



U.S. Department  
of Transportation

**National Highway  
Traffic Safety  
Administration**



---

DOT HS 812 314

October 2016

# **In-Vehicle Voice Control Interface Performance Evaluation**

## **Final Report**

## **Disclaimer**

This publication is distributed by the U.S. Department of Transportation, National Highway Traffic Safety Administration, in the interest of information exchange. The opinions, findings, and conclusions expressed in this publication are those of the authors and not necessarily those of the Department of Transportation or the National Highway Traffic Safety Administration. The United States Government assumes no liability for its contents or use thereof. If trade names, manufacturers' names, or specific products are mentioned, it is because they are considered essential to the object of the publication and should not be construed as an endorsement. The United States Government does not endorse products or manufacturers.

Suggested APA Format Citation:

Jenness, J. W., Boyle, L. N., Lee, J. D., Chang, C-C., Venkatraman, V., Gibson, M., ... & Kellman, D. (2016, October). In-vehicle voice control interface performance evaluation (Final report. Report No. DOT HS 812 314). Washington, DC: National Highway Traffic Safety Administration.

## Technical Report Documentation

1. Report No. DOT HS 812 314	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle <b>In-Vehicle Voice Control Interface Performance Evaluation</b>		5. Report Date <b>October 2016</b>	
		6. Performing Organization Code	
7. Authors <b>James W. Jenness, Linda Ng Boyle, John D. Lee, Chun-Cheng Chang, Vindhya Venkatraman, Madeleine Gibson, Kaitlin E. Riegler, Daniel Kellman</b>		8. Performing Organization Report No.	
9. Performing Organization Name and Address <b>Westat, Inc. 1600 Research Blvd. Rockville, MD 20850-3129</b>		10. Work Unit No. (TRAIS)n code	
		11. Contract of Grant No. <b>DTNH22-11-D-00237 /0007</b>	
12. Sponsoring Agency Name and Address <b>National Highway Traffic Safety Administration 1200 New Jersey Avenue SE. Washington, DC 20590</b>		13. Type of Report and Period Covered <b>Final Report</b>	
		14. Sponsoring Agency Code	
15. Supplementary Notes <b>NHTSA Task Order Manager: Ritchie Huang</b>			
16. Abstract <p>The purpose of the National Highway Traffic Safety Administration project titled, <i>In-Vehicle Voice Control Interface Performance Evaluation</i> was to conduct empirical research about the use of voice control systems (VCS) by drivers and potential measures that could be used for evaluating possible distraction from using these systems while driving. 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. In total, the research team noted 22 themes that characterized participants' interactions with VCS. 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. Interactions frequently included several types of 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.</p> <p>Two other studies were conducted in controlled laboratory environments. Participants interacted with a "Wizard of Oz" voice control system while they performed a surrogate driving task with a driving simulator or with a computer-based collision detection task (CDT). Cognitive load was measured by performance on the ISO Tactile Detection Response Task. Eye glance measures, based on Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices 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. The findings of these studies will inform development of Phase 3 of NHTSA's Driver Distraction Guidelines.</p>			
17. Key Words <b>Driver distraction, in-vehicle technologies, voice control system, speech recognition, driver vehicle interface</b>		18. Distribution Statement <b>This document is available to the public through the National Technical Information Service, <a href="http://www.ntis.gov">www.ntis.gov</a></b>	
19. Security Classif. (of this report) <b>Unclassified</b>	20. Security Classif. (of this page) <b>Unclassified</b>	21. No of Pages <b>112</b>	22. Price

**Form DOT F 1700.7** Reproduction of completed page authorized.

# TABLE OF CONTENTS

Abbreviations, Acronyms, and Symbols .....	v
Executive Summary .....	vi
1 Introduction .....	1
1.1 Background .....	2
1.2 Human factors challenges using VCS in vehicles .....	3
2 Study 1 – Contextual Interviews With VCS Users.....	6
2.1 Purpose.....	6
2.2 Method .....	6
2.2.1 Study design.....	6
2.2.2 Participants.....	7
2.3 Data collection sites .....	9
2.3.1 Maryland driving route .....	9
2.3.2 Washington driving route .....	9
2.4 Instrumentation .....	10
2.5 Data collection procedure .....	11
2.6 Data reduction.....	12
2.6.1 Video coding procedures .....	12
2.7 Results.....	15
2.7.1 Qualitative results .....	15
2.7.2 Quantitative results .....	24
2.8 Discussion.....	30
3 Study 2 – Simulated Driving when using VCS .....	32
3.1 Background .....	32
3.2 Method .....	33
3.2.1 Participants.....	33
3.2.2 Driving simulator task.....	35
3.2.3 Voice control tasks.....	36
3.2.4 Tactile detection response Task .....	40
3.2.5 Independent variables and experimental design .....	40
3.2.6 Procedure .....	42
3.2.7 Dependent measures .....	43
3.3 Data Reduction.....	45

3.4	Results.....	46
3.4.1	Demographics .....	46
3.4.2	VCS task performance: Total task duration.....	46
3.4.3	TDRT performance measures .....	48
3.4.4	Criterion 1: Percentage of long eyes-off-road glances ( $\geq 2$ sec) .....	49
3.4.5	Criterion 2: Mean glance duration .....	51
3.4.6	Criterion 3: Total eyes-off-road time.....	53
3.5	Discussion.....	54
4	Study 3 – Collision Detection Task When Using VCS.....	56
4.1	Background.....	56
4.2	Calibration study to select collision detection task parameters .....	58
4.3	Method .....	59
4.4	Results.....	60
4.4.1	Demographics .....	60
4.4.2	VCS task performance: Total task duration.....	60
4.4.3	TDRT performance measures .....	63
4.4.4	Criterion 1: Percent long EOR.....	64
4.4.5	Criterion 2: Mean glance duration .....	66
4.4.6	Criterion 3: Total eyes-off-road time .....	67
4.5	Discussion.....	69
5	General Discussion .....	70
6	Conclusions .....	74
	References.....	75
7	Appendix A.....	85
7.1	Study 1: Contextual interviews.....	85
7.1.1	Voice Control Tasks Experience Questionnaire .....	85
1.1	.....	85
7.2	Study 2: Driving simulator study in Seattle.....	88
7.2.1	Total task duration .....	88
7.2.2	Tactile detection response task performance measure: Reaction time .....	89
7.2.3	Criterion 1: Percentage of long eyes-off-road glances ( $\geq 2$ seconds).....	90
7.2.4	Criterion 2: Mean glance duration .....	91
7.2.5	Criterion 3: Total eyes-off-road time.....	92
7.3	Study 3: CDT Study in Madison.....	94

7.3.1	Total task duration .....	94
7.3.2	Tactile detection response task performance measure: Reaction time .....	96
7.3.3	Criterion 1: Percentage of long eyes-off-road glances ( $\geq 2$ seconds).....	97
7.3.4	Mean glance duration.....	98
7.3.5	Total eyes-off-road time .....	100

## **ABBREVIATIONS, ACRONYMS, AND SYMBOLS**

ANOVA	analysis of variance
CDT	collision detection task
DOT	Department of Transportation
EOR	eyes-off-road
HVAC	Heating, ventilation, and air conditioning
MGD	mean glance duration
NADS	National Advanced Driving Simulator
SD	standard deviation
SIM	simulator study
SMS	short message service
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

## EXECUTIVE SUMMARY

The overall objective of this project is to develop a protocol and performance metrics for evaluating voice control systems used in vehicles. The term, “VCS” as used in this report refers to systems that respond to the user’s spoken utterances (input). Usually, VCS also provide auditory prompts and responses in the form of audible speech (output). In the driving context, VCS include applications that are fully integrated into the vehicle’s on-board computer systems, applications that may or may not wirelessly access cloud-based computer systems, and applications that are physically brought into the vehicle via mobile, nomadic, or other connected devices. Other NHTSA documents have referred to VCS as, “auditory-vocal” driver-vehicle interfaces (e.g., NHTSA, 2013; Ranney et al., 2014), and automotive industry documents have used the term, “voice user interface” (e.g., SAE International, 2015). Conceptually, VCS may enable drivers to keep their hands on the steering wheel and their eyes on the road, and therefore may be less distracting than systems that demand visual-manual interaction. However, speech-recognition errors, complex interactions, and response delays might all draw drivers’ attention away from the road. Three related studies were conducted to provide an objective basis for the development of VCS evaluation protocols. The three studies included an on-road contextual interview study, a driving simulator study, and a laboratory-based collision detection task study. The CDT is an alternative driving surrogate task that requires the study participant to monitor a dynamic animation on a computer screen to identify spheres that appear to be on a collision course with the observer. Both the driving simulator study and the CDT study also included the tactile detection response task that requires the participant to respond as quickly as possible to a vibration stimulus. Performance on the TDRT is used as a measure the participant’s cognitive load.

The project focused on identifying practical assessment tests and protocols for voice control systems specific to vehicle applications. The project objectives were to:

- 1) Apply to evaluation of VCS the criteria concerning glance behavior and the driving simulator protocol described in the *Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices* (a.k.a. NHTSA visual-manual guidelines).
- 2) Develop an evaluation protocol to measure the cognitive demands of VCS and determine whether the TDRT is suitable for this protocol.
- 3) Develop a low-cost evaluation method for VCS which includes a driving surrogate task such as the CDT.

The contextual interview study was conducted in Rockville, Maryland, and Seattle, Washington, to identify drivers’ existing patterns of use with VCS while driving. There were 64 participants recruited (32 at each site) who indicated that they were currently using some form of VCS while driving. The hour-and-a-half contextual interview took place in each participant’s own vehicle. The entire contextual interview was video recorded (with audio). Drivers demonstrated many different VCS and many different VCS-enabled tasks. Drivers only demonstrated VCS tasks that they said they were familiar with performing when driving, and unlike a formal usability assessment, there was no attempt by researchers to standardize the list of tasks performed or the manner in which they were to be performed. Differences were observed between those who used



original equipment VCS and those who used portable smart devices that were paired to the vehicle. In total, the research team noted 22 themes that characterized participants' interactions with VCS. 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. Interactions frequently included several types of 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.

The team examined voice control for many vehicle makes and models. Almost all vehicles since 2012 have VCS installed as original equipment, or provide drivers the capability to interact by voice by connecting vehicle systems with their portable devices. Most VCS are implemented as part of a multi-modal interface and provide visual information within a center console, but there is no standard location, content, or format across automobile manufacturers.

The driving simulator study was conducted in Seattle and was designed to follow the protocol outlined in the NHTSA Visual-Manual Guidelines. The study included three tasks performed with a VCS: a radio task, a navigation task, and a calendar task. A total of 48 participants were included in the study. Participants performed each task while also responding to a TDRT per the ISO recommended guidelines (ISO NP/WD 17488, 2012). The dependent variables included measures used in the NHTSA Visual-Manual Guidelines and the ISO TDRT performance measure. More specifically, mean total task duration, mean TDRT reaction time, TDRT miss rate, and three eye glance criteria: percent of long eyes-off-road glances, mean glance duration, and total eyes-off-road time were used.

The third study used a driving surrogate task called the collision detection task and was conducted in Madison. The CDT requires the participating observer to view a dynamic display on a computer screen and to identify moving spheres that appear to be on a collision course with the observer. This study used the same voice tasks and TDRT protocol as in driving simulator study. A total of 48 participants were recruited for this study. The same dependent measures used in driving simulator and CDT studies to provide a comparison.

Results from these three studies help address several questions that motivated the project:

**1. Are commonly experienced VCS usability issues such as task complexity, interaction errors, and response delays potentially distracting for drivers?**

Increases in task complexity were found to result in longer task durations in all three studies. In addition, recognition accuracy and time delays in the VCS were also found to affect the task duration. Visual feedback was found to offset cognitive load in conditions where the VCS had poor voice input recognition accuracy, however, VCS designers might have to take into careful consideration the balance between cognitive and visual-manual distraction in designing VCS interaction interfaces. The pacing of the interaction could be important, given that the studies found that interaction with shorter delays resulted in higher cognitive load, as evidenced by performance on the TDRT.

**2. Are the NHTSA Visual-Manual Guidelines criteria relevant to VCS?**

These laboratory results, in combination with the contextual interview study, clearly demonstrate that drivers are more likely to look away from the road when engaged in voice-based tasks, suggesting that the NHTSA Visual-Manual Guidelines are relevant for VCS. Because voice interactions might produce cognitive load that is not reflected in glance

behavior, conforming to the NHTSA Visual-Manual Guidelines represents part of the needed assessment for VCS. However, the NHTSA Visual-Manual Guidelines alone are not sufficient for evaluation of VCS.

**3. As part of a VCS evaluation protocol, is the TDRT sensitive to task complexity, system delay, and recognition errors?**

There were differences observed in visual and cognitive demands depending on the design parameters of the VCS. A decrease in recognition accuracy and an increase in system delay increased visual demand as measured by glance behavior and reduced cognitive demand as measured by TDRT reaction time. These results have important implications for evaluating VCS, suggesting that TDRT measures of cognitive demand are a useful complement to measures of visual demand.

