio	-0.10	1.05	-0.06	-2.44	2.31	0.95
Aud Nav Easy - Hybrid Nav Easy	2.00	0.87	2.24	0.09	3.82	0.04
Aud Nav Easy - Hybrid Nav Hard	-0.60	0.99	-0.59	-2.76	1.58	0.56
Aud Nav Easy - Hybrid Radio	1.80	0.94	1.92	-0.26	3.88	0.08
Aud Nav Easy - N-Back	-0.30	0.91	-0.32	-2.24	1.67	0.76
Aud Nav Hard - Aud Radio	0.00	0.88	-0.02	-1.87	1.84	0.98
Aud Nav Hard - Hybrid Nav Easy	2.00	0.83	2.41	0.30	3.70	0.02
Aud Nav Hard - Hybrid Nav Hard	-0.50	0.92	-0.59	-2.51	1.42	0.56
Aud Nav Hard - Hybrid Radio	1.90	0.86	2.15	0.06	3.66	0.04
Aud Nav Hard - N-Back	-0.20	0.89	-0.27	-2.14	1.65	0.79
Aud Radio - Hybrid Nav Easy	2.00	1.07	1.89	-0.38	4.42	0.09
Aud Radio - Hybrid Nav Hard	-0.50	1.12	-0.47	-3.05	1.99	0.65
Aud Radio - Hybrid Radio	1.90	0.88	2.14	0.04	3.71	0.05
Aud Radio - N-Back	-0.20	0.87	-0.26	-2.07	1.61	0.80
Hybrid Nav Easy - Hybrid Nav Hard	-2.50	0.90	-2.84	-4.46	-0.63	0.01
Hybrid Nav Easy - Hybrid Radio	-0.10	0.99	-0.14	-2.29	2.00	0.89
Hybrid Nav Easy - N-Back	-2.20	1.07	-2.11	-4.64	0.15	0.06
Hybrid Nav Hard - Hybrid Radio	2.40	1.19	2.02	-0.28	5.08	0.07
Hybrid Nav Hard - N-Back	0.30	1.15	0.26	-2.27	2.87	0.80
Hybrid Radio - N-Back	-2.10	0.87	-2.43	-3.96	-0.25	0.03

The following list summarizes statistically significant differences ($p < .05$) between tasks in the on-road and driving simulator contexts.

On-Road Context Findings:

- Hybrid Navigation Easy had a smaller lag than Audio Navigation Easy
- Hybrid Navigation Easy had a smaller lag than Audio Navigation Hard
- Hybrid Navigation Easy had a smaller lag than Hybrid Navigation Hard
- Hybrid Radio had a smaller lag than Audio Navigation Hard
- Hybrid Radio had a smaller lag than N-Back

Simulator Context Findings

- Hybrid Navigation Easy had a smaller lag than Hybrid Navigation Hard
- Hybrid Radio had a smaller lag than N-Back
- Hybrid Radio had a smaller lag than Radio Audio
- Hybrid Radio had a smaller lag than Hybrid Navigation Hard
- Baseline had a smaller lag than Hybrid Navigation Hard

The overall difference between contexts, estimated by a model that included main effects of context and tasks but no interaction, was approximately 1 second. That is, the simulator data required a lag approximately one second longer than the on-road data. We speculate that speed changes of the LV may be slightly more difficult to detect visually in the simulator than on-road.

A similar analysis was conducted on a combined set of simulator and on-road data to model the magnitude of the correlation at each optimal lag. Estimated parameters for this model indicated a significant effect of context, such that simulator data generally had speed correlations that were weaker than those observed with on-road data. Multiple comparisons between tasks indicated significant differences ($p < .05$) between some pairs of tasks as shown in the list below.

- Hybrid Navigation Easy had a smaller lag than Audio Navigation Easy
- Hybrid Navigation Hard had a smaller lag than Audio Navigation Hard
- Hybrid Navigation Hard had a smaller lag than Audio Radio
- Hybrid Navigation Hard had a smaller lag than Hybrid Navigation Easy
- Hybrid Navigation Hard had a smaller lag than N-Back

4.4.4 Brake Light Response Times – Brake Light Scenario

In the Brake Light scenario, participants were instructed to tap the brake pedal as quickly as possible whenever they saw the LV's brake lights illuminate ahead of them. Conceptually, this task is similar to the modified RDRT that was used in Study 1.

The LV did not actually slow down when its brake lights were triggered, so the participants responded strictly to the onset of the illuminated brake lights and not to any looming cues that would normally be present if the LV were braking. There was one brake light event for each task. It occurred between 5 and 20 seconds after the start of the task. Brake light events also occurred twice in between tasks at predefined locations. The response times are displayed in Figure 31. The two brake light events that occurred in between tasks were considered baseline responses (labeled as the "None" task in Figure 31). Only one or two valid responses for the "None" task periods were obtained and included in the analysis for each participant in each context.

Mean response times per task were calculated for each participant by averaging responses across two attempts (if valid data were available). Response times that exceeded 5 seconds were not considered valid and were not included in the analyses because it was not clear whether long brake response times were made in response to the LV brake light illumination or were simply a vehicle control action made by the participant. The mean response time and standard error were then calculated across participants for each task. These are shown in Figure 31.

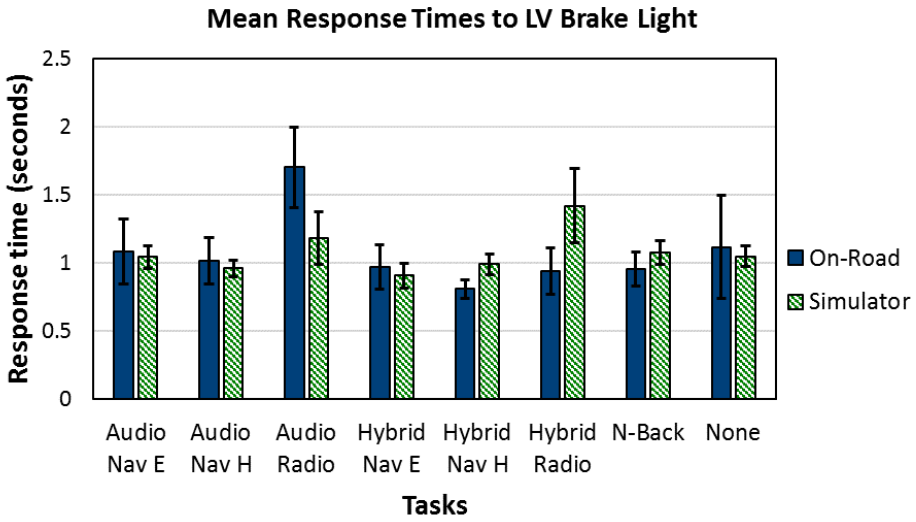


Figure 31. Comparison of mean brake response times to LV light by task and context.

Brake light response times were similar between on-road and simulator contexts across most tasks. One exception was the Audio Radio task, where the on-road responses were about 0.5 seconds longer than the simulator responses. There was also a difference in response times with the Hybrid Radio task in the opposite direction, with the simulator responses being about 0.5 seconds longer than the on-road responses. The large standard error of the baseline (none) task may be an indication of confusion during the on-road trials. Researchers noted that a few participants seemed to either not see the brake light in between tasks or were slow to respond. While participants were instructed that the brake light events would occur throughout the run, it is possible that participants let their guard down once they finished a voice task. If the researcher noticed that a participant missed a brake light event, the researcher reminded the participant to respond.

Individual response times were modeled with SAS 9.4 Proc Mixed. Because brake light events occurred only once per task attempt and only occasionally between tasks, the data sample was small. A mixed model was estimated with fixed effects of task and context, and random effects of participants. There was a significant effect of Task $F(7, 211) = 2.20$, $p = .04$, but not of context $F(1, 211) = 0.06$, $p = .81$. A separate model, which included an effect for the interaction between Context and Task was estimated, but in that case the Context*Task effect was not significant $F(7, 204) = 1.14$, $p = .34$. Another model included as a covariate the inverse of following distance, defined as 1/distance gap in feet, between rear bumper of LV and front bumper of

following vehicle. The inverse following distance was not significant F(1,210) = 0.26, p = .61. This finding was not consistent with prior research (Donmez, Boyle, & Lee, 2007).

Paired comparisons were conducted to examine differences between tasks combined across contexts. The following list provides a summary of significant (p < .05) findings with brake light response times.

- Audio Navigation Easy had a shorter response time than Audio Radio
- Audio Navigation Hard had a shorter response time than Audio Radio
- Hybrid Radio Easy had a shorter response time than Audio Radio
- Hybrid Radio Hard had a shorter response time than Audio Radio
- N-Back had a shorter response time than Audio Radio
- Baseline had a shorter response time than Audio Radio

4.4.5 Comparison of TDRT Response Times and Brake Light Response Times

For both the TDRT and the Brake Light responses, the mean response time per task was calculated for each participant and in each context (on-road or simulator). The correlation between these two response time measures obtained in the on-road context is shown in Figure 32 and simulator data in Figure 33. In Figure 32 and 33, the response times for each task are plotted by participant to illustrate individual differences and patterns. Some clustering of data by participant is evident. Correlations between the two response time measures were not statistically significant for the on-road (r(69) = -.055, p = .54) or simulator (r(66) = .224, p = .071) context.

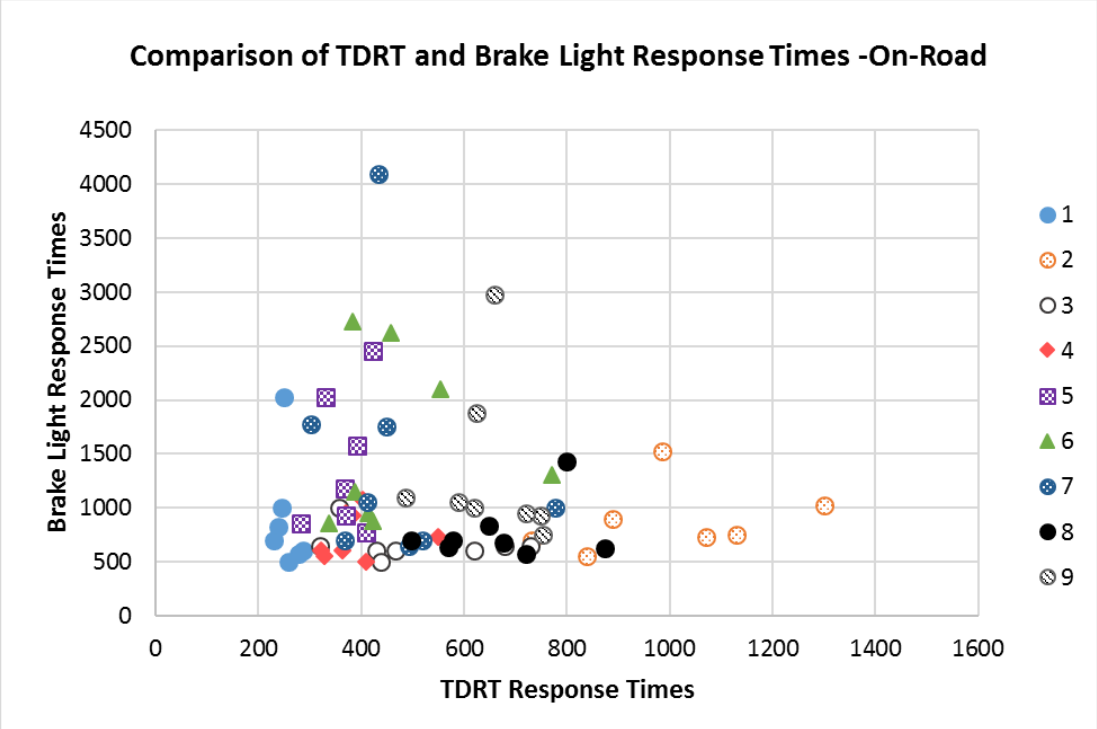


Figure 32. Comparison of RDRT and Brake Light response times (ms) for on-road context

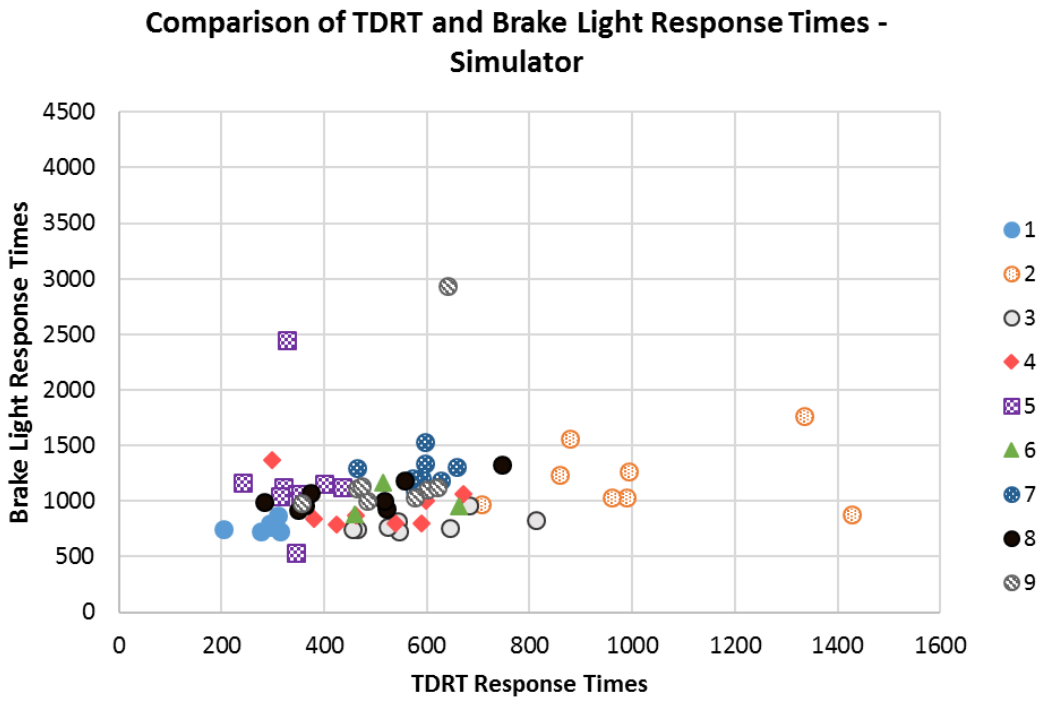


Figure 33. Comparison of TDRT and Brake Light Response times (ms) for simulator context

4.4.6 Eye Glance Measures – All Scenarios

Eye glance coding was performed manually by viewing driver face videos collected on-road and in the driving simulator. Off-road glance durations were summed over the duration of each task attempt. For each context and scenario, the off-road glance sums for the two attempts per task were averaged and then the mean and SEM were calculated across the 9 participants. Figure 34 shows the mean TEORT for each task by scenario and by context. Distributions of off-road eye glance durations are provided in this report under the description of Study 3.

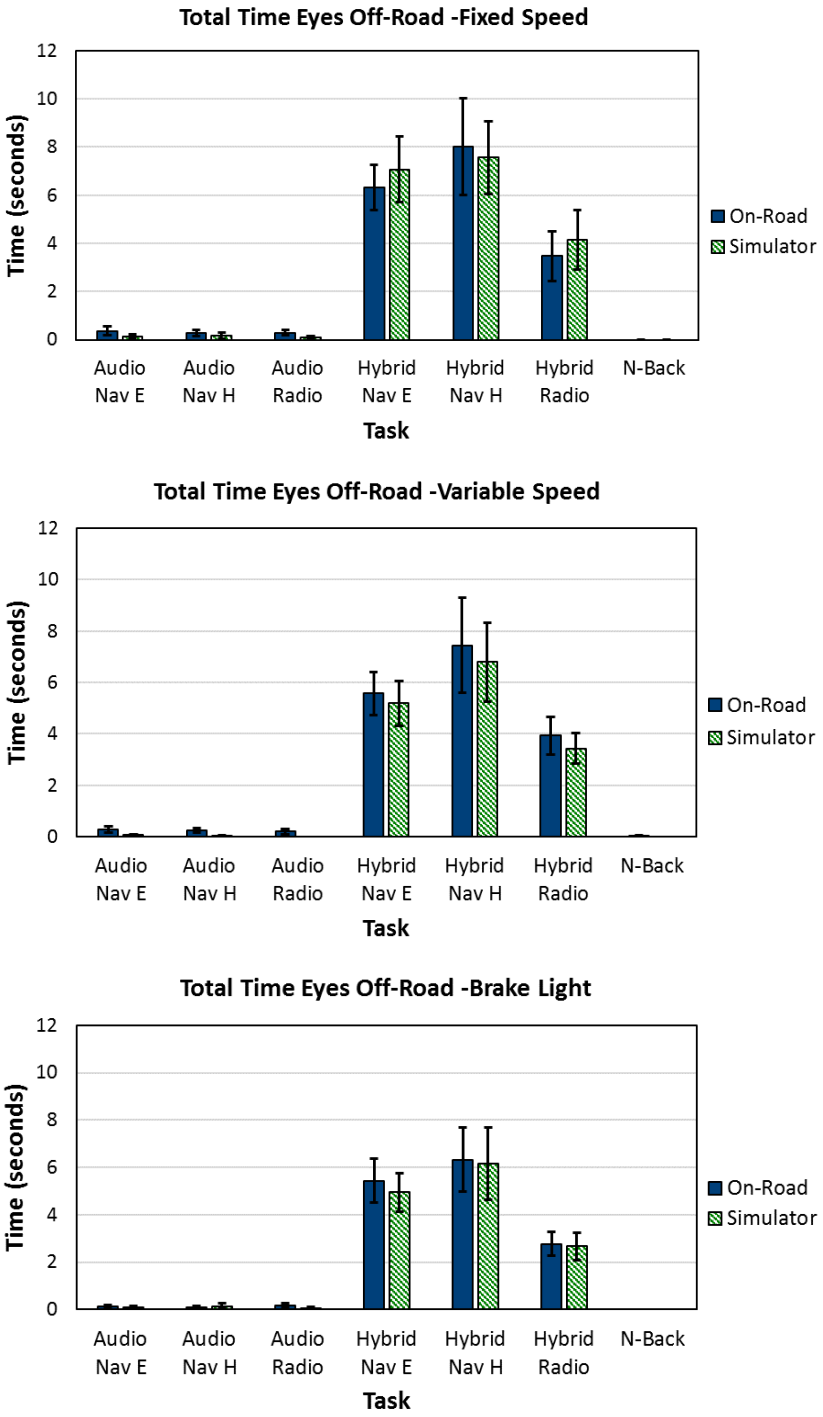


Figure 34. Average Total Eyes Off-Road Time, by task for (from top) Fixed Speed, Variable Speed and Brake Light scenarios

For each scenario, the on-road and simulator TEORT data are similar. In addition, the three scenarios show the same pattern of TEORT by task, with the auditory and N-Back tasks resulting in hardly any glance time. Together, the similarity of the results suggests that the TEORT measure is reliable, even with a small sample of test participants.

While the three scenarios resulted in overall similar patterns of TEORT, the Brake Light scenario resulted in less TEORT for each Hybrid task. This may be because participants were more focused on the roadway during this scenario, given they were anticipating the LV's brake lights going off in front of them.

Further analyses of glance data were carried out to calculate the number of long off-road glances (> 2 sec), and mean duration of glances per task. These results were compared to criteria provided in NHTSA Distraction Guidelines (NHTSA, 2013). According to the NHTSA Distraction Guidelines, an in-vehicle task should be locked out from performance by drivers while driving unless the following three criteria are met:

> **A**: For at least 21 of the 24 test participants, no more than 15 percent (rounded up) of the total number of eye glances away from the forward road scene have durations of greater than 2.0 seconds while performing the testable task one time.
>
> **B**: For least 21 of the 24 test participants, the mean duration of all eye glances away from the forward road scene is less than or equal to 2.0 seconds while performing the testable task one time.
>
> **C:** For at least 21 of the 24 test participants, the sum of the durations of each individual participants' eye glances away from the forward road scene is less than or equal to 12.0 seconds while performing the testable task one time.

The NHTSA Distraction Guidelines criteria are applied to only a single attempt of each task. Therefore, we considered only the first attempt of each task to evaluate conformance to NHTSA Distraction Guidelines (NHTSA, 2013). All our testing conditions deviated from the exact protocol specified by NHTSA for evaluating glances. The NHTSA Distraction Guidelines require the participant to drive on a straight road following a fixed speed LV. In the current study, participants performed testable tasks while driving on relatively straight, real road segments, and on a perfectly straight road in a driving simulator. However, in the fixed speed scenario used in this study participants were simultaneously performing the TDRT protocol, which is not part of the NHTSA protocol. In the other scenarios, participants were responding to a LV brake light signal, or to changes in the speed of the LV. Thus, the current study added task dimensions beyond those specified by protocols in the NHTSA Distraction Guidelines.

Although only nine participants took part in Study 2, we extrapolated the pass/fail percentages such that any task where seven or fewer participants passed a particular NHTSA criterion (NHTSA, 2013), we considered that task to be not in conformance with that aspect of the Distraction Guidelines. Table 18 shows the number of participants who passed each of the three NHTSA glance criteria (A, B, C) for each task, within each of the three scenarios for the on-road context. Table 19 shows the number of participants who passed in the simulator context. Cells shaded yellow had eight participants pass that criterion, whereas cells shaded pink had seven participants or fewer pass the criterion.

For the on-road sessions, the only tasks that did not pass the NHTSA glance criteria were Hybrid tasks (involving both Auditory and Visual displays). This is to be expected because the auditory

tasks did not involve the presentation of any information on the secondary display. While all participants passed Criterion B, the Hybrid Navigation Easy task failed to pass in both the fixed speed and variable speed scenarios. One possible explanation for this could be that the easy version of the task required fewer total glances at the screen. Even just one long glance would cause a participant to fail Criterion A if they looked at the screen four times or fewer.

Table 18. Number of the nine participants passing NHTSA Distraction Guidelines glance criteria for on-road tasks

On-Road Scenarios									
	Fixed Speed			Variable Speed			Brake Light		
	A	B	C	A	B	C	A	B	C
Auditory Nav Easy	9	9	9	9	9	9	9	9	9
Auditory Nav Hard	9	9	9	9	9	9	9	9	9
Auditory Radio	9	9	9	9	9	9	9	9	9
Hybrid Nav Easy	7	9	9	7	9	9	8	9	9
Hybrid Nav Hard	8	9	7	9	9	8	9	9	8
Hybrid Radio	8	9	9	9	9	9	9	9	9
N-Back	9	9	9	9	9	9	9	9	9

Note: To pass criteria "A," less than 15 percent of eye glances from the road have durations < 2.0 seconds. To pass criteria "B," the mean duration of all eye glances from the road ≤ 2.0 seconds. To pass criteria "C," the sum of the durations of eye glances from the road ≤ 12.0 seconds.

Table 19. Number of the nine participants passing NHTSA Distraction Guidelines glance criteria for simulator tasks

Simulator Scenarios									
	Fixed Speed			Variable Speed			Brake Light		
	A	B	C	A	B	C	A	B	C
Auditory Nav Easy	9	9	9	9	9	9	9	9	9
Auditory Nav Hard	9	9	9	9	9	9	9	9	9
Auditory Radio	9	9	9	9	9	9	9	9	9
Hybrid Nav Easy	7	9	8	9	9	9	8	9	9
Hybrid Nav Hard	9	9	8	9	9	8	8	9	8
Hybrid Radio	9	9	9	9	9	9	7	8	9
N-Back	9	9	9	9	9	9	9	9	9

Note: To pass criteria "A," less than 15 percent of eye glances from the road have durations < 2.0 seconds. To pass criteria "B," the mean duration of all eye glances from the road ≤ 2.0 seconds. To pass criteria "C," the sum of the durations of eye glances from the road ≤ 12.0 seconds.

For the simulator scenarios, all tasks passed Criteria B and C. Once again, the only scenarios that failed Criterion A were Hybrid tasks. In this instance, the Hybrid Navigation Easy task only failed Criterion A during the fixed speed scenario and the Hybrid Radio only failed during the brake light scenario. As with the on-road context, one explanation for the failure of Criteria A for the Hybrid Navigation Easy task is that participants glanced over fewer times, but a greater proportion of those glances were classified as "long." The Hybrid Radio task's failure on Criterion A for the brake light context however was unique to the simulator context.

4.5 DISCUSSION

In Study 2, data were collected from nine participants to explore several potential measures for evaluating the demands placed on the driver from performing voice control tasks while driving. Although this is a small sample of drivers, the within-subjects design allowed for meaningful comparisons to be made between data collected in a driving simulator and data collected on-road. However, more definitive conclusions could be made based on the greater statistical power enabled by a larger sample.

The same set of voice tasks was performed in two data collection contexts (on-road and driving simulator) using the same protocols and same set of test participants. We sought to identify measures that are sensitive for discriminating between voice tasks of different difficulty levels and baseline driving and to identify test protocols and measures that provide consistent results

across testing contexts and driving scenarios. A commonly used task (N-Back) was included in the set of voice tasks so that data collected in this study could be compared to other datasets in the future.

Task completion times were similar across contexts and the pattern of completion time differences across tasks was also consistent across the three different driving scenarios, which were performed on different days. These results suggest that task completion time may be a robust measure for comparing differences between voice tasks, particularly regarding risk exposure. The Hybrid tasks were completed much faster than similar tasks performed using Audio Only. This may indicate that audio-only interfaces place longer lasting, and therefore riskier demands on the driver.

The TDRT response time measure distinguished between voice tasks and baseline driving and the results did not depend significantly on context, nor was there any significant interaction between context and task. Each of the voice tasks led to significantly longer TDRT response times than driving without doing a VCS task, indicating the increased cognitive load imposed by the voice tasks. The TDRT response times for the N-back task, which is thought to impose a moderate cognitive load, were significantly greater than TDRT response times for each of the VCS tasks except the Hybrid Radio task. All of these results support the use of TDRT response time as a measurement tool for assessing VCS tasks in terms of the cognitive load that they impose on the driver. A limitation of TDRT is that it may not be appropriate for assessing VCS tasks that have short task completion times (i.e. < 5 seconds).