**4. Is a low-cost evaluation method, such as the collision detection task, a reasonable method for assessing VCS?**

The results suggest that the CDT might be a useful surrogate for driving or for a driving simulation study based on the protocol used. The trends in data obtained with the CDT were generally similar to those obtained with the simulator, but the CDT had greater sensitivity in differentiating between conditions. The CDT was more visually demanding when compared to the driving simulator protocol, and required active search for frequent hazards, whereas the simulator protocol required only visual attention to the road to maintain lane position and speed.

**5. What are the implications of the present findings for developing a protocol for assessing VCS?**

Some of these results have immediate application to evaluation, such as the need to consider the visual demand associated with even “pure” voice-based systems that do not have any visual displays. Other results require additional effort to integrate into an evaluation protocol. Focusing only on TDRT reaction time might lead evaluations to neglect potentially distracting systems that encourage drivers to take their eyes off the road. Focusing only on glance behavior might neglect potentially distracting systems that impose a high cognitive load. The combined findings imply that evaluation of voice-based systems cannot be achieved by relying on any single metric. Given that multiple metrics are needed, a protocol to “score” the various components should be developed. The consequence of driver errors (e.g., saying the wrong command or speaking at the wrong time) emerged as an important outcome that needs to be carefully scored in assessing a system. The contextual interviews showed that system interaction times increase as the number of errors increase. However, by some measures, such as TDRT reaction time, poor designs that induce errors might appear to be less distracting. These results suggest it may be useful to broaden the assessment protocol beyond a single summative pass/fail test.

In summary, the findings suggest that all VCS tasks as studied in this project conformed to NHTSA Visual-Manual Guidelines, demonstrating the substantial benefit of VCS relative to visual-manual interaction. Interaction errors and system delays with VCS are common and a normal part of current users’ experience. Therefore, it is important to consider these challenges in evaluating systems. In both the on-road and laboratory studies, VCS users often look away from

the forward roadway during user-system interaction errors. In fact, the typical interactions with multimodal VCS often include looking at a visual display and require manual inputs. Hence, the criteria based on the NHTSA visual-manual guidelines are appropriate for evaluation of VCS, but are not sufficient given the cognitive demands of voice interaction. The studies showed that increasing VCS error rate or increasing system delays is associated with increased glances away from the forward roadway, and decreased TDRT reaction time, which is a measure of cognitive load. In other words, glance measures and TDRT appear to assess different aspects of distraction. The CDT protocol appears to be a viable assessment method for driver distraction, yielding results similar to the driving simulator protocol in the NHTSA Visual-Manual Guidelines. The findings of the contextual interviews and laboratory evaluation also were complementary.

# 1 INTRODUCTION

The term “VCS” in this report refers to systems that respond to the user’s spoken utterances (input). Usually, VCS also provide auditory prompts and responses in the form of audible speech (output). In the driving context, VCS include applications that are fully integrated into the vehicle’s on-board computer systems, applications that may or may not wirelessly access cloud-based computer systems, and applications that are physically brought into the vehicle via mobile, nomadic, or other connected devices. Other NHTSA documents have referred to VCS as, “auditory-vocal” driver-vehicle interfaces (e.g., NHTSA, 2013; Ranney et al., 2014), and automotive industry documents have used the term, “voice user interface” (e.g., SAE International, 2015).

The primary objective of this project, *In-Vehicle Voice Control Interface Performance Evaluation,*” was to develop voice control assessment protocol and performance metrics specific to vehicle applications; these can include applications embedded within the vehicle’s on-board system, accessed from the “cloud” via cellular phone networks, or integrated within devices brought into the vehicle. Three empirical studies were conducted to provide an objective basis for developing VCS evaluation protocols.

NHTSA sought to investigate five core human-interaction areas related to VCS:

- 1) Visual display: Typically used to complement voice controls for instructional, menu, and recognition results.
- 2) Manual controls: For some operation such as a talk switch, or push-to-speak button.
- 3) Auditory display: Feedback that is provided in conjunction with VCS (i.e., not as independent interfaces).
- 4) Cognitive demands: A certain level of attention and cognitive ability is necessary to accomplish the varying levels of VCS tasks (simple to complex).
- 5) System performance: The user-experience as affected by recognition errors and potential response time delays due to network speeds, hardware, or implementation architecture.

Within this context, the general aims of this project included:

- 1) Provide an overview of the current products and research related to in-vehicle VCS.
- 2) Identify drivers’ existing patterns of voice control systems.
- 3) Examine alternative evaluation protocols that can be considered for evaluation of VCS.

From these general aims, the project focused on the following specific objectives:

- 1) Consider whether VCS usability challenges, such as task complexity, interaction errors, and response delays may be distracting for drivers.
- 2) Apply to evaluation of VCS the criteria concerning glance behavior and the driving simulator protocol described in the *Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices* (a.k.a. NHTSA Visual-Manual Guidelines) (NHTSA, 2013).
- 3) Develop evaluation protocols that can be considered for evaluation of VCS, particularly the tactile detection response task.

- 4) Develop a low-cost evaluation method for VCS that includes a driving surrogate task such as the CDT.

## 1.1 BACKGROUND

Nearly all new vehicles made in the United States since 2012 have some voice interface capabilities. As vehicle sensors, intelligence, and communication functions continue to advance, in-vehicle information systems become increasingly more accessible to drivers. Vehicle manufacturers are competing to offer more communication opportunities to drivers. Many manufacturers offer these features with voice control to minimize distraction, but speech-recognition errors, complex interactions, and response delays might all draw drivers' attention away from the road. The U.S. Department of Transportation is particularly concerned about the safety implications of distraction due to drivers' use of electronic devices and in-vehicle displays. NHTSA has begun issuing nonbinding, voluntary driver distraction guidelines to promote safety by discouraging the integration of excessively distracting tasks and displays.

VCS have the potential to reduce visual-manual distraction by keeping drivers' eyes on the road and hands on the steering wheel—drivers can control the functions of the device hands- and eyes-free (Putze & Schultz, 2012). This apparent ease of interaction makes it possible to offer drivers more complex features than may be safely possible with a visual-manual interface. However, in practice, VCS may require additional button presses and glances toward the in-vehicle display. Even without the button presses and off-road glances, interactions with VCS may also place unnecessary demands on drivers' attention such that driving performance and safety are compromised. In the case of a navigation destination entry on a production vehicle, the supposedly hands-free and eyes-free operation led to an average task completion time of over 90 seconds and an average of over 30 seconds of off-road glance time (Reimer & Mehler, 2013). Certain VCS perform better than others with respect to minimizing distraction.

Task complexity, interaction delays, user error, and VCS error all contribute to create potentially distracting interactions. Many in-vehicle systems require specific voice commands. Such systems can lead to user errors when drivers forget commands, which is more likely with complex tasks. Recent advances in speech recognition allow users to speak more naturally, but such systems are prone to recognition errors. For example, the Apple speech recognition software that is integrated into their mobile phones can analyze a user's speech and interpret the context and meaning of the request rather than using specific voice commands (Apple, 2013). In driving, many automobile manufacturers are incorporating more natural language interfaces as part of their vehicles' original equipment. Fitchard (2012) describes such a system for BMW vehicles, and other car companies are moving toward natural interactions. Natural language interaction is achieved by integrating onboard systems with cloud-based systems to achieve better recognition accuracy than might be possible with only the car's onboard computer. Use of cloud-based systems introduces a tradeoff between recognition accuracy and potential transmission delays associated sending data to the cloud-based system. Both delays and recognition errors might increase the cognitive demands as drivers identify and recover from the infelicity. Errors and delays can also lead drivers to look to a visual display to confirm their commands. Such errors and delays are a major focus of the study designs developed in this project.

## **1.2 HUMAN FACTORS CHALLENGES USING VCS IN VEHICLES**

Creating a VCS that enables drivers to access rich information sources without distraction faces several challenges. First, many interactions such as navigation destination entry are complicated, and voice interaction can be cognitively demanding. Second, accurate voice recognition in a vehicle is a substantial challenge and errors can distract drivers. Third, supporting natural language interactions to mitigate errors and task complexity is computationally and theoretically challenging. All of these factors lead to systems that might not conform to drivers' expectations, which is particularly likely given the range of drivers these systems must serve. Major issues that are addressed in the three studies of this project include task complexity, accuracy of the speech recognition system, naturalistic interactions with VCS, as well as individual differences.

### **Task Complexity**

VCS interactions can be cognitively demanding and the complexity of the task has been found to affect event detection. VCS system design is known to affect task performance, hence the same task can be differently demanding when implemented on different systems (Ranney, Baldwin, Parmer, Domeyer, Martin, & Mazzae, 2011). VCS demands have also been directly linked to slower response time to a lead vehicle braking (Lee, Caven, Haake, & Brown, 2001). The increase in time was attributed to the use of the same cognitive resources to interact with speech-based interfaces as those needed for driving. Engstrom, Johansson, and Ostlund (2005) also showed a reduction in standard deviation of lane position when participants were engaged in an auditory-vocal task.

The SAE Recommended Practice J2364 (SAE, 2004) standard recommends a 15-second maximum task completion duration, specific to visual-manual interfaces to minimize excess task completion times that could result in greater numbers of off-road glances. The standard was not intended to prescribe a safe limit of interaction. Similar to visual-manual interfaces, it is necessary to understand if task completion times are important for structuring and evaluating VCS interactions.

### **Speech Recognition Accuracy**

Speech recognition performance is critical because failures to understand drivers' commands increase the distraction potential of the system relative to error-free performance. Even a perfect speech-based system might distract drivers. Boril et al. (2012) indicated that the interaction with a speech recognition device itself usually leads to a more complex cognitive task than the interaction with passenger. Due to the limited vocabulary and the requirement of clear communication, drivers have to concentrate on their interaction with the system. Recognition accuracy is an important safety factor (Kun, Paek, & Medenica, 2007). Drivers tend to physically move closer to the device or microphone if the device did not understand the input correctly, leading to more frequent lane departures. A high error rate could lead drivers to visually verify commands and revert to visual-manual interaction. In addition, the cognitive demands and associated distraction potential associated with recovering from misinterpreted commands may be substantially greater than that associated with error free performance.

The automobile environment is a particularly challenging environment for VCS as there are many direct sources of noise both inside (e.g., passengers, climate control system, music/news

players) and outside the vehicle (e.g., honking horns, ambulances, passing trucks). Hence, VCS face substantial challenges in recognizing drivers' commands. There are many factors that can impact the error rate of voice control and the distraction potential of these systems including vocabulary size and confusability, isolated, discontinuous, or continuous speech, task and language constraints, use of scripted or spontaneous speech interface, and even adverse weather conditions. Background noise can also increase error rates in voice recognition systems and impact the user and system's ability to accurately discern the words spoken. Adding additional sound sources could also increase driver stress and annoyance.

### **Natural Interaction**

Properly implemented, a natural, conversational interaction with a VCS could reduce several of the challenges posed by task complexity and recognition errors. Natural conversation allows for pauses so that the participants can pace the conversation to accommodate other activities. Natural conversation is also robust to the occasional missed word. Unfortunately most in-vehicle voice systems require the user to push a button before using VCS. These push-to-talk systems help reduce recognition errors, but can increase drivers' workload (Fodor et al., 2012). This is particularly problematic if the driver attempts to activate the speech control while merging in and out of traffic, on curves, or in high pedestrian areas. Drivers often begin talking before the voice system is ready to analyze the speech. Drivers may also respond with increased stress and annoyance with each repeated push.

Another important aspect is the pacing of the interaction. As in conversation with a passenger, an intelligent voice controlled system will pace its speech based on the demands placed on the driver. Forcing a driver to respond very quickly, as well as forcing the driver to hold information in memory too long can increase heart rate and perceptions of workload (Reimer & Mehler, 2013). However, it is not clear how the pacing of the system should be operationalized. Cues such as pauses and speech inflections of the driver ("uhh," "uhmm"), which are likely the result of increased attention to the roadway demands, can be picked up by the VCS and used to modulate the pacing of the interaction. Further, match and mismatch of emotion and voice tones used by the VCS with the driver's current emotional state can have significant effects on driving behavior and distraction potential (Nass & Brave, 2005). Matching the tone of the VCS could improve driving performance.

### **Individual Driver Differences**

The diversity of drivers compounds the challenge of understanding the distraction potential of VCS. There are limited data on VCS usage in cars. All drivers can be considered a potential user, but not all drivers are technology savvy (Lo & Green, 2012). Although many users will adapt to the technology, drivers are a very diverse user population and segments of that population (e.g., older drivers) may be particularly vulnerable to speech recognition errors and distraction associated with recovering from those errors. A study by Reimer and Mehler (2013) found that the use of a VCS for address entry navigation task resulted in only 13.3 percent of participants conforming to the total eyes-off-road duration (less than 12 seconds) criterion of the visual-manual guidelines. Of even greater interest was that all the older adults in the study failed to conform to the criterion. Hence VCS interaction could differentially impact different age groups. In addition, substantial cultural differences regarding preferences and response to gender and accents can complicate the understanding of the driver response to VCS.