In Study 2, the TDRT hit rate was close to 100 percent for most tasks performed on the road but was slightly less for some of the Hybrid tasks and the N-Back task performed in the driving simulator. The ceiling effect in the on-road data and differences in data across contexts do not support the use of this measure for evaluating VCS tasks.

An advantage of TDRT over the Brake Light response task implemented in this study is that multiple responses were recorded during a single attempt of a voice control task. The mean of several responses provides a more stable (less variable) estimate of response time per task than only a single response. It may be possible to conduct the Brake Light response task with more frequent stimuli and responses to make it more similar to a RDRT task.

Like TDRT, the Brake Light responses did not depend significantly on context. However, we noted a very low correlation between the Brake light response times per task and the mean TDRT response times. This may indicate that they measure different aspects of the driver's attention, for example, visual and cognitive. The set of differences between tasks that were detected using the Brake Light response task was quite different from the set detected using TDRT. In fact, TDRT response times were significantly greater for the N-Back task as compared to the Audio Radio task, but the opposite result was found using the Brake Response task where response time was significantly less for the N-Back task as compared to the Audio Radio task.

The variable speed scenario used in Study 2 was designed to measure how well the participant could maintain a fixed distance gap between vehicles by responding to speed changes (without braking) initiated by the LV. The lag in speed changes by the participant relative to speed

changes by the LV was assessed across tasks as a potential measure of distraction from VCS tasks. Greater lags were hypothesized for more demanding VCS tasks relative to baseline driving. Cross correlations of the two vehicles' speed were performed to determine the lag which produces the highest correlation in speed. In some cases, substantial differences were observed in the correlation by lag functions between a participant's two attempts on the same task. Also, on some attempts, the maximum correlation that could be achieved by shifting the speed data in time was low. We are not confident that this is a reliable measure. Although some statistically significant differences between tasks were detected, these differences were not consistent across contexts. There was only one significant difference between baseline driving and a voice task. The lag in the driving simulator for data collected in the Hybrid Navigation Hard task was significantly greater than the lag in the Baseline driving task.

The glance data collected in Study 2 was sensitive to differences between voice tasks and reliable across contexts. Although NHTSA glance criteria generally require a larger sample size to get meaningful results, we did see differences between Hybrid tasks and Voice-Only tasks with the limited sample size. The details of passing or failing individual criteria per task varied between contexts. The TEORT measure for the Hybrid tasks showed an interesting trend across the three driving scenarios such that the Fixed Speed scenario produced slightly more off-road glance time than the Variable Speed scenario, and the Variable Speed scenario produced slightly more off-road glace time than the Brake Light scenario. We speculate that the demands of the driving related tasks were lower in the Fixed Speed scenario as compared to the other scenarios, which allowed participants to allocate more of their attention to the in-vehicle screen used for the Hybrid voice tasks. Distributions of off-road eye glance durations for both Study 1 and Study 2 are provided in this report under the description of Study 3.

5 STUDY 3 – MODELING RELATIVE CRASH RISK FOR VCS TASKS USING CONTERFACTUAL SIMULATION

5.1 PURPOSE

Off-road glance measures quantify the visual demand of potentially distracting tasks. Longer and more off-road glances might increase the likelihood of drivers missing critical roadway events, such as a LV braking, a red light, or a pedestrian crossing the road. Missing such events, or being slow to respond to them, could increase crash risk. A fundamental question regarding driver distraction, however, is to quantify the degree such glance behaviors increases the risk of crash and injury. Suppose that we compare the effect of a 1.9-second glance and 2.1-second glance. We know that the latter exceeds the threshold of being "risky," because it is over 2.0 seconds (Klauer et al., 2006; 78 FR 24817, 2013), but the difference between the two is only 0.2 seconds. At 50 mph, a vehicle would travel approximately 14.67 ft in that time, but this distance cannot directly tell us how dangerous it is. Likewise, it is unclear how the risk of 2 glances of 1.5 seconds each compares to a 1 glance of 2.5 seconds. The relationship between the pattern of glances away from the road and crash risk is further complicated by the dynamically changing driving environment.

There are studies that directly associate observed crash risk with glance durations. In SHRP 2, researchers at data collection centers in six States conducted a naturalistic driving study with more than 3,000 volunteer participants and their vehicles over a 3-year period (Blatt et al., 2015). From naturalistic driving data, continuously collected from vehicles in the SHRP 2 driving study, researchers have calculated the odds ratio of crash given a certain behavior such as phone conversation (Victor et al., 2015). It was also found that longer glance duration is likely to be involved in rear-end crashes when it coincides with higher closure rate between LV and following vehicle (Young, 2017). However, this method is retrospective and limited to a small number of observations.

Counterfactual simulation can be used to overcome the limitations. Counterfactual simulation takes a real-world observation, and explores how alternate events (e.g., different glance patterns, driver braking behaviors) might influence outcomes like crash frequency and crash severity. The underlying assumption is that the observed outcome from real-world observation, such as a crash from naturalistic driving data, is only one possible realization of a stochastic process. By applying different behaviors instead of the original driver's behavior, we can produce a better estimate of the full distribution of likely outcomes. In transportation safety research, counterfactual simulation has been used for a variety of purposes to identify how surrogate events (e.g., traffic conflicts) are related to the actual crashes (Davis, Hourdos, Xiong, & Chatterjee, 2011), to assess the performance of ADAS systems (McLaughlin, Hankey, & Dingus, 2008), or to estimate the effect of off-road glances (Bärgman, Lisovskaja, Victor, Flannagan, & Dozza, 2015; Victor et al., 2015).

Here we use counterfactual simulation to estimate the crash risk associated with different in-vehicle systems. We estimate the risk of distraction when the secondary tasks overlap critical roadway events, such as a braking LV. The risk estimation is an important step, as the ultimate

goal of research concerning driver distraction is to assess the safety consequence of distraction and ultimately alleviate the risk through less distracting designs.

5.2 METHODS

Counterfactual simulation takes kinematics from real world observations and simulates alternative behaviors, resulting in counterfactual outcomes that differ from the observed outcome. Assume that we have one observation of a crash event, where the driver looked away from the road when the LV started braking. Counterfactual simulations can estimate what could have happened if the driver was looking forward and therefore was able to start braking earlier than in the actual event, or if the driver looked away longer and starting braking later than in the actual event. Depending on the severity of the event, these different glance patterns might affect crash likelihood and severity. Different glance patterns can be tested with the same observation (crash event) to see how different glances influence the outcome in that safety critical situation.

In the current study, we assessed glance sequences associated with potentially distracting activities with rear-end scenarios from the SHRP 2 naturalistic driving data. The data include 34 crash events and 211 near-crash events, identical to the data used by Bärgman, Boda, & Dozza (2017). Each event includes a 20-second of time-series record: 15 seconds before collision (crashes) or the time of minimum time to collision (near-crashes), and 5 seconds after this point. The data was available at 10Hz, but up-sampled to 50Hz using linear interpolation, for higher resolution of simulation. Any braking maneuver of the original following vehicle drivers was removed, and the speed profile of the following vehicle was determined by a driver model in the counterfactual simulations (see Bärgman, Boda, & Dozza, 2017).

The anchor point of the simulation, the point where the driver model in the following vehicle replaces the original kinematics is $\dot{\theta} > 0.0065$. Theta (θ) is the visual angle of the rear end of the LV, and $\dot{\theta}$ is the derivative of theta with respect to time. This represents the rate of optical expansion of the rear end of the LV. $\dot{\theta}$ also guided the behavior of the driver model. When $\dot{\theta} > 0.003$ the simulated driver would not look away from the road. The optical expansion rate ($\dot{\theta}$) has widely been used in driving studies. Studies have reported that drivers perceive the relative velocity of a LV when $\dot{\theta}$ exceeds 0.003 (Mortimer, 1990; Mortimer, Hoffmann, & Kiefer, 2014), or 0.0021 or 0.0038 depending on the following distance (Lambel, Laakso, & Summala, 1999). When $\dot{\theta}$ exceeds such thresholds, the drivers perceive that the distance to the LV is closing. From this information, we assumed that if $\dot{\theta}$ exceeds 0.003, the driver does not look away to engage in a secondary task. In summary, the optical expansion rate $\dot{\theta}$ increases as situation evolves, and when it exceeds 0.003, the drivers do not look away from the forward vehicle to assess the situation, and when $\dot{\theta}$ exceeds 0.0065, the drivers act (e.g., brake) to avoid a collision.

Studies have also reported that $\dot{\theta}$ guides drivers' perception of and response to threats. Figure 35 shows the braking response time of drivers, reported in different studies. Maddox and Kiefer (2012) reconstructed response times from crashes, and Muttart, Messerschmidt, and Gillen (2005) conducted meta-analysis of response times. Muttart et al. reported that $\dot{\theta} > 0.0065$ is when the drivers' response time drops and plateaus at 1.6 seconds. We assumed that drivers initiate braking at a threshold crossing $\dot{\theta} > 0.0065$, as indicated with dashed grey line in Figure

35, and the braking response time differs by situation. The braking of the following vehicle follows equation 1 below. As $\dot{\theta}$ increases, response time decreases and converges to the asymptote 0.2s. This equation and its parameters represent a best fit to the data from the cited studies.

$$\mathrm{RT} = \frac{1}{80 * \dot{\theta}} + 0.2 \qquad \text{Eq. 1}$$

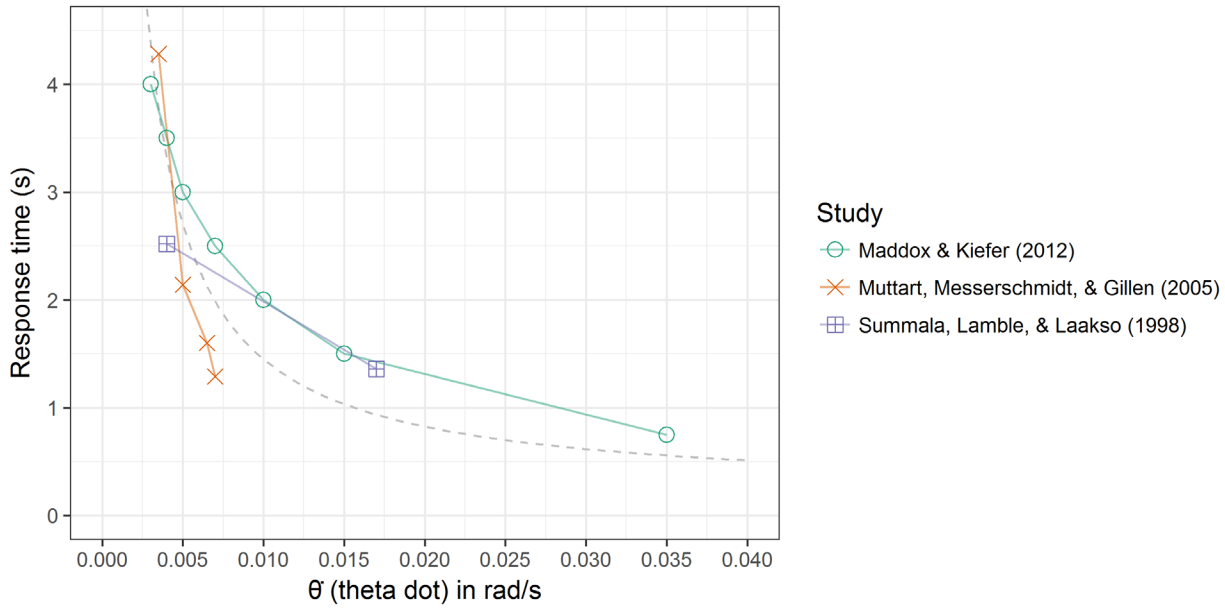


Figure 35. The relationship between $\dot{\boldsymbol{\theta}}$ and response time reported in previous studies. Grey dashed line shows the response time used for the current counterfactual simulation.

According to equation 1, if a driver is attentive when crossing the threshold, the response time is 2.12 seconds after the point of crossing the threshold ($\dot{\theta}$ = 0.0065). If a driver is distracted and the situation is more severe when the driver looks back at the LV, for example, $\dot{\theta}$ = 0.02, the response time becomes 0.825 seconds. However, in particularly severe situations, a high $\dot{\theta}$ can trigger an even faster response. As the situation evolves ($\dot{\theta}$ increases), response time decreases following equation 1. Because response time decreases quickly as $\dot{\theta}$ increases we update the response time as the situation evolves and select the earliest occurring response.

Figure 36 demonstrates this situation. It shows time along the horizontal axis and $\dot{\theta}$ on the vertical axis. The left curve shows how $\dot{\theta}$ increases as the following vehicle approaches the LV. The right curve shows the initial braking response, and the distance between the two curves is the reaction time. The reaction time, which decreases with increasing severity, defines this curve. The dotted line shows the threshold of response. The response time of a driver who is looking at the road at the time of the time when $\dot{\theta}$ exceeds the threshold of 0.0065 is 2.12s, which is indicated as the "initial response" in the graph. If the situation evolves according to the curve on the left, it will force a faster response as shown by the earlier "actual response".

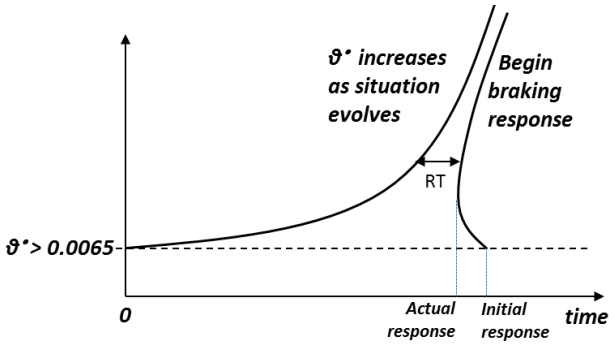


Figure 36. A timeline showing the influence of the evolving situation (curve on the left) on when the driver initiates a braking response (curve on the right).

Figure 37 shows the change of $\dot{\theta}$ in SHRP 2 vehicle data over time. As a following vehicle approached a LV, $\dot{\theta}$ increased. Drivers responded at various $\dot{\theta}$ (black points), but note that the drivers in the data might have been looking away while the situation was evolving or used other visual cues to predict the situation. In the data, after crossing the first threshold ($\dot{\theta} > 0.003$; about 3.1 seconds before critical event), it took about 0.4 seconds on average to cross the second threshold ($\dot{\theta} > 0.0065$; about 2.7 seconds before critical event).

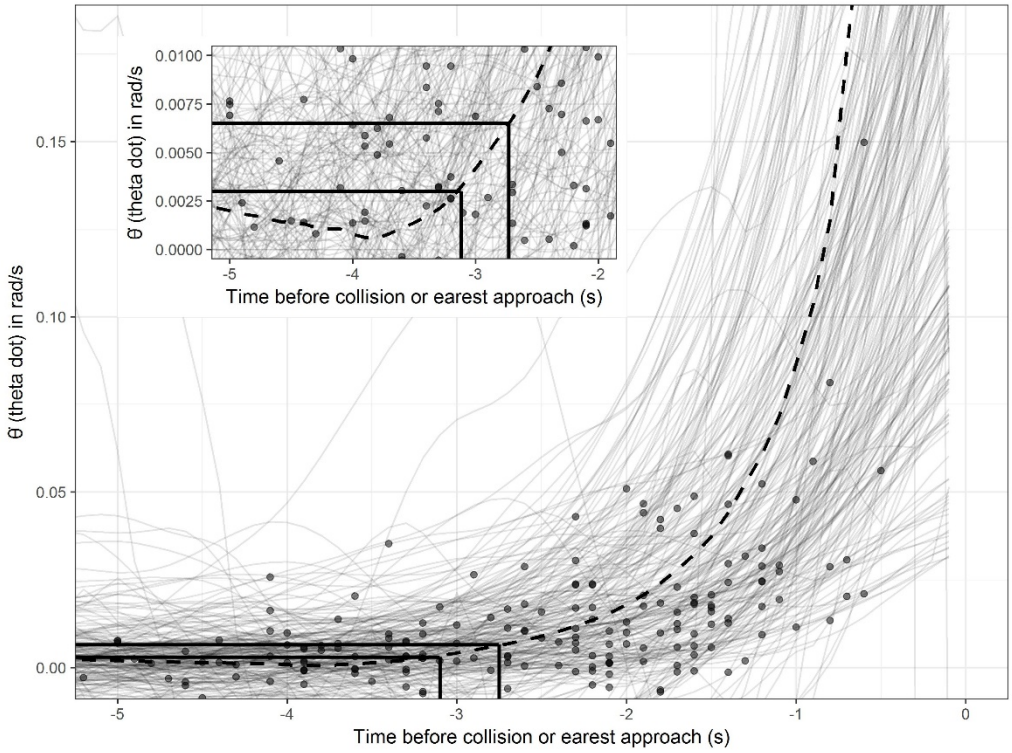


Figure 37. Change of $\dot{\boldsymbol{\theta}}$ in SHRP 2 data over time, from five seconds before the collision (crash events) or the nearest approach (near-crash events). Black points indicate the time when the drivers in the events initiated a response, and dashed line indicates the mean $\dot{\boldsymbol{\theta}}$ at each time point. Two horizontal lines show the thresholds (0.003 and 0.0065) and the two vertical lines show the corresponding times. They were about 0.4s apart.

For the model, we assume that a threat is perceived only when the driver looks at the forward roadway. If a threat emerges while the driver is looking away from the road, we modeled the driver as responding to cues only after returning visual attention back to the road. However, if the driver perceives that the LV is braking, the driver does not look away. Once the threat is perceived, the driver begins braking with a 1.23 g/s jerk until a maximum braking of 0.68 g is achieved, which are the mean values reported in Markkula et al. (2016).

Figure 38 shows the behavior of the counterfactual simulation with different off-road glances. If the off-road glance that starts before the $\dot{\theta} > 0.003$ threshold ends before crossing the $\dot{\theta} > 0.0065$ threshold, the driver attends to the roadway to observe the evolving situation and does not look away again. In this case, the braking response begins 2.12 seconds (i.e., reaction time) after the $\dot{\theta} > 0.0065$ threshold crossing, as specified in equation 1. If the off-road glance ends after the $\dot{\theta} > 0.0065$ threshold, the glance "overshoots" the threshold. Overshoot is the time between the threshold crossing and the driver looking back to the road. The overshoot delays the driver's perception of the situation. This delay can lead to the situation being more severe when the driver looks back to the road. The driver can partially compensate for the overshoot delay because the response time gets shorter, following the curve in Figure 35 and equation 1.

The response time may change as cognitive load is added to driving. A meta-analysis of Caird, Willness, Steel, and Scialfa (2008) and Horrey and Wickens (2006) showed that cell phone conversation while driving increases reaction time, about 0.25 seconds. Response time can increase with some roadway events even when the task does not require that the driver look away from the road, suggesting that cognitive load associated with secondary tasks might undermine driving safety.

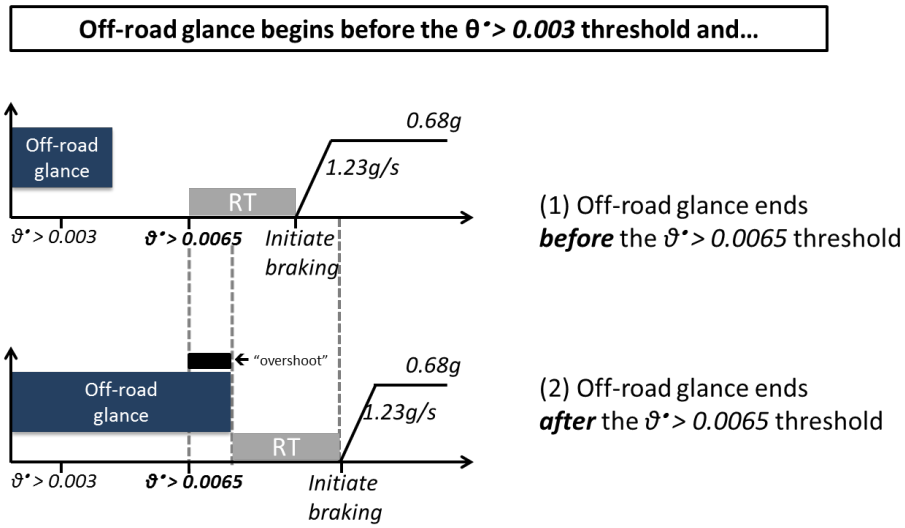


Figure 38. If an off-road glance ends after the $\dot{\theta} > 0.0065$ threshold then the drivers' response to the looming cue gets delayed.

5.2.1 Using counterfactual simulation to interpret data from the visual-manual evaluation

To investigate the effects of different glance patterns observed in Study 1 and Study 2, we ran counterfactual simulations for the glance patterns we observed with participating drivers performing various potentially distracting voice-based tasks. Participants completed these tasks, which required different time durations (total task time) and produced a range of glance patterns. There are several options in simulating these tasks in the counterfactual simulation. We could simulate critical events only within a total task time, but it would bias drivers with short total task time to be riskier than drivers with long total task time. Consider two drivers who each complete a task with 10 seconds TEORT, but one driver takes 20 seconds and the other takes 60 seconds to complete the task. A critical event that is simulated within the 20-second task will be more likely to overlap with an off-road glance than a critical event simulated within the 60-second task. The former driver, however, might be attentive to roadway after completing the task. To overcome this dilemma, we assumed that the drivers are attentive to the roadway after completing each task, filling up the 80-second time window. The 80 seconds window is selected because the maximum task time was 75 seconds.

Within the 80-second time window of each glance pattern, a LV braking event was simulated. In the simulation, initiation of the event is distributed uniformly between 0 to 80 seconds, and once the event is created within this time window, the kinematics trajectory and the driver response plays out until the end, even if this exceeds the 80-second window. The kinematics of the original event is used until the perception of the threat – $\dot{\theta} > 0.0065$. During the response time, the speed of the following vehicle is fixed at the speed at the moment of threshold crossing. After the response time passes, simulated braking is applied. The LV kinematics remains as that vehicle's original state throughout the simulation.

The result of the simulation was reported in terms of crash risk and crash severity. Crash risk is the proportion of crashes out of all counterfactual simulation runs. Crash severity is reported in two measures: delta V and MAIS2+. Delta V is the speed difference before (V_0) and after the crash (V_1 and V_2), as in the equation 2 below. We assumed that the two vehicles involved in the crash are the same mass ($m_1 = m_2$) and had perfectly inelastic collision ($V_1 = V_2$).

$$\Delta V = V_0 - \frac{m_1 * V_1 + m_2 * V_2}{m_1 + m_2} \quad \text{Eq. 2}$$

MAIS2+ is the probability of experiencing level 2 or higher injury, when all injury is categorized from 0 (no injury) to 6 (fatal) as in AIS scale (Gennarelli & Wodzin, 2006). MAIS 2+ is calculated based on NASS CDS data from 2001 to 2014: logistic regression between delta V and injury observed in crashes that a passenger car rear-ended a LV. Equation 3 below shows the logistic relationship between the delta V and injury risk MAIS2+.

$$MAIS2+ = \frac{e^{-3.28+0.064*\Delta V}}{1 + e^{-3.28+0.064*\Delta V}} \quad \text{Eq. 3}$$

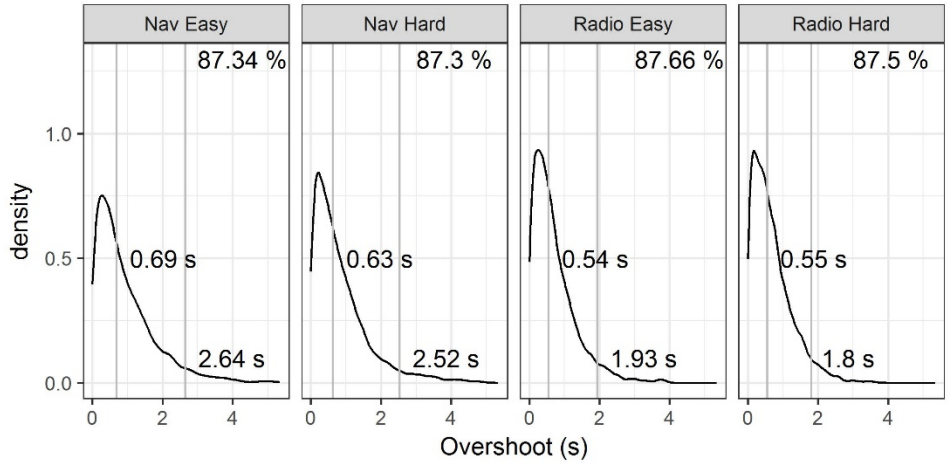
5.3 RESULTS

We used 34 crash events and 190 near-crash events from the SHRP 2 naturalistic data in the counterfactual simulation. We used the empirical data from two studies that each included nine drivers. In Study 1, conducted at the University of Washington, nine drivers experienced simulated driving while performing DRT in two different modes, four voice task types (, i.e., Radio Hard, Navigation Hard, Radio Easy, Navigation Easy), and had three trials of each combination. In Study 2, nine drivers at Westat experienced both driving in the simulator and on the road. The results are identified as University of Washington and Westat and graphs for each are presented with the University of Washington first.