## **Summary**

The following three studies, aimed at developing evaluation protocols for VCS, were informed by these four basic issues concerning driver interaction with VCS: task complexity, recognition accuracy, natural interactions, and individual differences. The first study uses on-road contextual interviews to understand how people currently use VCS in natural driving environments. This study provides an indication of the types of tasks that drivers choose to perform using VCS in their vehicles. It also documents the length of time that these tasks require and how often drivers encounter errors in their interactions with VCS. The second study uses a driving simulator protocol to collect data on distraction associated with VCS design and performance, specifically task complexity, recognition error, and time delay in system response. These three variables capture some of the most salient aspects of variability in VCS performance, and to the extent that they contribute to driver distraction, evaluation protocols should be sensitive to them. The third study uses a desktop driving surrogate task with the same VCS design as the simulator study to understand if the desktop task can be a viable alternative protocol for VCS evaluation.



## **2 STUDY 1 – CONTEXTUAL INTERVIEWS WITH VCS USERS**

### **2.1 PURPOSE**

The goal of the contextual interview study was to identify drivers' use patterns with voice control systems that exist in current automobiles and to document any usability issues with VCS that may affect driving safety. This exploratory study was not designed to evaluate any specific VCS, nor was it designed to directly compare performance of different VCS. Rather, this was a qualitative study designed to document the types of non-optimal user/system interactions that occur for experienced VCS users who have systems manufactured prior to 2014. Both original equipment VCS and VCS implemented through portable devices (primarily cell phones) connected to the vehicle or operating independently within the vehicle were included in this study. The findings of this research will inform development of VCS evaluation methods and development of NHTSA guidelines for driver distraction.

A literature review highlighted several research needs and questions that are relevant to evaluating VCS. There were some key points from the literature review that were considered when developing the protocol for the contextual interview study:

- It is important to define the driving situations and associated tasks that may make the distracting effects of the VCS more prominent.
- There is a need to define representative driving situations to measure the frequency and types of speech recognition failures.
- There is a need to quantify cognitive demands of error-free interaction with VCS.
- There is a need to quantify cognitive demands of error recovery given the speech-recognition accuracy in representative driving situations.
- There is a need to quantify visual-manual demand associated with different VCS. Under what circumstances do drivers feel comfortable using particular voice interface functions while driving? Do drivers appropriately self-regulate their use of VCS when driving demands are high?
- Newer vehicles may include onboard, original equipment voice systems, but many drivers still feel more comfortable using a cloud-based system from a portable device, that can be easily synched to the vehicle. There are different ways that users may interact with these systems.

### **2.2 METHOD**

#### **2.2.1 Study design**

Drivers in Seattle and Rockville who currently use some form of VCS while driving were recruited and interviewed in the context of their own vehicles about their experiences using VCS. A large part of the video recorded interview consisted of having the participant demonstrate how he or she typically uses VCS while driving. The study was designed to capture drivers' typical behaviors as they interacted with their voice control system. During the contextual interview, a researcher rode along with the participant to provide navigation instructions through a

predetermined route. The researcher observed and took notes on the participant's interactions with the VCS, and asked clarifying questions as the driver demonstrated performance of voice control tasks. Contextual inquiry methods such as this often have been used as part of user-centered design processes (Bayer & Holtzblatt, 1997).

### **2.2.2 Participants**

A total of 64 drivers 19 to 65 years old, 34 women and 30 men, were interviewed for this study. All held valid U.S. driver licenses for at least two years. Participants in Maryland were recruited through advertisements on the website Craigslist, WesInfo (an internal Westat website for employees), the Gazette (a local newspaper), and by posting recruitment flyers on community bulletin boards around Montgomery County. Participants in Washington were recruited via Craigslist, and by posting flyers on bulletin boards around King County. Prospective participants from both sites completed a screening questionnaire by telephone. The screening questions concerned the participant's age, gender, driver license status, and details regarding their voice control system use. Only experienced and regular VCS users were included in the study. Participants at both study sites were compensated with \$100 for their time and travel expenses.

A broad range of user experience levels, vehicle models, and voice control system types were included in this study. Researchers found it more difficult in Washington than in Maryland to recruit participants who use original-equipment, vehicle-based VCS. As a result, the Washington sample includes a higher percentage of drivers who use cell-phone-based VCS. A breakdown of participants by system type and by site can be found Table 1. The numbers in parentheses represent participants for whom video data were corrupted or missing. Analyses for these four participants were based only on the interviewer's notes.

Table 1: Voice control systems used by study participants

<b>System Name</b>	<b>Type</b>	<b>Washington Participants</b>	<b>Maryland Participants</b>
Ford SYNC	Original Equipment	1	8
Uconnect	Original Equipment		3
Lexus Premium Total Technology Package	Original Equipment	2	1
Entune	Original Equipment	1	6
BlueLink	Original Equipment	1	2
Infiniti System	Original Equipment	1	1
Nissan	Original Equipment		1
BMW I - System	Original Equipment		2
Honda Hands Free BlueTooth	Original Equipment		1 (1)
AcuraLink	Original Equipment		2
Tesla Model S	Original Equipment	1	
Subaru	Original Equipment	1	
VW System	Original Equipment	1	
OnStar	Original Equipment	(1)	
Smartphone linked to OEM vehicles systems with after-market device	Portable, Aftermarket	3	3 (1)
iPhone Siri	Portable	13	
Samsung Galaxy S-Voice	Portable	5	
Google Now	Portable	1	
Windows Phone	Portable	(1)	
<b>Total</b>		<b>30 (2)</b>	<b>30 (2)</b>

## 2.3 DATA COLLECTION SITES

Two data collection sites were used in this study. One site was based at Westat's offices in Rockville, Maryland and the other site was based at the University of Washington in Seattle, Washington. The driving routes used at each site are described below.

### 2.3.1 Maryland driving route

The on-road driving portion of the interview was approximately 30 minutes. Two similar driving routes were used (Figure 1). These included a mixture of driving on a limited access highway, arterial streets, and low speed residential and commercial streets. Participants were randomly assigned (with counterbalancing) to drive either Route 1 or Route 2. Route 1 was 14.6 miles and participants drove on arterials first and then on I-270 heading back to Westat. Route 2 was 14.1 miles and participants drove on I-270 first and arterials back to Westat.

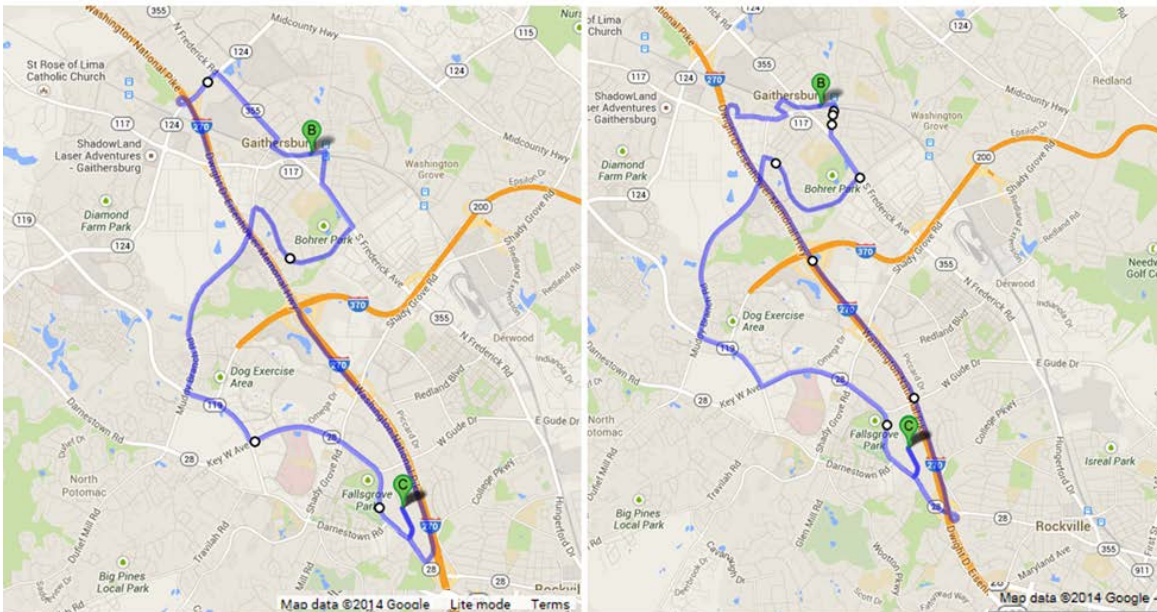


Figure 1. Maryland driving routes. Route 1 (left) and Route 2 (right) from Google Maps

### 2.3.2 Washington driving route

The route in Washington (Figure 2) was 13.2 miles long and took approximately 30 minutes to complete in light traffic. Washington assigned participants to drive in the clockwise or counter clockwise direction. Similar to the route in Maryland, the Washington route consisted of mixture of highway, heavy arterials, residential, and commercial streets.

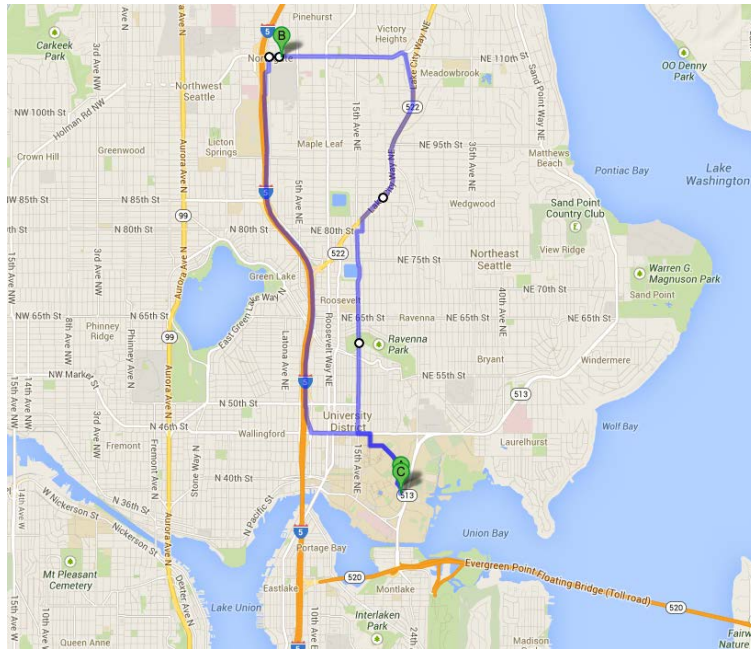


Figure 2. Washington driving routes (from Google maps)

## 2.4 INSTRUMENTATION

At the beginning of each contextual interview session, two battery-operated video cameras were installed into participants' personal vehicles. One camera (Contour Roam, model 1600) was mounted to the windshield and captured a view of the participant's face. The other camera (GoPro Hero 3) captured an over-the-shoulder view of the participant's hands on the steering wheel, the dashboard, and a view of the forward roadway. This camera was firmly mounted on a fixture attached to the headrest posts on the driver's seat (Figure 3).

Video and audio from each camera was recorded on 64GB SD media cards in the cameras. After each session the data from these cards was copied and stored on a computer and the SD cards were cleared for the next participant.

The experimenter and the participant each wore a lavalier microphone clipped onto their shirt, as close to their mouth as possible. This allowed for clear audio recording of all interactions with the system and between the experimenter and the participant. The lavalier microphones were plugged into the GoPro 3 camera and audio recordings were linked to the GoPro video footage.



Figure 3. Composite camera views of the steering wheel, dashboard, forward roadway, and participant's face

## 2.5 DATA COLLECTION PROCEDURE

Each participant completed an informed consent process prior to engaging in any data collection activities. The hour and a half contextual interview took place in the participant's own vehicle. For the duration of the data collection session, the participant was seated in the driver's seat and a researcher was seated in the front passenger seat. The entire contextual interview was video recorded (with audio) for later analysis.

The interview consisted of three segments:

1. An initial VCS-use survey with equipment set up period
2. A 30-minute drive on a pre-determined route while demonstrating how they typically use the voice control features in their vehicle
3. A short, final period after the drive to complete some final questionnaires and answer any follow-up questions from the drive

Prior to equipment setup, the experimenter reviewed the informed consent form with the participant. During the process, the experimenter showed the participant a map of the study route to give them an idea of the types of roadways they would be driving on.

The participant filled out a voice control use questionnaire (Appendix A) while the researcher installed the video cameras in the participant's vehicle. The participant was asked to report whether or not he or she used their voice control system to perform each task and if yes, to report the frequency and driving conditions under which they typically performed that task. The researcher and the participant then sat in the participant's parked car, and the researcher asked the participant to demonstrate a few of the voice control tasks. The experimenter then chose up to six voice control tasks for the participant to perform during the drive. These were selected from the set of tasks that the participant had reported doing while driving. The researcher told the participant to perform these tasks at any point in time during the drive without being prompted

by the researcher, whenever they were comfortable and in driving situations when they would most likely perform the tasks in real life. To summarize: Participants only demonstrated VCS tasks that they said they were familiar with performing when driving, and unlike a formal usability assessment, there was no attempt by researchers to standardize the list of tasks performed or the manner in which they were to be performed.

During the drive, the researcher provided step-by-step navigation instructions to the participant. While driving, the participant spontaneously performed the six voice control system tasks, only being reminded by the researcher of the tasks they had remaining to perform. The researcher asked the participant about their strategies for using VCS, errors encountered, and overall experience with the system.

At the end of the drive, the participant and the researcher returned to the point of origin and parked the car. Next, the participant filled out one NASA-TLX assessment of task load (Hart & Staveland, 1988; Hart, 2006; NASA, 2015) for each task performed during the drive. This instrument was administered as a paper and pencil questionnaire with six rating scales. Only the raw (unweighted) scores for the six subscales of the NASA-TLX were used in this study.

## **2.6 DATA REDUCTION**

Researchers' field notes were analyzed to find common themes that characterized participants' experiences using their VCS. This qualitative analysis included identifying examples and quotations where appropriate to support the themes.

All video data were synchronized, formatted, and composited by research staff in Maryland before being coded by researchers in Maryland and Washington. A researcher used Plural Eyes 3 software to synchronize the audio from the two video cameras. Once synchronized, the video files were imported into Adobe Premier video editing software where the researcher combined the video views from the two camera angles and rendered out a single composite video suitable for coding.

### **2.6.1 Video coding procedures**

Researchers in Maryland and Washington reviewed and coded the video recordings of the contextual interview drives. All coders were trained to follow the same protocols, which included using Morae Manager (TechSmith) software.

Within Morae Manager, a project template was created to define marker and task definitions. The markers used in Morae included a start of task marker, an end of task marker, an attempt marker, six error type markers, three confirmation type markers, and two types of manual input markers. Additionally, coders categorized each use of the VCS into one of five task types.

Begin and End markers were used to indicate the beginning and end of each discrete task-based interaction with the VCS and the elapsed time between these markers was subsequently computed as the system interaction time. The task began when the participant initiated an interaction with the system, usually by manually pressing a button or by speaking to the system. The task ended when the participant's goal for the task had been achieved or when the participant gave up and disengaged with the system.



Coders marked each attempt to complete a task that occurred over the course of the task. One attempt was recorded each time a new task began. If a task was successfully completed without error then there would only be one attempt. Otherwise, a new attempt was counted each time the participant needed to start the task over again by repeating the original command. Coders defined attempts as starting a task over from the beginning, and excluded cases where the participant was answering system questions in the “pathway” to completing a task.

The successful or unsuccessful completion of each task interaction was coded with an outcome marker. The three possible outcome markers assigned were “Yes” if the task that the driver set out to complete was successfully completed, “No” if the task that the driver set out to perform was not successfully completed, and “Kind of” if the driver completed a task (and seemed to accept the outcome) even though that task was not exactly the task that they set out to do. For instance, when demonstrating VCS tasks, participants occasionally accepted the consequences of a system error such as choosing to listen to the (wrong) radio station that the system heard rather than the station the participant intended. Other examples include accepting a different climate temperature than the one specified by the participant, or even accepting a phone call to a person other than the person that the participant clearly intended to call. In these cases, the participant was satisfied that they had demonstrated the task, even though there was clearly an error in the outcome.

Based on preliminary review of the data, six categories were created for subsequent detailed coding of non-efficient system interactions. For simplicity, we refer to these inefficient interactions as “errors” even though in some cases the system has performed as it was designed to perform. Table 2 lists the error types coded with a description of each type.

Table 2: Error Types

<b>Error Type</b>	<b>Description</b>
Clarification	System has a “misrecognition” and asks the participant to clarify what was said. This requires the participant to respond with some sort of “open ended” input. For instance, the system might say “pardon,” “excuse me,” “please repeat,” “system doesn’t understand” or some variation of that.
Premature Speaking	Participant speaks before prompted by the system. This can cause the system to cut the participant off (starts to provide information) or the system picking up only a portion of what the person said. For instance, hits the button and speaks immediately instead of waiting for the tone to sound or speaks before hitting the button at all. Person speaks out of turn (any time that they speak and the voice system is not “open”).
Read Options	This usually occurs after the system has a “misrecognition.” System provides list of alternative/potential options for the driver to select



	from. Requires the driver to respond with a “canned” specific option. For instance, you can say Bluetooth audio, navigation, radio commands, etc. Also if the system visually presents a list and says something like “please select a line number.”
System Timeout	User does not respond quickly enough and the system moves on to another prompt.
Mode Uncertainty	Mode confusion occurs when the participant tries to perform a task in one modality while still in another modality. They have not properly “backed out” of a system and therefore can only perform tasks within that modality until backing out. For instance, participant tries to perform a navigation task while still in the phone menu and the system will not execute. Usually says “voice command not recognized” or “that task is not available in navigation/communication,” etc.
Wrong Task	System has “misrecognition” and executes the wrong task. Confuses the auditory input from the participant with some acoustically similar command. For example, the participant says to call Mom and it calls Tom. Participant says phone commands and instead turns on the radio or participant wants to navigate to a certain radio station and the system leads them to a different station.

Researchers also coded each task performed into one of five general task categories: navigation, communication, entertainment, information, or climate control (HVAC). Table 3 provides examples of the types of tasks that fell into each category.

Table 3: Task Types and Examples

<b>HVAC</b>	<b>Communication</b>	<b>Entertainment</b>	<b>Information</b>	<b>Navigation</b>
Any task related to climate control. Set car temperature, turn on fan, adjust fan speed, defrost on, AC on, AC off	Place a phone call, send a text message, enter a new contact, redial, check voicemail, check e-mail	Turn on entertainment systems, change station (change radio frequency, go to preset), play music on an mp3 device, Pandora, etc., turn on CD, adjust the volume of music, listen to audiobooks, retrieve/update social media accounts, plan for a future activity (restaurant reservations, movie times, etc.)	Get vehicle related information (vehicle health report), trip status, miles driven, miles before fuel is on empty. Get fuel prices, weather, news, traffic information	Enter a known address, navigate to a favorite address, search a point of interest, get estimated time of arrival

After being coded, each video was checked by a second experienced coder for quality control (QC) purposes. During the QC process, the video coder checked to ensure that for each task there was a begin marker, an end marker, and an outcome marker. QC coders then watched the entire video to ensure that they agreed with the markers for errors, button presses, and attempts. For cases where there was a disagreement between the original coder and the QC coder, the two parties met to discuss the issues and come to an agreement. In rare cases, a third coder was enlisted to help resolve disagreements.

## 2.7 RESULTS

### 2.7.1 Qualitative Results

It should be noted that among the 64 participants in this study, many different VCS were used and many different VCS-enabled tasks were demonstrated. We were interested in both the variety and commonality of user experiences with VCS while driving.

Experimenters' notes and videos from both the Washington and Maryland site were analyzed to find common themes that related to the behavior of participants and usability of VCS. Themes were created to describe collections of similar observations from multiple participants. A total of 22 themes were identified. Most of the themes apply to use of original equipment VCS in vehicles as well as to use of VCS on cell phones while driving. The themes were grouped into five major categories shown conceptually in Figure 4. In this figure, the size of the ovals roughly represents the number of distinct themes in each category. Overlap between ovals is meant to suggest categories that were related to each other. All 22 themes are listed in Table 4 with supporting examples and interpretations derived from the qualitative analysis.

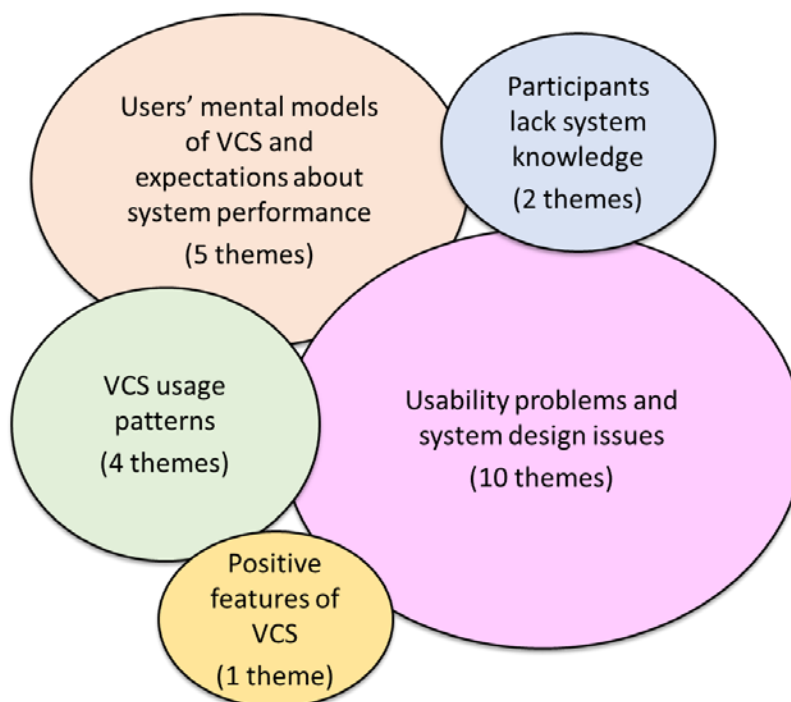


Figure 4. Conceptual representation of five categories of themes identified from qualitative analysis of contextual interviews

Table 4: Themes Emerging from Analysis of Contextual Interviews

Users' mental models of VCS and expectations about system performance	
Theme 1 – Anthropomorphism of VCS	<ul style="list-style-type: none"> <li>• It was common for users to reference VCS as another human being. This seems to influence their mental model of system function.</li> <li>• This may be an obstacle to overcome for people to learn to use the system. Some participants seemed to use system errors as an excuse for their own lack of training and understanding of the system.</li> <li>• Users have a need for immediate and frequent system feedback in their interaction with VCS, as if in a conversation.</li> <li>• One user said that he needed to speak loudly and clearly to the system, <i>“almost as if you are talking to an idiot.”</i></li> <li>• Other participants said things like, <i>“Sometimes she is stubborn,” “She doesn’t listen very well,”</i> and <i>“What’s wrong with you? Bad Navi!”</i></li> <li>• Throughout the drive, one participant kept referring to her VCS as <i>“Isabella,”</i> and when the system didn’t work properly, she said that this behavior was, <i>“One of Isabella’s moods.”</i></li> </ul>
Theme 2 - Users' expectations for system performance are modest – they tend to find some level of errors acceptable.	<ul style="list-style-type: none"> <li>• <i>“Voice systems are still in their infancy.”</i></li> <li>• <i>“If there is a technology and it is hard to use it, I’ll use it anyway.”</i></li> <li>• <i>“The errors that the system makes are to be expected. This is a computer after all.”</i></li> </ul>
Theme 3 - System performance seems worse on the demonstration day as compared to other days. Users claim that their system normally works better.	<ul style="list-style-type: none"> <li>• <i>“This isn’t the way it normally works. This typically works pretty smoothly but the system is crashing today.”</i></li> <li>• <i>“Maybe your [data collection] equipment is screwing with my Navi’s head. This [poor performance] is unusual.”</i></li> <li>• May indicate that users’ perception of system performance exceeds actual performance.</li> </ul>
Theme 4 - Users believe that noise in their vehicle reduces	<ul style="list-style-type: none"> <li>• Windows down</li> </ul>

recognition by VCS	<ul style="list-style-type: none"> <li>• Sunroof open or top down on convertible</li> <li>• Passengers making too much noise</li> <li>• This leads to users to try to compensate for perceived noise problems by shouting at the system and turning head to speak directly into the microphone.</li> </ul>
Theme 5 - Users tend to blame themselves for non-optimal user/system interactions	<ul style="list-style-type: none"> <li>• <i>“That could be human error where, you know, I may not say it clearly enough.”</i></li> <li>• <i>“It was my fault. I need to get out of USB mode before going into directions.”</i></li> <li>• <i>“It’s a robot. You should accept mistakes and try again.”</i></li> <li>• <i>“I’m sure by the time I get my next car the system will be much more advanced”</i></li> <li>• People seem to understand the limits of the technology.</li> </ul>
Participants lack system knowledge	
Theme 6 - Lack of knowledge of system features – not knowing what can be done by voice	<ul style="list-style-type: none"> <li>• Participants were often unsure which vehicle features were controllable by voice.</li> <li>• After filling out the Part 1 survey concerning their use of various features many participants said, <i>“Oh, I didn’t know you could do that with a system.”</i></li> <li>• Several people commented that after participating in the study they wanted to further explore their own VCS to see what features they could use that they currently didn’t know how to use.</li> </ul>
Theme 7 – People are not familiar enough with how to use their VCS - memory and learning issues	<ul style="list-style-type: none"> <li>• <i>“I don’t have the enough time or a desire to memorize all of the sequences of commands that are possible.”</i></li> <li>• Most users said that they learned to use VCS by trial and error.</li> <li>• I know the ones that I use on an everyday basis, but the other ones I usually need to use the help command or a menu structure to tell me the right commands.</li> <li>• Some people find VCS is not beneficial, since some tasks can be performed much quicker manually (e.g, selecting a radio station). This may provide a disincentive to learn about VCS.</li> <li>• For people who are not familiar with the system, it can be difficult to issue voice commands correctly. This may indicate that the system is not “natural” enough.</li> </ul>