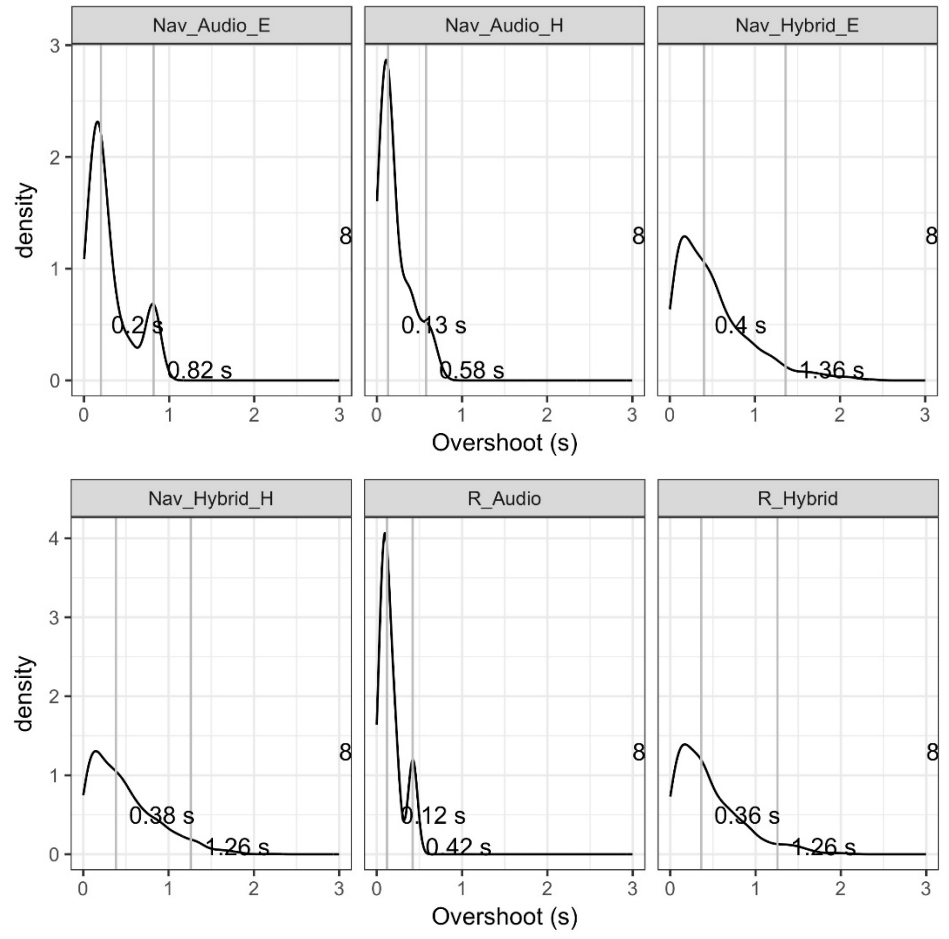
Ten counterfactual simulations were completed for each driver in each trial. Because the simulation was repeated 10 times, there were 483,840 simulation records in total. The number of simulated crashes was 60,724, and the number of simulated near-crashes was 423,116. Thus, the proportion of crashes was 12.55 percent, which is similar, but slightly lower than the crash rate observed in original SHRP 2 data (34 out of 224, 15.1%).

5.3.1 Overshoot and response time

Glance patterns in each of four task types generated different overshoot duration.
Figure 39 shows the distribution of overshoot, excluding zero overshoot. The percentage of zero overshoot (where driver was attentive at the $\dot{\theta} > 0.0065$ threshold crossing) is marked on the corner of each grid. Because the simulated drivers were attentive to the roadway most of the time (i.e., completed task in 10 seconds and be attentive to roadway during the 80-second window), the rate of zero overshoot was high, around 87 percent. The average duration of overshoot was higher in Navigation Easy task (932ms), followed by Navigation Hard task (854ms), Radio Easy task (705ms), and Radio Hard task (688ms). The graph shows the 50th and 95th percentile values. A similar graph for the Westat data shows a narrower distribution with fewer long off-road glances. The glance distributions for the Westat on-road and simulator data are very similar, and so the data were combined so that all graphs show the combined simulator and on-road data for Westat.



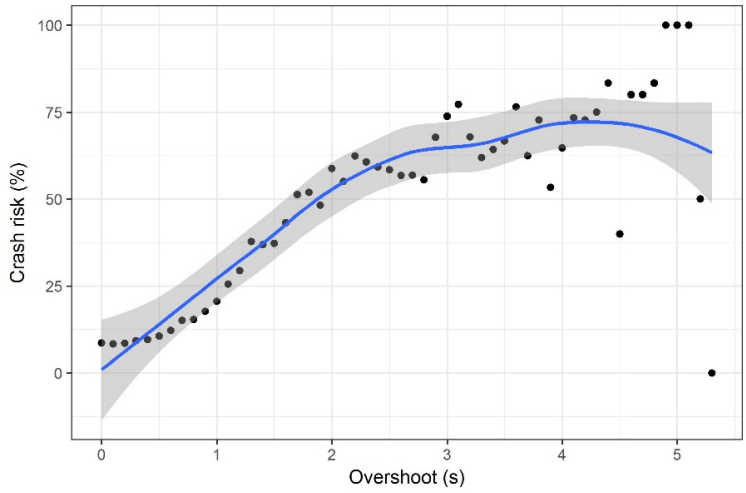
University of Washington: Simulator (Note: Scaling differences)



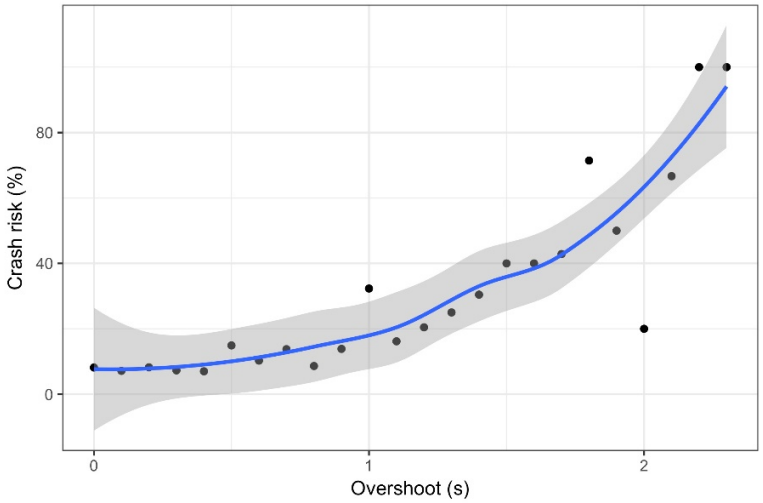
Westat: On-road and Simulator (Note: Scaling differences)

Figure 39. Overshoot of four different tasks. Overshoot is aggregated over two DRT modes. The 50th and 95th percentile positive overshoot of each task type is marked with vertical grey line.

Crash risk increases as overshoot increases (Figure 40). Figure 40 is the data of all drivers and all tasks combined. The overshoot times were combined into 100ms intervals and for each interval we calculated the proportion of crashes. For instance, the proportion of crash in simulation of overshoot between [0ms, 100ms] is around 10 percent (the left most point). Long overshoot means that the driver model misses the initial evolution of the situation because gaze is not directed at the road and responds to the situation when the crash is more imminent. If an overshoot happened, average response time after returning attention back to road was 1.40 s. The overshoot is important because it reflects how the overall glance pattern (i.e., timing, duration, and number) combine to increase driver response to braking LVs. The line in Figure 40 shows the smoothed fit to the data and the grey ribbon indicates the 95th percentile confidence interval. The notable difference between the two datasets is that the Westat data has fewer long glances. Although the shape of the functions seems quite different, considering only the zero to 2-second range in the University of Washington data, the function is quite similar.



University of Washington: Simulator (Note: Scaling differences)



Westat: On-road and Simulator (Note: Scaling differences)

Figure 40. Relationship between overshoot and crash risk. Each point represents a single crash event.

5.3.2 Crash risk by scenario – unavoidable events

The SHRP 2 data set includes safety critical events (34 crashes and 190 near-crashes) with initial speed of the following vehicle that ranges from 0 km/h to 45.6 km/h and to 33.6km/h for the LV, and initial distances that range from 1 meter to 64.0 meters. This initial condition is the state of the vehicles at the precipitating event, such as when the LV began changing lanes or decelerating. There was no significant difference in the initial speed of the following vehicle or LV for crash and near-crash events ($F(1, 221) = 2.71$, $p = .10$; $F(1, 220) = 1.29$, $p = 0.26$) and in the initial distance ($F(1,220) = 1.41$, $p = .16$). This similarity of initial conditions and behavior of LV in the two event types suggest that the drivers' responses to the situation played a critical role in creating more or less severe situations. In this analysis, we do not separate crash and near-crash events.

SHRP 2 events showed different crash risks. As Figure 41 shows, there were "unavoidable" events in our counterfactual simulation (7 in crash events and 20 in near-crash events). In these events, the LV decelerated quickly while the following vehicle was at a short distance, and thus the collision was unavoidable with the currently modeled braking maneuver (1.23g/s jerk and 0.68g maximum braking) even for a driver model with no off-road glances. With zero overshoot time, all the simulated crashes produced by each of these events generated the same injury severity because the braking maneuver was fixed.

Figure 42 shows the distribution of delta-V at crash. The graph on the left includes all crashes in simulation, and the graph on the right excludes unavoidable events. The crashes associated with each unavoidable event with zero overshoot resulted in the same collision speed, indicated by the spikes in the graph on the left. These figures show the effect of the broader glance distribution in the University of Washington data. The general shape of the distributions does not differ, but the number of crashes is much greater with the University of Washington data.

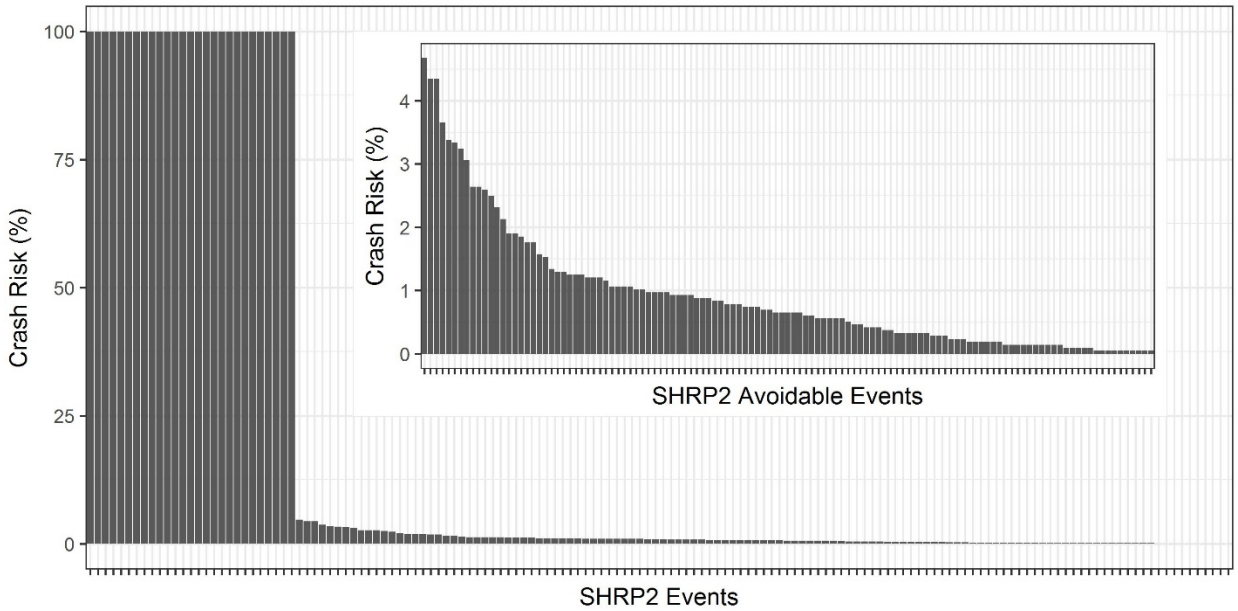
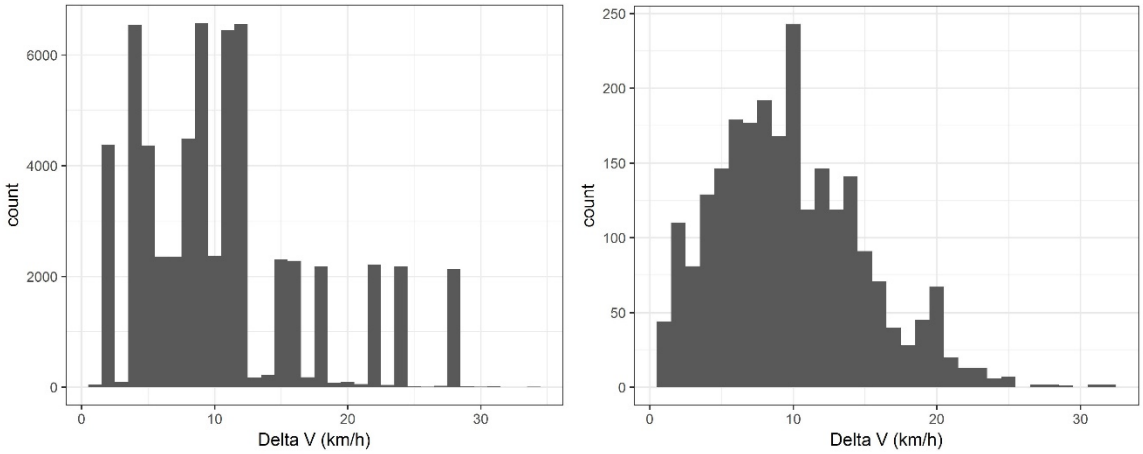
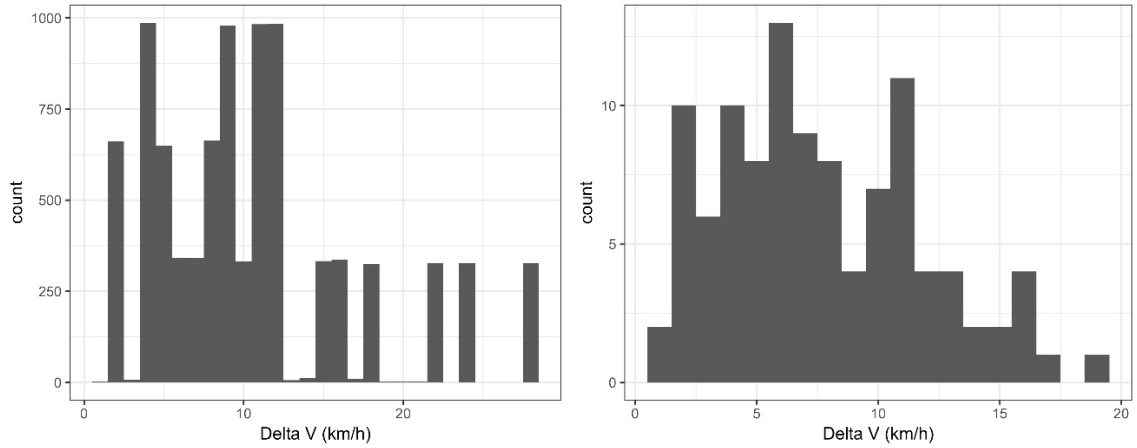


Figure 41. Crash risk by events. Twenty-seven out of 224 were "unavoidable" with currently modeled maneuver. Within the graph, crash risks of "avoidable" events are presented separately.