VCS usage patterns	
<p>Theme 8 - Users' adapt their behavior to VCS</p>	<ul style="list-style-type: none"> <li>• Turn face toward (perceived) microphone location when speaking</li> <li>• Try to speak more loudly, clearly, or slowly than normal</li> <li>• Intentionally mispronounce certain names to achieve better system recognition</li> <li>• One participant noted that sometimes you need to change the words of a search query in order to get the answer that you want.</li> <li>• <i>"I end up just pulling over and doing it manually."</i></li> <li>• <i>"[The VCS] trained me more than I trained it. I'm speaking like a robot. I'm not speaking in a normal voice."</i></li> <li>• Acceptance of system mistakes <ul style="list-style-type: none"> <li>○ Some people changed their end goal to adapt to system mistakes. For instance, if they were trying to tune to a certain radio station or to change the temperature in the vehicle, they would accept if the system tuned to the wrong station or changed to the wrong temperature.</li> <li>○ <i>"I'm proud of her [the VCS] she at least did a task related to what I asked."</i></li> </ul> </li> </ul>
<p>Theme 9 - VCS use typically involves multimodal interaction</p>	<ul style="list-style-type: none"> <li>• Extensive use of visual displays when using VCS <ul style="list-style-type: none"> <li>○ Verbally choosing an option from a visual list is a common UI design (e.g., address entry) <ul style="list-style-type: none"> <li>▪ Select a line number or "please look at the list" and choose.</li> </ul> </li> <li>○ Help with valid commands <ul style="list-style-type: none"> <li>▪ Prompts such as, "Here's a list of things that you can say. . ."</li> </ul> </li> <li>○ Confirmation of what system heard and system state <ul style="list-style-type: none"> <li>▪ People indicated that they liked when they had an option to prevent the system from executing the wrong task. For example, when the system asks, "You want to Call Mom, is that right?" rather than dialing and calling the wrong number.</li> </ul> </li> </ul> </li> </ul>

	<ul style="list-style-type: none"> <li>▪ Prevents added distraction of needing to terminate a wrong task</li> <li>▪ Users seemed to have a strong need to know the system state and this need is not always met by VUI designs.</li> </ul>
<p>Theme 10 - Notable VCS interactions indicating variability between systems</p>	<ul style="list-style-type: none"> <li>• It was very rare to accomplish tasks with VCS without some use of manual inputs and visual displays.</li> <li>• Some systems do not support “barging.” They do not respond to the user’s attempts to interrupt a series of long prompts or a list of options.</li> <li>• For entering phone numbers, one participant mentioned that she wanted to be able to say the whole string of numbers at once rather than saying groups of three or four digits at a time.</li> <li>• One participant demonstrated how she always searches for directions to an address by using the destination phone number.</li> <li>• One participant noted that his system was configurable with respect to whether confirmations were turned on or off.</li> <li>• Interaction to get navigation directions using Siri involved a complex interaction. User unplugged phone from auxiliary cord connected to vehicle, pressed button on phone, asked Siri for directions and then plugged auxiliary cord back into phone and then manually selects “start” to begin navigation and hear directions read out through vehicle’s speakers.</li> <li>• One participant’s system did not provide much audio feedback but it did provide visual confirmations. He said that he likes this and would find audio feedback annoying.</li> </ul>
<p>Theme 11 – People vary in their use of smartphones while driving</p>	<ul style="list-style-type: none"> <li>• Some people had to password unlock in order to activate Siri, some people did not. <ul style="list-style-type: none"> <li>○ For a password locked setup, participant had to power on phone, input password, then hold the “home” button, issue voice command (4 steps).</li> <li>○ For an unlocked setup (no password), participant just has to hold the home button, issue voice command (2 steps).</li> </ul> </li> <li>• Many used Siri without any connection or pairing, while some people paired Siri via Bluetooth or connected iPhone to car speakers via AUX cable.</li> </ul>

	<ul style="list-style-type: none"> <li>○ At the Maryland site participants used Bluetooth, AUX cable, and several “after-market” devices meant to integrate a smartphone into the vehicle. For example, the BlueAnt system, Belkin, or Jabra.</li> </ul>
<p>VCS usability problems and system design issues</p>	
<p>Theme 12 - Mode confusion</p>	<ul style="list-style-type: none"> <li>• Mode confusion among participants was observed on several occasions. For example, some participants would try to locate a point of interest (POI) when still in an “entertainment” subsystem which would result in an apparent system error. Many people didn’t realize this mode was causing the error.</li> <li>• May be related to incomplete system integration</li> <li>• May be related to lack of feedback on system state</li> <li>• One participant commented that her system was “<i>mode stuck,</i>” meaning that when resuming interactions with VCS, the system stays in the last mode that it was in during the previous interaction. This can cause some mode confusion for users who initiate a new task.</li> </ul>
<p>Theme 13 - Users experience problems when trying to cancel a system action or exit by using voice commands</p>	<ul style="list-style-type: none"> <li>• System prompts are not always interruptible.</li> <li>• “Cancel” is not always accepted as a command by the VCS. <ul style="list-style-type: none"> <li>○ Is misinterpreted as another command, causing even more errors/issues/distraction.</li> </ul> </li> <li>• “<i>I’ll have to wait to get to a stopping point to turn this off.</i>”</li> <li>• Often, as an exit strategy, manual input is used to restart system from some initial state.</li> <li>• “<i>You can’t use the voice button to cancel when she [the system] is talking.</i>”</li> <li>• “<i>I say ‘No!’ but the system still doesn’t get it.</i>”</li> </ul>
<p>Theme 14 - Help functions</p>	<ul style="list-style-type: none"> <li>• Commands to get help differ between systems and sometimes do not match user expectations.</li> <li>• Users say things like, “<i>What can I say?</i>” “<i>Help;</i>” “<i>Tell me my choices.</i>”</li> <li>• Automatically playing help prompts is common as part of error handling routines.</li> <li>• Help messages are sometimes not interruptible.</li> <li>• Some users make extensive use of help functions as part of their normal interaction with the system. They use a help command first at each interaction point. This may indicate a</li> </ul>

	<p>need for a novice mode to reduce help requests.</p>
<p>Theme 15 - Unintended system inputs</p>	<ul style="list-style-type: none"> <li>• Conversations with passengers may be interpreted by VCS as commands.</li> <li>• Spoken directions provided by the navigation system were picked up by VCS. One participant reported that this happens frequently.</li> <li>• One participant believes that the VCS has trouble recognizing contacts stored in her phone because she included emoticons with the text of the names. This issue was not observed with original equipment VCS.</li> <li>• Button confusion was observed for users with multiple buttons on steering wheel.</li> <li>• During the drive, one participant accidentally pressed a button that started his VCS.</li> <li>• Confirmation steps were controversial. Some participants mentioned that confirmation steps were unnecessary or annoying, but several participants said that they liked having a confirmation step before placing a call so that they did not call the wrong number. One participant said that she didn't need to have a confirmation step because if she called the wrong number she could just cancel the call with a button press.</li> </ul>
<p>Theme 16 - System timing and pacing, timeouts</p>	<ul style="list-style-type: none"> <li>• System timeouts are frequent, perhaps too short for some users. (Maybe timeout thresholds should be adjustable by user.)</li> <li>• <i>“The system is in charge.”</i></li> <li>• <i>“I wasn't fast enough. The system timed out there.”</i></li> <li>• When driving conditions demand the driver's attention, the system will timeout.</li> <li>• Users are frustrated by frequent timeouts.</li> <li>• <i>“You go through all those steps and you get one thing wrong and the whole thing ends and you need to start over.”</i></li> </ul>
<p>Theme 17 – Inefficient command requirements and systems delays</p>	<ul style="list-style-type: none"> <li>• Most original equipment systems require a hierarchal sequence of commands in order to retrieve simple data (i.e., Searching for known address requires “State,” “City,” and “Confirmation” before proceeding to actual task), which create long delays, distraction and user frustration.</li> <li>• For most smartphone devices, a straightforward command (i.e., “Search Starbucks”), can be performed without</li> </ul>



	<p>unnecessary steps.</p> <ul style="list-style-type: none"> <li>• There are sometimes long delays in redirecting/recalculating navigation tasks which can lead to confusion about system status and frustration.</li> </ul>
<p>Theme 18 – Inability for VCS and user to communicate with “human slang”</p>	<ul style="list-style-type: none"> <li>• When giving a command, using typical slang terms (such as “UW” instead of University of Washington) could not be understood by VCS (original equipment or smartphone-based).</li> <li>• Cannot say a contact name in native language (e.g., “Ng” from Linda Ng)</li> </ul>
<p>Theme 19 – Difficulties and inconsistencies with use of terms within the voice user interface</p>	<ul style="list-style-type: none"> <li>• System commands often do not match user expectations.</li> <li>• Seemingly inflexible commands and command structure. Often there seems to be only one acceptable command for each situation (e.g., “end” is not an acceptable substitute for “cancel”).</li> <li>• Acceptable system commands often do not match terms used in feedback prompts.</li> <li>• One participant pointed out that a button on his steering wheel is labelled as the “Media” button so he refers to it as the “media button,” but the VCS refers to it as the “voice button.”</li> <li>• Another participant was having trouble with a navigation task and she wanted to cancel so she said, “Cancel.” The VCS did not respond to this command, but then she said, “Suspend route guidance,” and the system confirmed by saying, “Cancel.”</li> <li>• Another participant was entering a contact phone number for a new contact called, “Test Number Two.” After verbally entering the number he said, “Store” to keep the number, but the system interpreted “Store” as “Star.” To store a contact number, the correct command was “Enter.” However, after the participant successfully entered the number and then said, “Enter,” the system responded, “Test number 2 has been stored.”</li> </ul>
<p>Theme 20 – Technical problems pairing and synching phone and vehicle systems</p>	<ul style="list-style-type: none"> <li>• Difficulty with the synching procedure</li> <li>• Losing sync connection</li> <li>• Difficulties knowing if phone is synched with vehicle</li> <li>• Problems arise when multiple phones are synched in one vehicle.</li> </ul>

	<ul style="list-style-type: none"> <li>• Sync problem: Sometimes participants could not find the number they thought they had on the system.</li> <li>• Users return to using smartphone only functions and manipulation when driving even when synched because many common smartphone applications are not available through the link with the vehicle.</li> </ul>
<p>Theme 21 – Potential for driver distraction</p>	<ul style="list-style-type: none"> <li>• Participants tended to reach for, hold, and/or bring their smart mobile device closer to avoid background noises.</li> <li>• Some systems cannot be interrupted (e.g., barged in), even if the output is unwanted.</li> <li>• Distracting steps: unlocking phone, searching for voice control app (e.g., Samsung Galaxy S-Voice that does not have a dedicated hardware button to activate voice control)</li> </ul>
<p style="text-align: center;">Positive features of VCS</p>	
<p>Theme 22 - Positive features of VCS noted by participants</p>	<ul style="list-style-type: none"> <li>• Training of system to make recognition better</li> <li>• One participant noted that pausing her VCS is an option. This gives the user more time to think about what task she wants to perform or to make decisions, and it allows passengers to have a conversation that is not picked up by VCS.</li> <li>• Another participant said, <i>“One good thing about this system is that if it loses connection it refreshes and remembers where you left off. You don’t need to go back through the tasks again.”</i></li> <li>• <i>“The system seems to do well with [recognizing spoken] numbers.”</i></li> <li>• One participant mentioned liking the safety feature that her system doesn’t allow her to set up a new Bluetooth connection unless she is parked.</li> <li>• “Push to talk” feature increases user confidence and comfort with the system. <ul style="list-style-type: none"> <li>○ When first initiating a task, a single push to talk feature is preferred over a multiple step process to activate the system. <ul style="list-style-type: none"> <li>▪ For smartphones, if locked, the process of unlocking the phone leads to discomfort while traveling in various traffic scenarios (i.e., heavy traffic, higher speeds).</li> </ul> </li> </ul> </li> </ul>

The qualitative analysis of contextual interviewer notes and videos uncovered a wide variety of usability problems and insights regarding participants' behavior with VCS. Key findings relative to driving while using VCS are:

- 1) VCS are typically multimodal and there may be multiple ways to accomplish the same task (Theme 9). Visual-manual interactions require glances away from the forward roadway.
- 2) Drivers often do not know how to use their VCS system (Themes 6, 7) and this can result in doing tasks inefficiently (e.g., frequently asking for help, or encountering interaction errors). This prolongs the length of time that the driver spends interacting with the system.
- 3) Less than perfect of integration and consistency of the user interface between different VCS subsystems (infotainment, phone, navigation) can lead to usability problems and can distract the driver (Themes 10, 11, 12, 13, 19, 20, 21).
- 4) User acceptance of VCS may be partially attributable to the tendency for users to blame themselves rather than the system for interaction errors (Themes 1, 2, 3, 5). We speculate that incorrect or unrealistic mental models of user-system interaction may lead users to underestimate the task demands, interaction time required, and potential distraction for using VCS while driving.
- 5) There is wide diversity among drivers in their success with, and methods of using VCS (Themes 8, 10, 11, 16, 18). It is not clear whether VCS can support efficient interactions for all potential users.

## **2.7.2 Quantitative Results**

### **System interaction times**

For each VCS task demonstrated by a participant, system interaction times were calculated from analysis of video data. System interaction time was calculated as the elapsed time between the Begin marker and the End marker for a given task. These data included interactions where the task was successfully completed as well as interactions where the task was not successfully completed and the participant gave up, disengaged with the system. Often, non-efficient interactions (errors defined in Table 2) occurred and these typically resulted in longer system interaction times as compared to interactions that did not have any errors. Two examples of non-efficient interactions are provided here:

Example 1 - The user tries to end route guidance within a navigation application.

User: <Voice button press>

System: “Phone. Please say a command” (followed by a beep)

User: “End route”

System: (beeps twice) “I did not understand you. Pleas-“ (cut off by user hitting the voice button)

User: (mutters under breath) “What is it?”

User: (Speaks up) “Stop route”

System: (beeps twice) “I did not understand you. Please try again (followed by a beep)

The user then made one more attempt to end navigation. He hit the voice button again and told the system to “end route,” just like he did the first time. The system then asked him if he wanted to “cancel the route.” He replies “yes” and the route guidance ended.

Example 2: The user tries to place a phone call to a personal contact

User: <Voice button press>

System: (ding) “Sync. Please say a command”

User: “Phone”

System: (ding) “Phone. Please say a command”

User: “Call BPC”

System: (ding) “Read message. Is that correct?”

User: “No”

System: (ding) “Phone. Please say a command”

User: “Call B-P-C!”

System: (ding) “Calling BPC”

Bars in Figure 5 show the mean (+/- 1 SEM) system interaction times for 499 tasks demonstrated while participants were driving. The number of cases observed is shown inside each bar. Of the tasks observed, 242 (48.5 percent) had at least one non-efficient interaction (error). The mean system interaction time for all 257 interactions without errors was 22.6 seconds. System interaction times tended to increase as the number of errors encountered when performing the VCS task increased. For 15 percent of the cases observed, interaction time exceeded 65 seconds. Also, 13.6 percent of the cases observed included 3 or more errors.

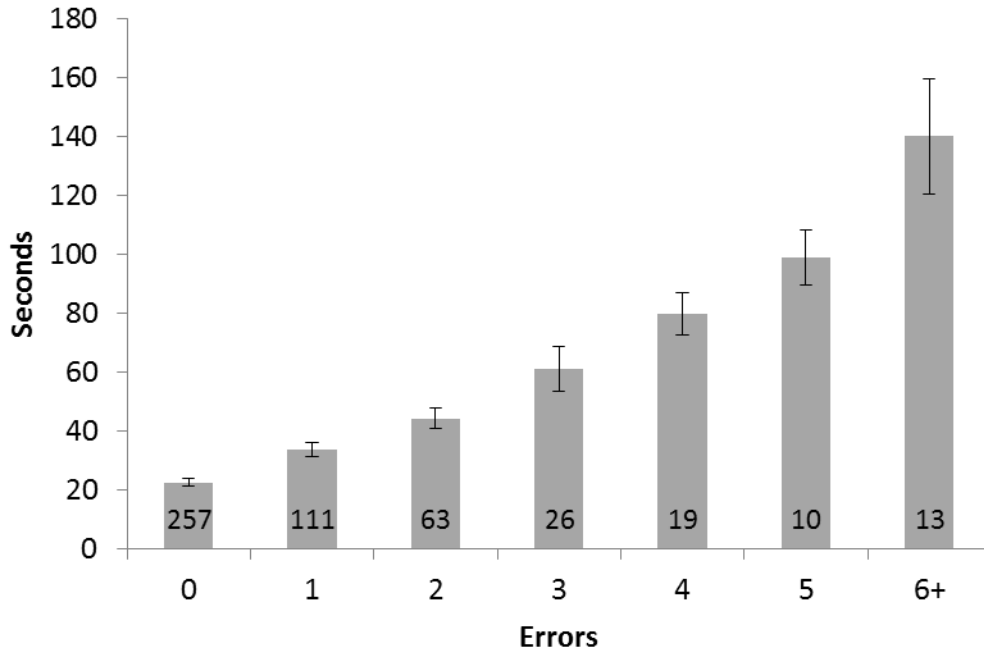


Figure 5. Mean interaction time with VCS for all attempted tasks grouped by number of interaction errors encountered

Mean interaction times are shown by task type in Figure 6. The two series of bars indicate mean interaction times for all interactions (dark bars) and for the subset of error-free interactions (light bars). It should be noted that not every participant is represented in each task type category and that certain participants demonstrated multiple tasks within a given category. Nevertheless, the data seem to indicate that participants tended to spend the greatest amount of time interacting with their VCS when performing navigation-related tasks and the least amount of time when performing tasks related to climate control (HVAC). Navigation tasks took an average of 33.3 seconds when no errors were encountered, but took an average of 51.0 seconds for all observed navigation task interactions. By contrast, HVAC tasks had a mean interaction time of only 11.2 seconds without errors and 14.7 seconds for all observed HVAC interactions.

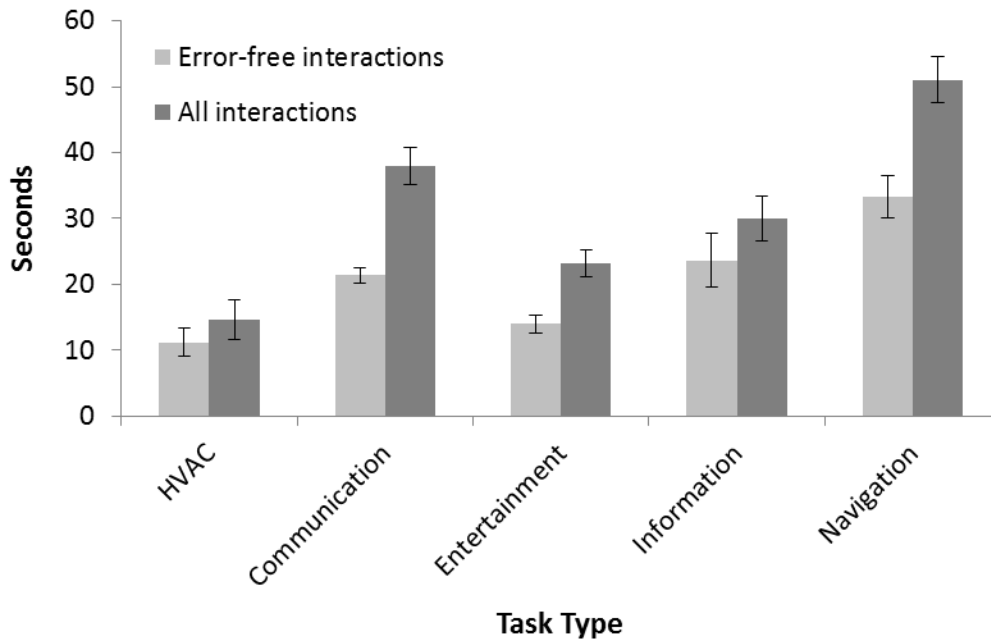


Figure 6. Mean interaction time with VCS by task type

### Errors and inefficient interactions with VCS

The total number of tasks analyzed within each task type is shown in Table 5 along with the overall error rates (errors per task) and error rates for the six specific types of inefficient system interactions (errors) defined in Table 2. The highest overall error rates were observed for communication tasks and navigation tasks, which were also the task types that were most frequently demonstrated.

Table 5: VCS tasks observed by type and error rates by error type

	HVAC	Communications	Entertainment	Information	Navigation
Tasks	n = 21	n = 148	n = 112	n = 65	n = 153
Error types					
Clarification	0.19	0.22	0.25	0.14	0.37
Premature speaking	0.05	0.15	0.12	0.11	0.12
Read options	0	0.14	0.04	0.12	0.21
System timeout	0	0.33	0.21	0.09	0.19
Mode uncertainty	0.24	0.07	0.02	0.05	0.08
Wrong task	0.28	0.39	0.23	0.18	0.32
All errors	0.76	1.30	0.87	0.69	1.29

### Task Completions With VCS

The overall observed rate of successful task completion was 76.4 percent. Table 6 shows additional details about completion outcomes for each task type. Information tasks had the highest rate of successful completion at 84.6 percent. The modest success rates for VCS tasks found in this study suggest that for VCS tasks commonly performed while driving, drivers often fail to succeed in accomplishing their tasks.

Table 6: Successful VCS task completion by task type

	HVAC	Communications	Entertainment	Information	Navigation
Tasks	n = 21	n = 148	n = 112	n = 65	n = 153
Outcome					
Complete and correct	15 (71.4%)	112 (75.7%)	81 (72.3%)	55 (84.6%)	118 (77.1%)
Accepted by user but not strictly correct	1 (4.8%)	7 (4.7%)	8 (7.1%)	1 (1.5%)	5 (3.3%)
Not completed	5 (23.8%)	29 (19.6%)	23 (20.5%)	9 (13.8%)	30 (19.6%)

### Subjective Ratings of Workload for VCS Tasks

After completing the contextual interview drive, each participant recorded their subjective ratings of task load using the NASA-TLX for each task type that they had demonstrated during their drive. Ratings on the six NASA-TLX subscales were scored from 0 – 100 and the means for each subscale were calculated across all participants for each of the five VCS task types defined in this study. There was a possible score of 600 (given the six ratings).

It should be noted that not every participant attempted tasks in all five task categories and that the task interactions included in each category varied considerably between participants due to the variety of specific tasks attempted and the variety of VCS interfaces used. With these caveats in mind, the mean task load scores are compared in Figure 7. Each segment of the stacked bars contains the mean value (rounded to nearest whole number) for the NASA-TLX subscale indicated. Higher values indicate greater task demand, frustration and effort, and poorer perceived performance. In general, the mean task load subscale scores were moderate to low (on the 100 point scales) for the VCS tasks performed in this study. Navigation types of tasks tended to have the highest task load scores and information tasks had the lowest. The rank ordering of mean subscale scores across task types was similar for the six subscales except for HVAC tasks which had relatively lower scores on Physical Demand.

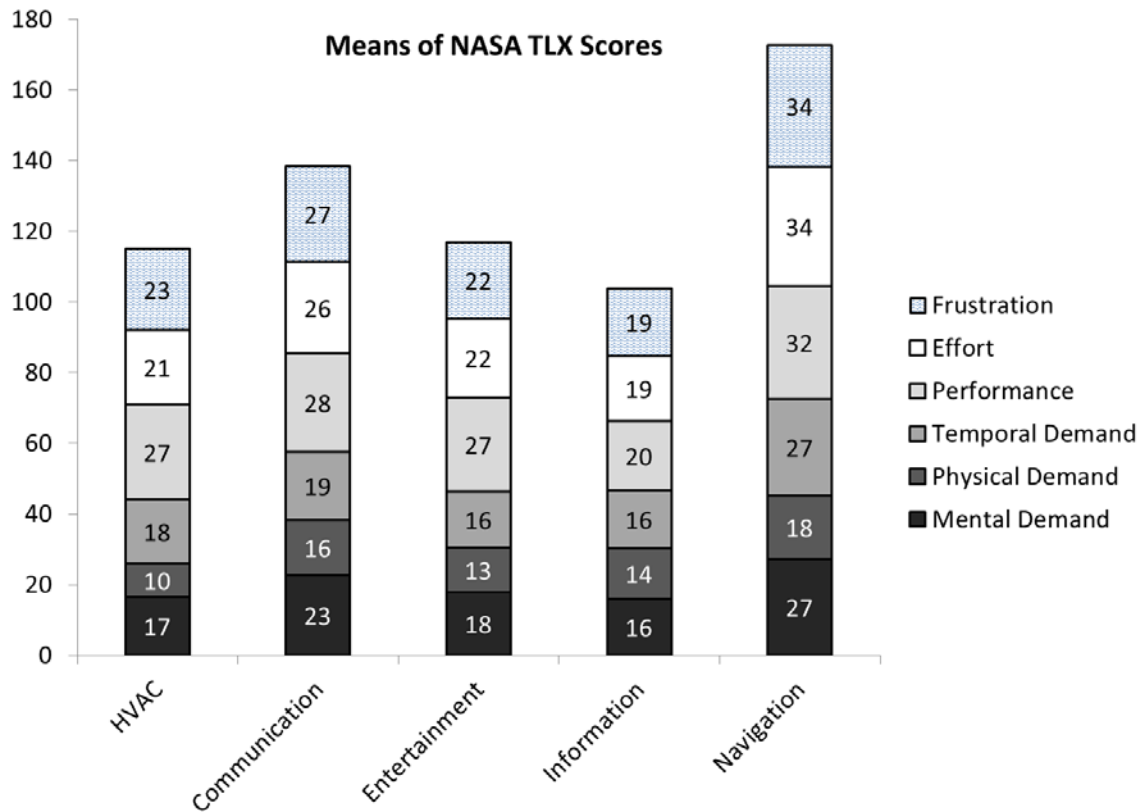


Figure 7. Mean NASA-TLX subscale measures of task load by task type



## 2.8 DISCUSSION

Conducting on-road contextual interviews with drivers using their own VCS was a useful method for learning about everyday use patterns, users' perceptions and understanding of VCS operations, system design features, and identifying usability problems. The 22 themes that emerged from qualitative analyses of interviewers' notes suggest that drivers often have trouble using their VCS but do not necessarily blame the system for the error prone and lengthy system interactions that they experience.

Detailed coding of interaction errors suggests that inefficient interactions are common. Only 51.5 percent of the interactions observed were error free and the time spent interacting with VCS when attempting to accomplish a task ranged from 14.7 seconds for HVAC tasks to 51.0 seconds for navigation tasks. Interaction times with VCS increased with the number of interaction errors encountered. Approximately 25 percent of interactions with VCS failed to result in task completion. The apparent mismatch between users' generally positive opinion of VCS and actual system performance suggests that drivers may underestimate the amount of time and workload imposed by VCS tasks.