University of Washington: Simulator (Note: Scaling differences)

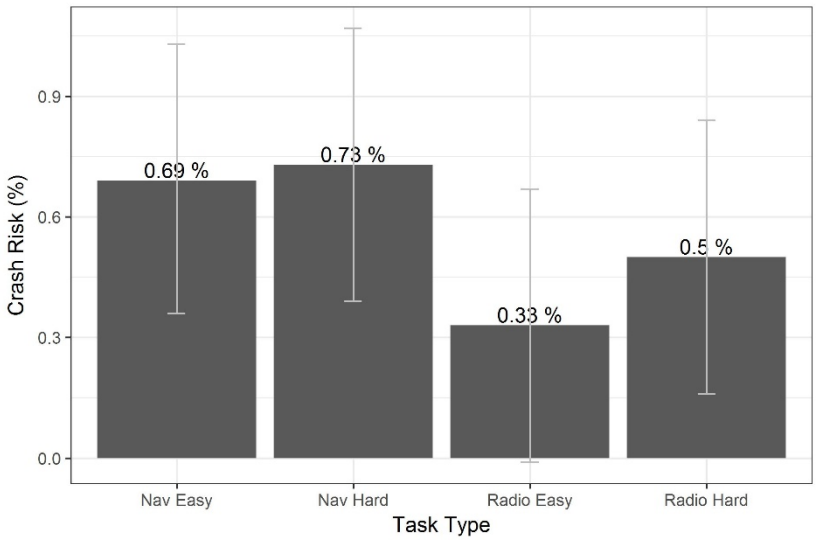


Westat: On-road and Simulator (Note: Scaling differences)

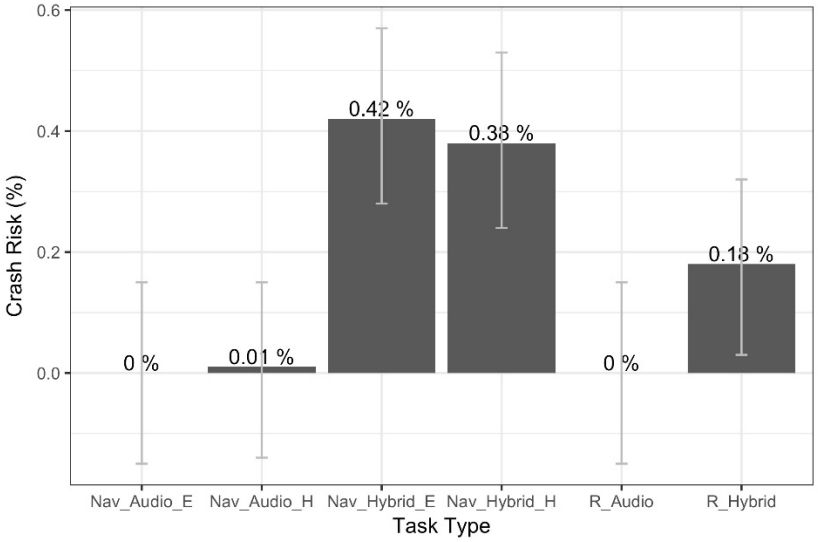
Figure 42. Delta V of crashed simulation outcomes. Unavoidable events are included (left) and excluded (right). The spikes in the left are due to the unavoidable events. Because we apply a fixed braking maneuver, unavoidable events crash at same delta V, when there is no overshoot.

5.3.3 Crash risk by experimental condition

Because the unavoidable events result in crashes in the simulation due to the kinematics of the situation, we excluded them from the analysis to focus on the effect of tasks and driver behaviors. There was significant effect of task type ($F(3, 4.2*10^5) = 66.1$, $p < .001$) when we analyzed crash (dichotomous; crash versus near-crash) with driver and DRT type as random effects and task type as a fixed effect. Figure 43 shows the crash risk by task types with 95th percentile confidence intervals. Here, the broader distributions observed in the University of Washington data led to a higher risk of crash overall. Westat data showed greater differences between tasks and found that Audio tasks had essentially no crashes.



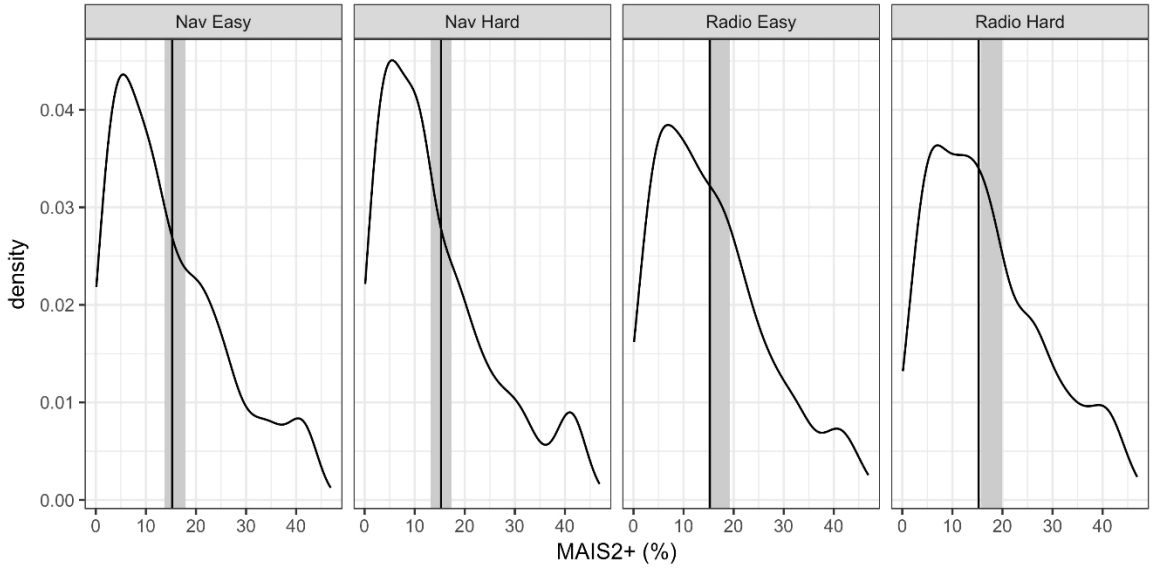
University of Washington: Simulator (Note: Scaling differences)



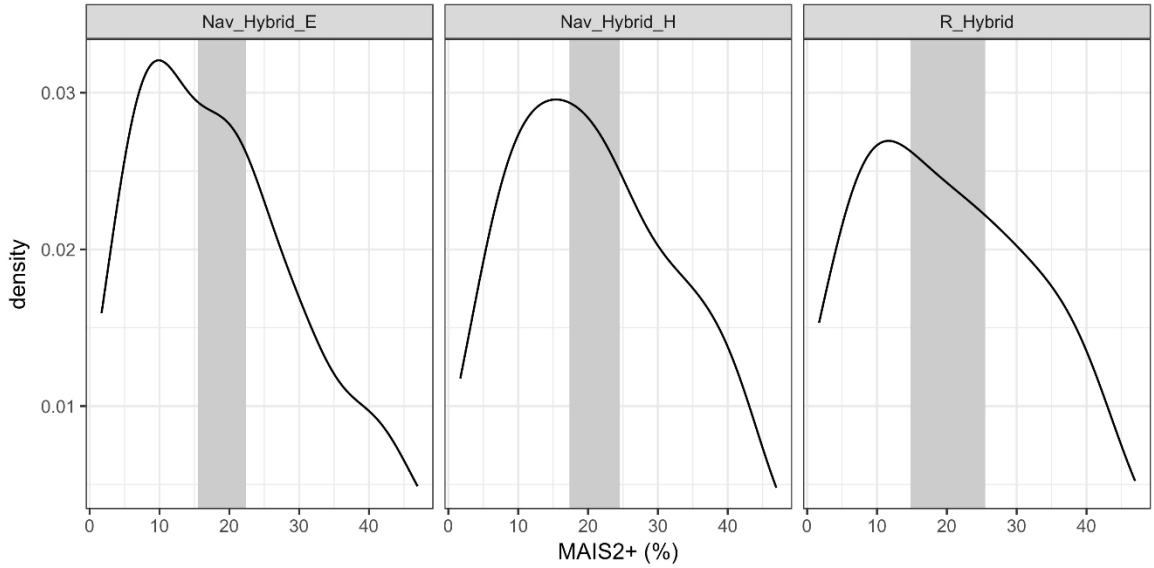
Westat: On-road and Simulator (Note: Scaling differences)

Figure 43. Crash risk of different task types. Vertical lines represent 95th percentile confidence intervals.

Task type also influenced crash severity ($F(3, 2397) = 9.85$, $p < .001$). Figure 44 shows delta-V (top) and MAIS2+ percentage (bottom) at crash. Grey areas show 95 percent confidence interval around the mean value.



University of Washington: Simulator (Note: Scaling differences)

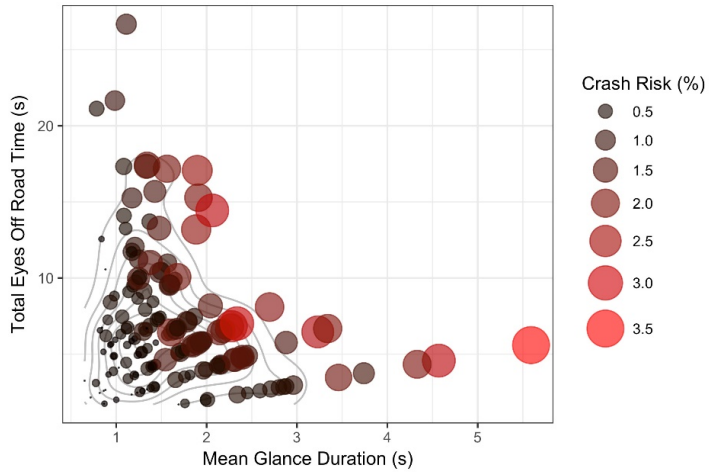


Westat: On-road and Simulator (Note: Scaling differences)

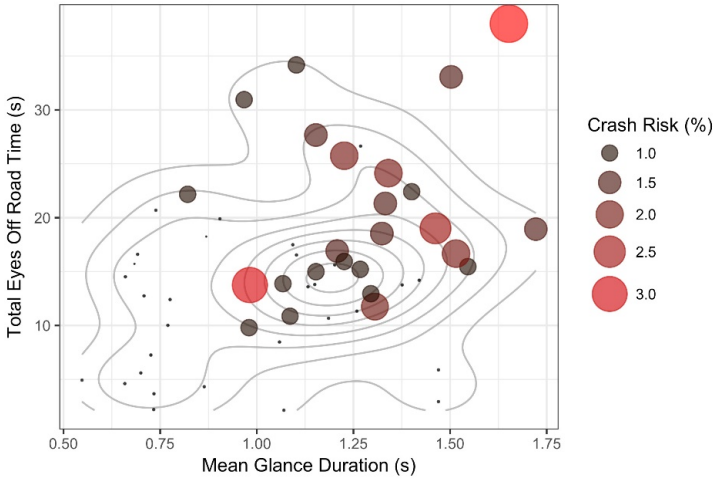
Figure 44. Crash severity (MAIS2+) by task types, with 95 percent confidence interval around the mean.

5.3.4 Crash risk by driver and relationship to glance measures

We used three glance measures to predict crash risk of each driver. To focus on the effect of behavior, unavoidable events were removed. Among three glance measures, (1) MGD, (2) TEORT, and (3) percentage of long (>2s) glances, the first two were the most predictive ($F(1, 166) = 114$, $p<.001$, $p\_eta2 = .41$; $F(1, 183) = 123$, $p <.001$, $p\_eta2 = .43$; $F(1,183) = 1.2$, $p = 0.28$, $p\_eta2 = .00$). The scatter plot in Figure 45 shows how the variables influence crash risk. The vertical axis represents TEORT and the horizontal axis represents MGD. The color and size of the point represents the crash risk predictions of the model. Combinations of high TEORT and MGD, the upper right of the distribution, produce the highest risk. One trend to note is that the cases with the longest MGD have relatively short TEORT. These data suggest that people tend to trade off TEORT and MGD, at least with the tasks in this study. The Westat data show a slightly different pattern, perhaps because the glance patterns in that dataset resulted in fewer collisions and so fewer data points in Figure 45.