The qualitative findings and insights gained from conducting this study have several implications for VCS design. Based on usability problems encountered by participants in this study, the following observations are offered to help VCS developers.

- A) **Easy error recovery:** As in human conversation, users tend to find occasional interaction errors with VCS to be acceptable. Despite this tolerance, errors observed in this study were clearly associated with increased system interaction times. System interaction times could be reduced by making error-handling routines more robust and efficient.
- B) **Controllable pacing:** System timeouts were frequently observed in the present study. Users were not able to give a command in the time allotted by the VCS. Conversely, some users tried to give commands before the system would accept and process it. Several users indicated that they would like VCS messages (especially help messages) to be interruptible and would like the VCS interaction to be user-paced.
- C) **Seamless integration of aftermarket systems:** Users reported some issues related to pairing their cell phones with vehicle systems. Making the pairing process easier, the pairing link more reliable, and the status of the pairing state more apparent may help alleviate these problems
- D) **Embodiment of human-like qualities:** Human verbal communication patterns may serve as an effective model for interactions with VCS. Users' in the study tended to anthropomorphize their VCS. Therefore, it may be useful to consider making VCS interactions more human-like in terms of turn taking behavior, use of natural speech patterns and vocabulary, and prosody. The design of efficient and highly acceptable interactions may benefit from a clearly defined human surrogate role and personality for the system (e.g., helpful personal assistant).
- E) **Consistency of the user experience:** Users often tried to use command words that were not recognized by the VCS and they pointed out several specific inconsistencies between terms used in VCS feedback and the recognizable command terms. User interface designs that have strict internal consistency would result in users being exposed to the same terms over and over which may help users to learn the system. While internal consistency of the

VCS user interface is important, users may need to be held to a lower standard of consistency than the system.

- F) ***Support for interruption and resumption of tasks:*** Given the length of the system interaction times, driving demands sometimes changed considerably during the interaction. Users had difficulty cancelling unwanted system operations, returning to a previous system state, and resuming tasks without backtracking. A simple, universal means of pausing an interaction and then resuming when the user is ready to continue may be helpful.
- G) ***Match users' mental models:*** Many users lacked knowledge of VCS function and often didn't understand which subsystem was currently active (e.g., navigation versus infotainment). This led to many inefficient interactions. More complete integration between subsystems may help solve this problem, but if that is not possible, clearer cues may be needed to convey which subsystem is active.
- H) ***Distraction potential of multimodal systems:*** Drivers who encountered errors with VCS often knew alternative ways to interact with their systems and they often reverted to manual input methods to accomplish their tasks. Additionally, visual displays were often used in conjunction with VCS. Users often looked toward these displays for visual confirmation of system status or to choose options from a list. Some users indicated that the visual and/or manual interactions incorporated in the VCS made tasks easier or faster, but they could also lead to potentially unsafe glances away from road. The demands of these alternative paths may be important to consider in overall VCS evaluation.

## 3 STUDY 2 – SIMULATED DRIVING WHEN USING VCS

### 3.1 BACKGROUND

Study 2 was conducted in controlled laboratory environments and consisted of drivers performing in-vehicle VCS tasks while driving in a driving simulator. This study had two goals: (1) to identify practical and credible measures to assess the distracting effects of voice control systems that can be used across many applications and (2) to validate the use of the simulator protocol described in the NHTSA Visual-Manual Guidelines for voice tasks.

To ensure that future NHTSA guidelines and evaluation protocols for voice control systems address actual and potential driver distraction issues, the research team studied the voice interaction behaviors of current vehicle owners' who have such systems (discussed in Chapter 2). As part of the front-end work, the study team also reviewed standards, guidelines, and best practices that are currently in place from:

- SAE standard (J2988: Guidelines for speech input and audible output in a driver vehicle interface)
- ISO 9921: 2003: Ergonomics – Assessment of speech communication
- ISO19358: 2002: Ergonomics – Construction and application of tests for speech technology
- ISO/IEC 2382-29: 1999: Artificial intelligence – Speech recognition and synthesis
- ISO 8253-3: 2012: Acoustics – Audiometric tests methods –Part 3: Speech Audiometry.

Many VCS guidelines generally consider voice interfaces that use onboard processing and a finite set of commands. To varying degrees, these guidelines are not well suited for the design of cloud-based processing systems that respond to more natural speech inputs. It should also be noted that some guidelines for vehicle interface design specifically exclude voice interfaces (e.g., European Statement of Principles on Human Machine Interface for In-Vehicle Information and Communication Systems; AAM Statement of Principles, Criteria and Verification Procedures on Driver Interactions With Advanced In-Vehicle Information and Communication Systems; Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices). However, these guidelines were still examined because drivers' interaction with VCS may still involve visual-manual components.

NHTSA developed voluntary driver distraction guidelines to encourage the design of in-vehicle devices that minimize driver distractions associated with secondary task performing during driving. The latest version (April 2013) was reviewed with respect to the acceptance criteria for the eye glance data, which is considered in this study as well (NHTSA, 2013).

The existing visual-manual guidelines provide some considerations applicable to the design and testing of VCS. For example, with respect to VCS with steering wheel-located voice activation button, NHTSA recommends that any tasks should be operable by using at most one of the driver's hands. This is particularly important given that most VCS also encompass some form of visual/manual controls. The guidelines also recommend that any device's "active display should be located as close as practicable to the driver's forward line of sight..." and specific calculation of the maximum downward angle is provided (page 38).

One consideration for examining the suitability of a driving task is the use of eye glance data and there are three acceptance criteria proposed in the visual-manual guidelines based on these measures (page 272 of the April 2013 guidelines):

- (1) 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 a testable task one time.
- (2) For at least 21 of the 24 test participants, the MEAN duration of all eye glances away from the forward road scene is less than or equal to 2.0 seconds while performing a testable task one time.
- (3) For at least 21 of the 24 test participants, the SUM of the duration of each individual participant's eye glances away from the forward road scene is less than or equal to 12.0 seconds while performing a testable task one time.

These acceptance criteria may help identify visual-manual distraction potential of VCS. Further, in examining these criteria along with measures of cognitive load, we can assess the applicability of these measures to VCS interactions. VCS can generate surprisingly high visual-manual demands, but complex VCS interaction, together with the recognition errors and system delays might produce substantial cognitive demands that could distract drivers. The ISO Tactile Detection Response Task was used as a measure of the cognitive load associated with VCS interaction. A recent NHTSA study comparing showed the TDRT task to be slightly more sensitive to other detection response task variants, such as head mounted visual cue (Ranney et al., 2014). This report also found that longer TDRT reaction time associated with the cognitive demands of the n-back task, but that the visual, manual, and cognitive demands of radio tuning produced greater reaction time increases than the one-back task. This study considers how the TDRT differentiates a range of voice-based interactions.

## **3.2 METHOD**

In this study, participants were asked to drive a simulated car while engaged in a VCS task and TDRT. The experiment was designed to also examine the impact of voice recognition error and system delay as the results from Study 1 (Contextual Interview) highlighted these as important VCS characteristics—Theme 2: Users' expectations for system performance are modest; Theme 5: Users tend to blame themselves for non-optimal user/system interactions; Theme 16: System timing and pacing, timeouts, and Theme 17: Inefficient command requirements and systems delays. There have also been several studies demonstrating that recognition error and system response delay undermine driving performance (Gellaty, 1998; McCallum, 2004; Kun, 2007).

### **3.2.1 Participants**

The research team in Seattle recruited 48 participants via campus e-mails, Craigslist, flyers, newspaper ads, and online advertisements. Interested individuals who contacted the research team were screened via a telephone interview. People who were willing to participate and met all inclusion criteria were scheduled for the laboratory study. Participants were compensated \$30 per hour for their participation. Parking validation also was provided for \$5 (weekend) and up to \$15 (weekdays) for participants who drove to the UW campus.

The visual-manual guidelines (pg. 264, April 2013) recommend recruiting participants from four age groups in equal numbers (18-24, 25-39, 40-54, and 55-75) balanced for gender. The

inclusion and exclusion criteria were based on a combination of previous experiences with simulator studies and the Visual-Manual NHTSA Driver Distraction Guidelines (Test Participant Recommendations, pg. 263). The screening requirements used from the April 2013 guidelines include:

- Be in good general health (no heart condition, seizure, epilepsy, Ménière's Disease, or narcolepsy),
- Be an active driver with a valid US driver's license, and
- Drive a minimum of 3,000 miles per year.

In addition, the study team also screened for:

- Be in the age range of 18 through 75 years of age, inclusive;
- Be comfortable using computer, touchscreens, and using voice control systems (in the home and car);
- Have any experience using automatic speech recognition systems;
- Be comfortable communicating via text messages using short message service (SMS), voice input, keypad input, or a combination;
- No participation in any driving simulator studies in the past 6 months; and
- Be a native English speaker.

We did not screen for the following, but was listed in the 2013 NHTSA Visual-Manual Distraction Guidelines:

- Have experience using a cell phone while driving, and
- Be unfamiliar with the devices being tested.

Participants were not asked about their experience with a cell phone while driving, as voice interactions do not necessarily need to be through a cell phone. However, participants were asked a similar question about their use of computers, touchscreens and voice control systems in the home and car. The team also did not ask about familiarity with the device, because the VCS used in the study was designed specifically for this study and familiar to nobody. That is, no participant had any prior experience with the VCS used in this study.

In addition, participants who used any special equipment to drive (i.e., booster seats, pedal extensions, hand brake or throttle, spinner wheel knobs, or seat cushions) or identify themselves as having a high likelihood of experiencing simulator sickness were excluded from participating in the simulator study.

Because voice tasks were used in the study, we also asked a series of situational questions associated with hearing. Respondents were asked to check all that applied. These questions did not serve as exclusion criteria but were used to generally assess participants' hearing ability.

- I sometimes feel that people are not speaking clearly (mumbling).
- When people address me from behind or from few feet away, I have difficulty understanding them.
- I have difficulty understanding people in meetings or group discussions.

- In situations with a high noise level (e.g., in restaurants, at parties or at the airport), I have difficulty understanding other people.
- I find it hard to hear birds singing, footsteps, running water, and other soft everyday sounds.
- I sometimes fail to hear the doorbell or telephone.
- I turn the television or radio up louder than other people. When someone else controls the volume, I have problems understanding.
- Other people have told me that I don't hear well.

Information was recorded on the types of voice control systems that participants were familiar with and how often they used these systems (i.e., daily for all possible use of voice, for select applications only [dialing, navigating, voice recorder], rarely or not at all). This questionnaire also assessed participants' knowledge and comfort in engaging in VCS while driving. Familiarity with VCS (based on frequency and length of use) was identified. The participant's experience with VCS did not necessarily have to be while driving, as this would have limited the number of drivers we could recruit. As part of the survey, information was collected on the participants' vehicle make and model, and how frequently they used voice systems while driving. This was similar to the questions asked in the contextual interview study (Study 1).

### **3.2.2 Driving Simulator Task**

Drivers were asked to maintain a 2-second time headway from the lead vehicle, drive safely, and stay in the center of the right lane during the entire drive (Figure 8). All simulator testing (even in the baseline condition of SIM only) was performed with the participant following a lead vehicle in the right lane of the simulated road. The simulated environment was similar to that in a previous NHTSA study of driver distraction from text reading and text input (Boyle et al., 2013) as follows.

- Four lanes, undivided with a solid double yellow center line
- Solid white edge lines
- Dashed white lines separating the two lanes that go in the same direction
- Flat, straight road (no horizontal or vertical curves)
- Posted speed limit of 55 mph..

The scenarios did not include any lead vehicle braking events, but the lead vehicle was traveling at a speed of approximately 50 mph, with variations from 50 mph according to a sinusoidal function, as previously defined in the text reading and text input study conducted by Boyle et al. (2013). The subject vehicle begins motionless in the right lane of the road, and proceeds to follow the lead vehicle for the remainder of the drive at an approximately two-second following distance.



Figure 8. NADS MiniSim setup

### 3.2.3 Voice Control Tasks

A Wizard of Oz testing protocol was used for the VCS interface, where a participant believes he or she is interacting with an automated speech recognition system, when in reality the behavior of the system is controlled by another human (called wizard) (Fraser & Gilbert, 1991). A screenshot of the wizard control interface is shown in Figure 9.

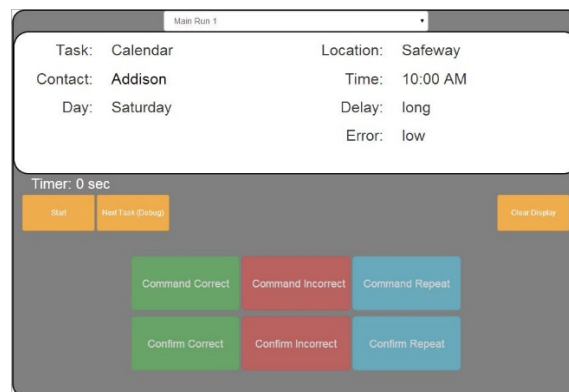


Figure 9. Screenshot of wizard interface (wizard view)

The Wizard of Oz (WOZ) methodology was selected over a cloud-based automatic speech recognition system like Google Now. Even though Google Now has fairly rapid and accurate speech recognition capabilities, the Wizard of Oz method allowed for precise and highly repeatable control of recognition error and system delay. This methodology has been widely applied to evaluate speech interfaces for in-vehicle applications. For example, Tsimhoni, Smith, and Green (2004) used a WOZ to compare speech recognition input with manual keypad input in a navigation system address entry task while driving. McCallum et al. (2004) used WOZ to assess driver distraction of speech interfaces in a simulated driving environment. Gellatly and Dingus (1998) used WOZ to examine speech recognition accuracy and recognition error type to see the impact it has on driving performance. Our voice-auditory tasks were designed to complement these previous studies.