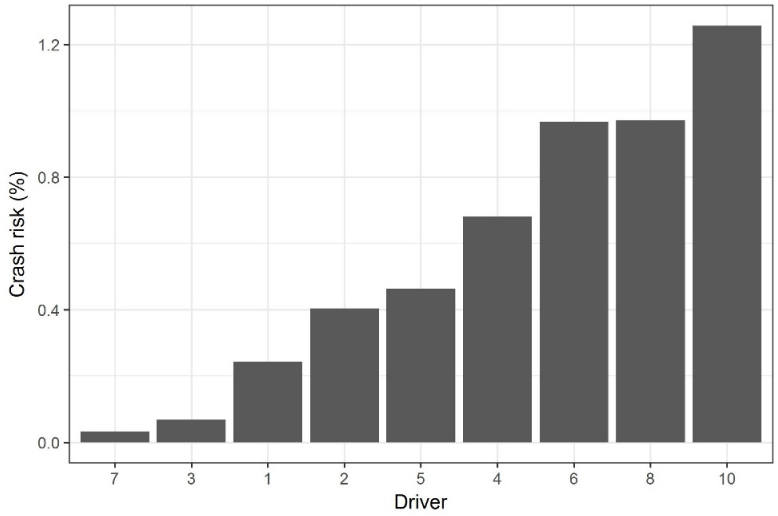
University of Washington: Simulator (Note: Scaling differences)



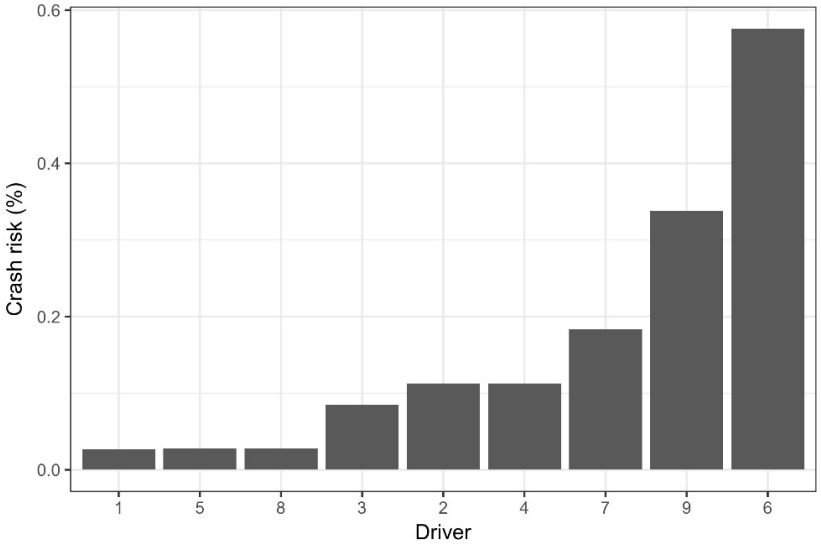
Westat: On-road and Simulator (Note: Scaling differences)

Figure 45. Observed MGD (x-axis) and total eye on road (y-axis) and crash risk (size and color of the points)

Just as crash risk can be calculated for each task, it can also be calculated for each driver. Rather than combining glance patterns and associated by task type we summmarized crashes by driver. Figure 46 shows the crash risk for each driver. Across the bottom of the figure is the driver identification number and the vertical axis shows the crash risk. Drivers differed by an order of magnitude in their crash risk. What this means is that drivers consistently adopt glance patterns that increase their risk. These substantial differences highlight how strongly the behavior of a few drivers could influence the outcome of a study. In this study the crash risk of drivers had a much larger range than that of the tasks—drivers varied by a factor of 10 and tasks varied by a factor of 2. The drivers in the Westat study were generally less risky, but showed a similar pattern of a few drivers having a much higher risk than the others.



University of Washington: Simulator (NNote: Scaling differences, and Driver Numbers refer to a Participant ID, where there was no #9]



Westat: On-road and Simulator (Note: Scaling differences)

Figure 46. Crash risk associated with each of the drivers.

5.4 DISCUSSION

Linking crash likelihood and crash severity to the visual demands of potentially distracting tasks represents an important consideration in guiding design to minimize distraction. The counterfactual analysis described here represents an initial attempt at deriving such risk estimates. Such an analysis endeavors to directly link glance patterns associated with particular devices to crash risk rather than using summary glance metrics, such as MGD or total eyes off road, to evaluate devices.

Glance data obtained in Study 1 and Study 2 were incorporated into the simulations of the SHRP 2 events and the outcome of these simulations was used to estimate crash risk for the various VCS tasks tested in this project. To investigate the effects of different glance patterns observed in Study 1 and Study 2, we ran counterfactual simulations for the glance patterns we observed with participating drivers performing various potentially distracting voice-based tasks. The results show that counterfactual simulation shows promise in linking glance patterns collected during simulator-based evaluation to potential risks of the device for drivers.

Despite the benefits of counterfactual simulation, this type of analysis also has several important caveats. Most importantly, the crash rates that result from the method depend on the sample of potential crash situations. The more severe these situations, the higher the rate of crashes. Consequently, the crash rates should be considered as indicators of relative rather than absolute crash rates. The SHRP 2 data used in this study are predominantly low-speed events and so might underestimate the risk that a larger sample that includes a more complete range of initial speeds.

Another important caveat concerns the driver model used in the simulation. The driver model makes assumptions about how an actual driver would respond to the situation given the visual demands associated with the secondary task. The simple driver model used in this analysis gives very little consideration to individual differences or to how drivers might adapt their driving when confronted by a demanding secondary task. Regarding individual differences, the distribution of crash risk showed that drivers vary substantially in their glance patterns and associated crash risk. Beyond this variability in glance patterns, the variability between drivers could be much higher if we model variability in driver braking responses. Modeling such individual differences could enhance the precision of the risk measures by making it possible to sample behavior with the model that is not feasible to sample directly from human subjects data collection because so many participants would be required. Regarding driver adaptation, there is some evidence that people reduce their speed when engaged in highly demanding tasks. Long periods where the driver's eyes are removed from the road might also provoke a state of hypervigilance and faster response time if the eyes return to the road as a LV brakes. If drivers adapt in such as way, then the simple driver model used in this study likely overestimates risk. More generally, drivers might adapt when they engage in demanding tasks based on the roadway situation, and similarly they might adjust when and for how long they look away based on the roadway situation.

Beyond driver adaptation, other elements of the driver model might not represent actual drivers well and so might lead to inaccurate risk estimation. In this model, to determine when the drivers

perceive threat we used optical expansion rate ($\dot{\theta}$) or inverse of tau (τ^{-1}). Tau (τ) calculated by dividing the optical angle of the rear width of the LV with the expansion rate ($\theta/\dot{\theta}$). It is an optically defined TTC (time to collision) to reflect characteristics of human perception. Inverse of tau (τ^{-1}) has been used, because the larger value of τ^{-1} represents more severe situation (i.e., closer to crash). People may rely on different sources of optical information (θ, $\dot{\theta}$, τ^{-1}) or use combination of information to judge whether an approaching object is far or near.

6 CONCLUSIONS

The objective was to explore potential empirical measures and use a modeling approach to evaluate risk with voice-based systems. Research efforts examined these issues across three studies, which are summarized as follows.

One of the goals of Study 1 involved examining the sensitivity of DRT tasks. This analysis did not find statistically significant differences for driving performance (i.e., SDLP or, SD speed), cognitive workload (i.e., response time or miss rate), or eye glance behavior (MGD), but TEORT was greater for TDRT than for RDRT on all voice tasks except the most complex task (i.e., Navigation Hard). This finding suggests that the modified RDRT task may encourage participants to focus visual attention in the forward direction.

One of the goals of Study 2 compared VCS evaluation measures across the simulator to the same measures collected on-road in a real vehicle. Results indicated that task completion time, TDRT performance, and eye glance measures distinguished between VCS tasks and provided similar results in the on-road and simulated driving contexts.

The glance data collected in Study 1 and Study 2 were sensitive to differences between voice tasks and were reliable across contexts. The details of passing or failing individual criteria per task varied between contexts. The Total Eyes-Off-Road Time (TEORT) measure for the Hybrid tasks showed an interesting trend across the three driving scenarios such that the Fixed Speed scenario produced slightly more off-road glance time than the Variable Speed scenario, and the Variable Speed scenario produced slightly more off-road glace time than the Brake Light scenario. We speculate that the demands of the driving-related tasks were lower in the Fixed Speed scenario as compared to the other scenarios, which allowed participants to allocate more of their attention to the in-vehicle screen used for the Hybrid Voice tasks. In Study 1, TEORT was also higher in the TDRT protocol as compared to the RDRT protocol, perhaps for a similar reason. The RDRT protocol may have encouraged more visual attention toward the forward roadway.

Study 3 used modeling and simulation techniques to model crash and risk severity between VCS tasks from driving simulator and on-road data. Findings suggest that potential applications for linking glance patterns from on-road or simulator-based evaluation to potential risks of the device for drivers. It is important to note that the crash rates that result from the method depend on the sample of potential crash situations, where more severe situations lead to higher rates of crashes in the simulation. This implies that crash rates were indicators of relative, as opposed to absolute measures of, crash rates. Furthermore, SHRP 2 data in this study predominately included low-speed events, and as risk of this type of event is low, data from this analysis might underestimate risk.

Overall, this project made contributions to the development of testing protocols for VCS in several key areas including:

- In Study 1, researchers developed the concept and initial data for a new, modified form of RDRT stimulus, which may provide a measure of visual attention toward the forward roadway. RDRT data were compared to TDRT data collected on the same set of VCS tasks and same set of participants.

- For Study 1 and Study 2, researchers developed the concept and pilot data for a voice-based Radio Tuning task that could be used as a task to be included in evaluation protocols for VCS. This task was compared to a just driving condition, to a 1-back task, and to other VCS tasks using TDRT, RDRT, glance data, and driving performance data (SDLP, SD speed).

- In Study 2, researchers collected appropriate data and made direct comparisons between on-road and driving simulator contexts for several potential measures related to driver workload and distraction including task completion time, TDRT response time and hit rate, glance data, speed correlation with a LV, and response time to onset of LV brake lights. The data were all collected on the same set of participants.

- In Study 3, researchers developed quantitative methods to link crash likelihood and crash severity to the visual demands of potentially distracting tasks such as VCS tasks.

- Researchers developing a driver model of braking responses to looming cues from a forward hazard in Study 3.

- Exercising a counterfactual simulation in Study 3, researchers combined existing data from crash and near crash scenarios recorded in the SHRP 2 naturalistic driving study with glance data patterns collected within Study 1 and Study 2 in the current project.

REFERENCES

78 FR 24817, 24817-24890, Docket No. NHTSA-2010-0053, Document Number 2013-09883. National Highway Traffic Safety Administration. (2013, April 26). *Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices*. Washington, DC: National Highway Traffic Safety Administration. Available at www.govinfo.gov/content/pkg/FR-2013-04-26/pdf/2013-09883.pdf

Bärgman, J., Boda, C. N., & Dozza, M. (2017). Counterfactual simulations applied to SHRP 2 crashes: The effect of driver behavior models on safety benefit estimations of intelligent safety systems. *Accident Analysis & Prevention*, *102*, 165-180.

Bärgman, J., Lisovskaja, V., Victor, T., Flannagan, C., & Dozza, M. (2015). How does glance behavior influence crash and injury risk? A 'what-if' counterfactual simulation using crashes and near-crashes from SHRP 2. *Transportation Research Part F: Traffic Psychology and Behaviour*, *35*, 152-169.

Blatt, A., Pierowicz, J., Flanigan, M., Lin, P-S., Kourtellis, A., Lee, C., … Hoover, M. (2015). *Naturalistic Driving Study: Field data collection* (SHRP 2 Report No. S2-S07-RW-1). Washington, DC: National Academy of Sciences.

Caird, J. K., Willness, C. R., Steel, P., & Scialfa, C. (2008). A meta-analysis of the effects of cell phones on driver performance. *Accident Analysis & Prevention*, *40*(4), 1282-1293.

Davis, G. A., Hourdos, J., Xiong, H., & Chatterjee, I. (2011). Outline for a causal model of traffic conflicts and crashes. *Accident Analysis & Prevention*, *43*(6), 1907-1919.

Donmez, B., Boyle, L.N., & Lee, J. D. (2007). Accounting for time-dependent covariates in driving simulator studies. *Theoretical Issues in Ergonomics Science, 9*(3):189-199.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189-211.

Gennarelli, T. A., & Wodzin, E. (2006). AIS 2005: A contemporary injury scale. *Injury*, *37*(12), 1083-1091.

Guo, F., Klauer, S. G., Fang, Y., Hankey, J. M., Antin, J. F., Perez, M. A., … Dingus, T. A. (2017). The effects of age on crash risk associated with driver distraction. *International Journal of Epidemiology, 46(1)*, 2017, 258–265.

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.

ISO. (2015). Road Vehicles –Transport Information and Control Systems –Detection-Response Task (DRT) for Assessing Attentional Effects of Cognitive Load in Driving (ISO/Dis17488), ISO/TC 22/SC 13/WG 08 (under development). Geneva: International Organization for Standardization.

Jenness, J. W., Boyle, L. N., Lee, J. D., Chang, C-C., Venkatraman, V., Gibson, … Kellman, D. (2015). *In-vehicle voice control interface performance evaluation* (Report No. DOT HS 812 314). Washington, DC: National Highway Traffic Safety Administration. Available at www.nhtsa.gov/document/vehicle-voice-control-interface-performance-evaluation-final-report

Jenness, J. W., Boyle, L. N., Guo, H., Lee, J. D., & Chang, C. C. (2015). *In-vehicle voice control interface performance evaluation: Bridge study*. (Unpublished manuscript). Washington, DC: National Highway Traffic Safety Administration.

Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweeks, J. D., & Ramsey, D. J. (2006, April). *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: National Highway Traffic Safety Administration. Available at https://vtechworks.lib.vt.edu/bitstream/handle/10919/55090/DriverInattention.pdf?sequence=1%26isAllowed=y

Lamble, D., Laakso, M., & Summala, H. (1999). Detection thresholds in car following situations and peripheral vision: Implications for positioning of visually demanding in-car displays. *Ergonomics*, *42*(6), 807-815.

Maddox, M. E., & Kiefer, A. (2012, September). Looming threshold limits and their use in forensic practice. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, No. 1, pp. 700-704). Thousand Oaks, CA: Sage Publishing.

Markkula, G., Engström, J., Lodin, J., Bärgman, J., & Victor, T. (2016). A farewell to brake reaction times? Kinematics-dependent brake response in naturalistic rear-end emergencies. *Accident Analysis & Prevention*, *95*, 209-226.

McLaughlin, S. B., Hankey, J. M., & Dingus, T. A. (2008). A method for evaluating collision avoidance systems using naturalistic driving data. *Accident Analysis & Prevention*, *40*(1), 8-16.

Mortimer, R. G. (1990, October). Perceptual factors in rear-end crashes. In *Proceedings of the Human Factors Society Annual Meeting* (Vol. 34, No. 8, pp. 591-594). Thousand Oaks, CA: SAGE Publishing.

Mortimer, R., Hoffmann, E., & Kiefer, A. (2014). The psychological and accident reconstruction “thresholds” of drivers' detection of relative velocity (SAE Paper No. 2014-01-0437).

Muttart, J. W., Messerschmidt, W. F., & Gillen, L. G. (2005). *Relationship between relative velocity detection and driver response times in vehicle following situations* (SAE Technical Paper No. 2005-01-0427). Warrendale, PA: SAE International.

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

Reed, M. P., & Green, P. A. (1999). Comparison of driving performance on-road and in a low-cost simulator using a concurrent telephone dialing task. *Ergonomics*, *42*(8), 1015–10.

Society of Automotive Engineers. (2000). SAE J2396 Surface vehicle recommended practice, definitions and experimental measures related to the specification of driver visual behavior using video based techniques. Warrendale, PA: Society of Automotive Engineers

Victor, T., Dozza, M., Bärgman, J., Boda, C. N., Engström, J., Flannagan, C., Lee, J. D., & Markkula, G. (2015). *Analysis of naturalistic driving study data: Safer glances, driver inattention, and crash risk* (Report No. S2-S08A-RW-1). Washington, DC: Transportation Research Board.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose, *The American Statistician*, *70*(2), pp. 129-133.

Young, R. A. (2017). Are rear-end crashes caused mainly by an interaction between glance duration and closure rate? In: Proceedings of Driving Assessment 2017: The 9th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, June 26-29, 2017, Manchester Village, Vermont. Iowa City, IA: Public Policy Center, University of Iowa, 2017: 347-354. Available at https://doi.org/10.17077/drivingassessment.1657

APPENDIX A. COUNTERBALANCE ORDERS FOR STUDY 2

Table 20. Order of counterbalance sequences

Trial	Sequence A	Sequence B
1	Hybrid Navigation Hard	Hybrid Radio Tuning
2	Hybrid Radio Tuning	Auditory Navigation Easy
3	N-Back	Hybrid Navigation Easy
4	Auditory Navigation Hard	Auditory Navigation Hard
5	Auditory Navigation Easy	N-Back
6	Auditory Radio Tuning	Auditory Radio Tuning
7	Hybrid Navigation Easy	Hybrid Navigation Hard
8	Auditory Navigation Hard	Auditory Navigation Easy
9	Hybrid Navigation Easy	Hybrid Navigation Hard
10	N-Back	N-Back
11	Auditory Navigation Easy	Auditory Radio Tuning
12	Auditory Radio Tuning	Auditory Navigation Hard
13	Hybrid Navigation Hard	Hybrid Radio Tuning
14	Hybrid Radio Tuning	Hybrid Navigation Easy

APPENDIX B. DRIVING INSTRUCTIONS TO PARTICIPANTS FOR STUDY 2

On-Road Fixed Speed Scenario Instructions

In this driving scenario, you will be following a lead vehicle. Do your best to leave a two second gap between you and the lead vehicle (using the lane divider, stay about five lane markings back from the lead vehicle). **The vehicle may vary in speed**. It is your job to keep pace with the vehicle so that you are always a 2-second gap behind.

Remember that driving safely is your top priority during today's session. During today's drive, you will be driving on the ICC. The majority of the drive, you will be driving in the right lane. Stay in the right lane and do not pass the lead vehicle. I will be in the car with you and I will give you instruction on when to change lanes and any other maneuvers you may need to do. In a moment, you will have a chance to drive a loop around the Westat campus, so you can get a feel for what it is like to drive this car. During this session, your most important responsibility is to always drive safely. OK, let's begin the practice drive.

[Instruct participant to make a circle around the Westat parking lot. Assist with putting the car into drive if necessary]

I will now start the tactor. Remember, the tactor is the thing attached to your shoulder and it will vibrate. When you feel the tactor vibrate, click the button that is attached to your finger against the steering wheel or with your thumb. Continue to click the button every time you feel the tactor vibrate. Do not hold down the button.

Are you comfortable using the tactor? Are you comfortable driving the car?

[If yes, direct them back to the parking space]

On-Road Variable Speed Scenario Instructions

In this driving scenario you will be following a lead vehicle. Do your best to leave a two second gap between you and the lead vehicle (using the lane divider, stay about five lane markings back from the lead vehicle). **The vehicle may vary in speed.** It is your job to keep pace with the vehicle so that you are always a 2-second gap behind.

Remember that driving safely is your top priority during today's session. During today's drive, you will be driving on the ICC. The majority of the drive, you will be driving in the right lane. Stay in the right lane and do not pass the lead vehicle. I will be in the car with you and I will give you instructions o when to change lanes and any other maneuvers you may need to know.

In a moment, you will have a chance to drive a loop around the Westat campus, so you can get a feel for what it is like to drive this car. During this session, your most important responsibility is to always drive safely. OK, let's begin the practice drive.
[Instruct participant to make a circle around the Westat parking lot. Assist with putting the car into drive if necessary]

Are you comfortable driving the car?

[If yes, direct them back to the parking space]

On-Road Brake Light Scenario Instructions

In this driving scenario you will be following a lead vehicle. Do your best to leave a two second gap between you and the lead vehicle (using the lane divider, stay about five lane markings back from the lead vehicle). **Occasionally, the brake lights of the vehicle in front of you will turn on.** When this happens, you need to TAP the brake as quickly as possible. You do not need to slam or depress the brake all the way. A light tap will do. When we start the practice, if you will hear a click when you depress the brake. That is all the pressure that is needed.

Remember that driving safely is your top priority during today's session. During today's drive, you will be driving on the ICC. The majority of the drive, you will be driving in the right lane. Stay in the right lane and do not pass the lead vehicle. I will be in the car with you and I will give you instructions on when to change lanes and any other maneuvers you may need to do. In a moment, you will have a chance to drive a loop around the Westat campus, so you can get a feel for what it is like to drive this car. You will follow the lead vehicle and tap the brake every time the vehicle in front of you illuminates their brake lights. The car in front of you might not necessarily slow down when this happens. During this session, your most important responsibility is to always drive safely. OK, let's begin the practice drive.

[Via radio, tell safety observer you are ready. The safety observer will instruct the platoon to drive around the Westat parking lot]

[Instruct participant to make a circle around the Westat parking lot. Assist with putting the car into drive if necessary]

- Trigger brake light approximately every 10-15 seconds

Are you comfortable responding to the brake light? Are you comfortable driving the car?

[If yes, inform safety observer and direct participant back to the parking space]

Simulator Fixed Speed Scenario Instructions

During the drive, you will also be asked to perform a series of different tasks, some of them using a voice recognition system. Examples of different tasks include changing the radio, selecting a restaurant for the navigational system, etc. I will provide more detail about these later.

We are about to start a practice drive using the driving simulator. In this driving scenario you will be following a lead vehicle. Do your best to leave a two second gap between you and the lead vehicle (using the lane divider, stay about five lane markings back from the lead vehicle). **Please treat the driving scenario as if you were driving your own vehicle. Remember that driving safely is your top priority during today's session.** Stay in the right lane and do not pass the lead vehicle. After I start the simulator I will help you put the car into drive. After driving for 1 minute, I will start the tactor. Remember, the tactor is the thing attached to your shoulder and it will vibrate. I will review what I need you to do when it buzzes later. After another minute, we will practice the tasks.

For the majority of this drive you will be driving around 60 MPH, but it is more important to maintain a 2-second gap between you and the lead vehicle than to drive at 60 MPH.

Assist the participant with starting the car and putting it into drive. Let the participant drive for 1 minute or until comfortable.

In a moment I will turn on the tactor attached to your shoulder. When you feel the tactor vibrate, click the button that is attached to your finger against the steering wheel or with your thumb. Continue to click the button every time you feel the tactor vibrate. Do not hold down the button.

After 1 minute of driving with the tactor, ask the participant if they are comfortable with the tactor. Make sure the participant understands how close/far away they should be from the lead vehicle.

Are you comfortable using the tactor? [*If yes, proceed*].

Simulator Variable Speed Scenario Instructions

We are about to start a practice drive using the driving simulator. In this driving scenario you will be following a lead vehicle. Do your best to leave a two second gap between you and the lead vehicle (using the lane divider, stay about five lane markings back from the lead vehicle). **The vehicle may vary in speed.** It is your job to keep pace with the vehicle so that you are always a 2-second gap behind.

Please treat the driving scenario as if you were driving your own vehicle. Remember that driving safely is your top priority during today's session. Stay in the right lane and do not pass the lead vehicle. After I start the simulator I will help you put the car into drive. After driving for 1 minute, we will practice the tasks.

For the majority of this drive you will be driving around 60 MPH, but it is more important to maintain a two second gap between you and the lead vehicle than to drive at 60 MPH. Remember, the vehicle may modulate speeds, so you will have to slow down or speed up accordingly.

Assist the participant with starting the car and putting it into drive. Let the participant drive for 1 minute or until comfortable. Give them feedback if they get too close or too far away from the car as the speed fluctuates.

After about a minute, ask the participant if they are comfortable with the driving. Make sure the participant understands how close/far away they should be from the lead vehicle.
Are you comfortable with driving? [*If yes, proceed*].

Simulator Brake Light Scenario Instructions

We are about to start a practice drive using the driving simulator. In this driving scenario you will be following a lead vehicle. Do your best to leave a 2-second gap between you and the lead vehicle (using the lane divider, stay about five lane markings back from the lead vehicle). **Occasionally, the brake lights of the vehicle in front of you will turn on.** When this happens, you need to TAP the brake as quickly as possible. You do not need to slam or depress the brake all the way. A light tap will do.

Please treat the driving scenario as if you were driving your own vehicle. Remember that driving safely is your top priority during today's session. Stay in the right lane and do not pass the lead vehicle. After I start the simulator I will help you put the car into drive. After driving for 1 minute, the brake lights will be triggered occasionally. After another minute, we will practice the tasks.

For the majority of this drive you will be driving around 60 MPH, but it is more important to maintain a two second gap between you and the lead vehicle than to drive at 60 MPH. While the brake lights will flash on the vehicle in front of you, the lead vehicle may not actually slow down.

Assist the participant with starting the car and putting it into drive. Let the participant drive for 1 minute or until comfortable.

In a moment, the brake lights will come on. When you see the brake lights illuminate, tap the brake as quickly as possible. We are timing your response. No need to hold down the brake. You may have to speed back up after responding to the brake light to ensure you are close enough to the lead vehicle.

- Press the brake light “mini sim test” button after 1 min
  - Once you click it once, it will display the time that it was last triggered
  - Trigger the brake light once every 10-15 seconds.
  - Make sure the participant experiences the brake light AT LEAST five times.
  - Ensure that the participant is close enough to the lead vehicle to see the brake lights.

After about a minute, ask the participant if they are comfortable with the brake light response. Make sure the participant understands how close/far away they should be from the lead vehicle. Are you comfortable responding to the brake lights? [*If yes, proceed*].

DOT HS 812 813
January 2020

U.S. Department
of Transportation

**National Highway
Traffic Safety
Administration**

14456-100419-v2