The verbal in-vehicle task was accompanied by visual information that drivers were able to view on a 7" screen display located on the right side of the driver in a location similar to the center stack of a passenger car. The instructions were given using a male voice and aimed to emulate interactions with a passenger. The in-vehicle system prompts were given as a *female voice* (as

commonly heard in U.S. vehicles). The volume level of the voice prompt was 15 dB above the ambient noise level, to ensure that the sound of interest would not be masked (Wickens et al., 2004).

The study included three interactive VCS tasks along with the one-back task. The one-back task serves as a point of comparison to other studies.

1. **Radio Channel Selection Task:** Participants were prompted to select a radio station (e.g., “Please tune to Light Jazz”). There were six radio tasks per error and delay condition, or a total of 24 radio tasks presented over the entire study. The same 24 radio tasks were randomly presented to each participant (Figure 10). In each voice task, there were two types of interactions, depending on whether a system recognition error was absent or present.

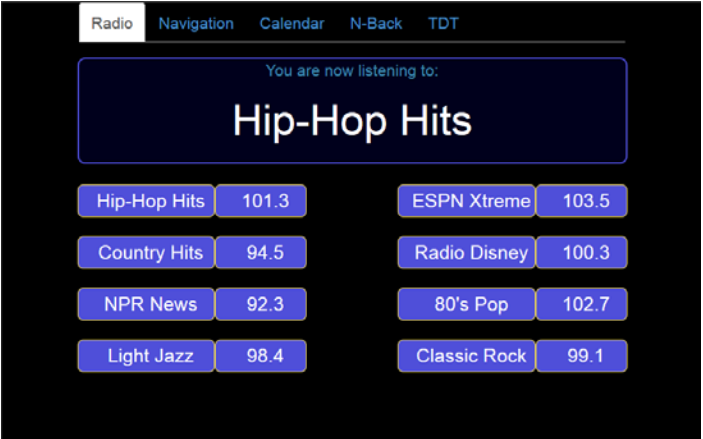
 <p>The screenshot shows a radio interface with a dark background. At the top, there are navigation tabs: 'Radio', 'Navigation', 'Calendar', 'N-Back', and 'TDT'. Below the tabs, a blue box displays 'You are now listening to: Hip-Hop Hits'. Below this, there are eight radio station buttons arranged in two columns. The first column contains: 'Hip-Hop Hits 101.3', 'Country Hits 94.5', 'NPR News 92.3', and 'Light Jazz 98.4'. The second column contains: 'ESPN Xtreme 103.5', 'Radio Disney 100.3', '80's Pop 102.7', and 'Classic Rock 99.1'.</p>	<p><b>RADIO EXAMPLE (RECOGNITION ERROR ABSENT)</b></p> <p><b>Male Voice:</b> <i>"Please tune to Hip-Hop Hits"</i> *Chime*</p> <p><b>Participant:</b> <i>"Play Hip-Hop Hits"</i></p> <p><b>Wizard:</b> [Presses "Command Correct" button]</p> <p><b>System:</b> (Text appears on 7" monition) <i>"You are now listening to Hip-Hop Hits. Is this the right station?"</i></p> <p><b>Participant:</b> <i>"Yes"</i></p> <p><b>Wizard:</b> [Presses "Confirmation Correct" button]</p>
--	---

Figure 10. Screen view of Radio task with example of voice interaction for the condition where without the voice recognition error.



**RADIO EXAMPLE (RECOGNITION ERROR PRESENT)**

**Male Voice:** *"Please tune to Hip-Hop Hits"*

\*Chime\*

**Participant:** *"Play Hip-Hop Hits"*

**Wizard:** [Presses "Command Correct" button]

**System:** (Text appears on 7" monition) *"You are now listening to Country Hits. Is this the right station?"*

**Participant:** *"No"*

**Wizard:** [Presses "Confirmation Correct" button]

**Male Voice:** *"Please tune to Hip-Hop Hits"*

\*Chime\*

**Participant:** *"Play Hip-Hop Hits"*

**Wizard:** [Presses "Command Correct" button]

**System:** (Text appears on 7" monition) *"You are now listening to Hip-Hop Hits. Is this the right station?"*

**Participant:** *"Yes"*

**Wizard:** [Presses "Confirmation Correct" button]

Figure 11. Example of voice interaction in radio task without the recognition error.

- 2. Navigation Task: Participants were instructed to verbally enter an address that consisted of a four-digit house number and a short generic street name (e.g., "Navigate to 5435 Main St"). There were three navigation tasks per error and delay condition, or a total of 12 navigation selections randomly presented to each participant over the entire study (Figure 12).

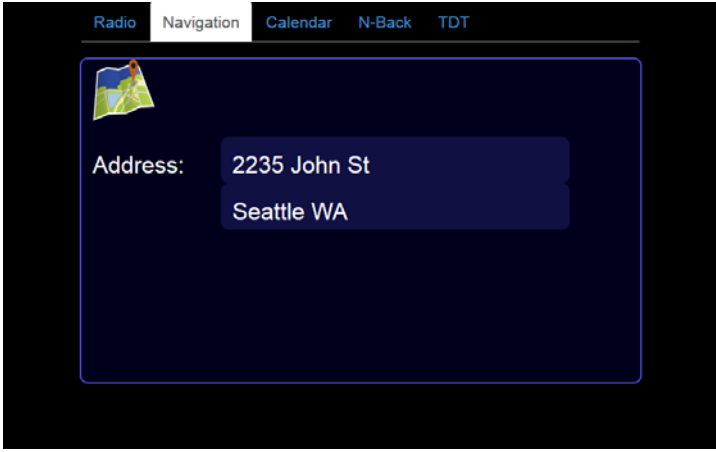
	<p><b>NAVIGATION EXAMPLE</b></p> <p><b>Male Voice:</b> <i>"Please go to 2235 John St., Seattle, Washington."</i></p> <p><b>Female Voice:</b> <i>"Where do you want to go?"</i></p> <p>*Chime*</p> <p><b>Participant:</b> <i>"Go to 2235 John St. Seattle, Washington"</i></p> <p><b>Wizard:</b> [Presses "Command Correct" button]</p> <p><b>System:</b> (Text appears on 7" monition) <i>"Do you want to go to 2235 John St. Seattle, Washington?"</i></p> <p><b>Participant:</b> <i>"Yes"</i></p> <p><b>Wizard:</b> [Presses "Confirmation Correct" button]</p>
---	---

Figure 12. Screen view of navigation task with example of voice interaction without the recognition error.

3. Calendar Entry Task: Participants were prompted to schedule an appointment with a specific contact name, at a specific location, time, and day (e.g., “Schedule an appointment with Luke at Starbucks on Friday 12 PM”). There were three calendar tasks per error and delay condition, or a total of 12 calendar tasks randomly presented to each participant over the entire study (Figure 13).

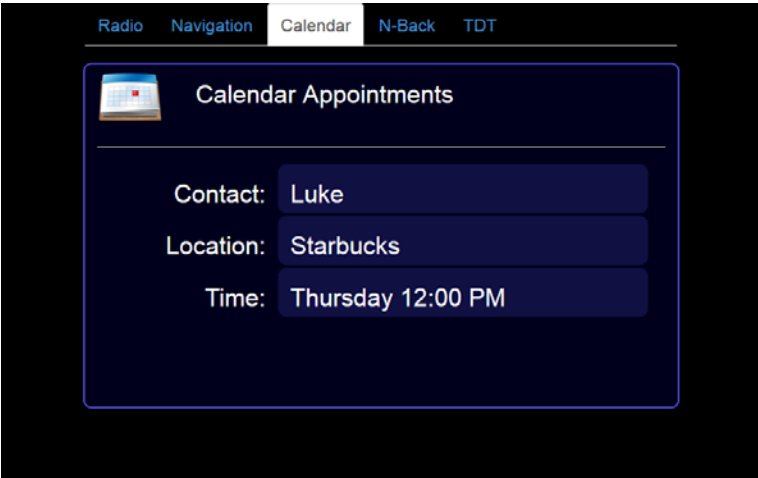
	<p><b>CALENDAR EXAMPLE</b></p> <p><b>Male Voice:</b> "Meet Luke at Starbucks on Thursday at 12PM"</p> <p><b>Female Voice:</b> "What would you like to schedule?"</p> <p>*Chime*</p> <p><b>Participant:</b> "Luke at Starbucks on Thursday at 12PM"</p> <p><b>Wizard:</b> [Presses "Command Correct" button]</p> <p><b>System:</b> (Text appears on 7" monition) "Do you want to schedule an appointment with Luke at Starbucks on Thursday at 12PM?"</p> <p><b>Participant:</b> "Yes"</p> <p><b>Wizard:</b> [Presses "Confirmation Correct" button]</p>
---	---

Figure 13. Screen view of Calendar tasks with example of voice interaction without the recognition error

4. 1-back (using the numerals 0 to 9). A 1-back task was used in the study. Table 7 shows the correct responses to 1-back task as compared to 0-back task. For this task, an automated female voice gave instructions to say the proceeding number. Participants responded to the system by saying the previously presented number. A sequence of twenty 1-back tasks was presented. Digits (of 0 to 9) were presented in a random order with replacement. Unlike the other three VCS tasks, no numbers (or visual feedback) was provided on the 7" monitor. After each correct response, the wizard pressed the "N-back correct" button to mark the approximate time and record the number of correct responses during the task.

Table 7: Stimulus and Response Sequence example for a 0-back and 1-back

Task	Digit Presented	9	4	2	3	5
0-Back	Correct Response	9	4	2	3	5
1-Back	Correct Response	-	9	4	2	3

The voice tasks are designed to have varying levels of complexity. The radio channel selection should be the least complex since the participant needs to recall only one chunk of information (name of station). The Calendar Scheduling task is considered the most complex because participants need to recall four chunks of information (contact name, location, time, and day). The Navigation tasks should fall between the Radio and Calendar because the participant needs to recall three chunks of information (street number, street name, and street type).

### 3.2.4 Tactile Detection Response Task

The Tactile Detection Response Task was used in place of the more widely used Peripheral Detection Task to assess the cognitive load of a secondary driving task. The peripheral detection task uses light stimuli, which makes it impossible to discern whether a missed signal is due to simply looking away or due to high cognitive load (Engstrom, 2010).

The TDRT uses a vibration stimulus produced by a small tactor that can be taped anywhere on the participant (Figure 14). There are different levels of sensitivity to tactile response given different users. The common protocol is to tape the stimuli to the participant's neck or wrist, with the responses provided through a micro-switch attached to the finger (see Engström, 2010). In this current study, the tactor was set up per Engström (2005). The tactor was taped to the neck (above the collarbone) and it vibrated randomly once every 3-5 seconds. The participant needed to respond to the vibration by pressing a micro switch mounted on the left index finger.

The draft ISO standards for Detection Response Tasks (ISO/NP WD 17488, 2012) state that reaction times greater than 2.5 seconds are to be marked as a "miss" and not included in the mean reaction time calculations. Similarly, reaction times lesser than 0.1 seconds should be regarded as invalid. The rate of misses during each task was recorded.

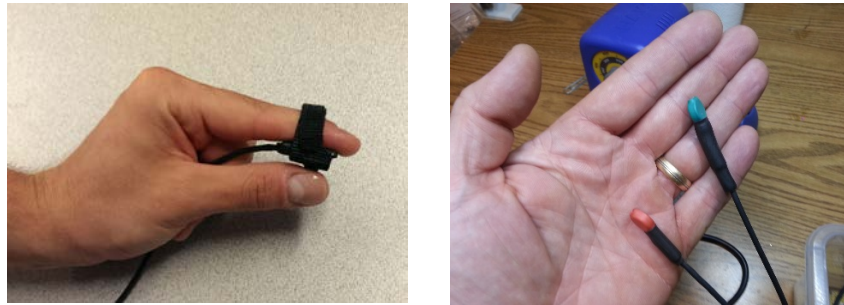


Figure 14. Hardware for tactile detection response task (for details, see <http://coglens.com/drt-device.html>)

### 3.2.5 Independent Variables and Experimental Design

The study used a mixed factorial, block design with two within-subject variables: the voice task (3 levels: Radio, Navigation, Calendar), and time delay (2 levels: Short, Long). One between-subject variable was included: recognition error (2 levels: present and absent). This 3x2x2 incomplete block design produced 14 different conditions (including the 1-back task in each between subject conditions as noted in Figure 15). Between these conditions, participants performed only the TDRT task and the order of the TDRT performance resulted in an additional between subject variable: order (2 levels: TDRT First, TDRT Second).

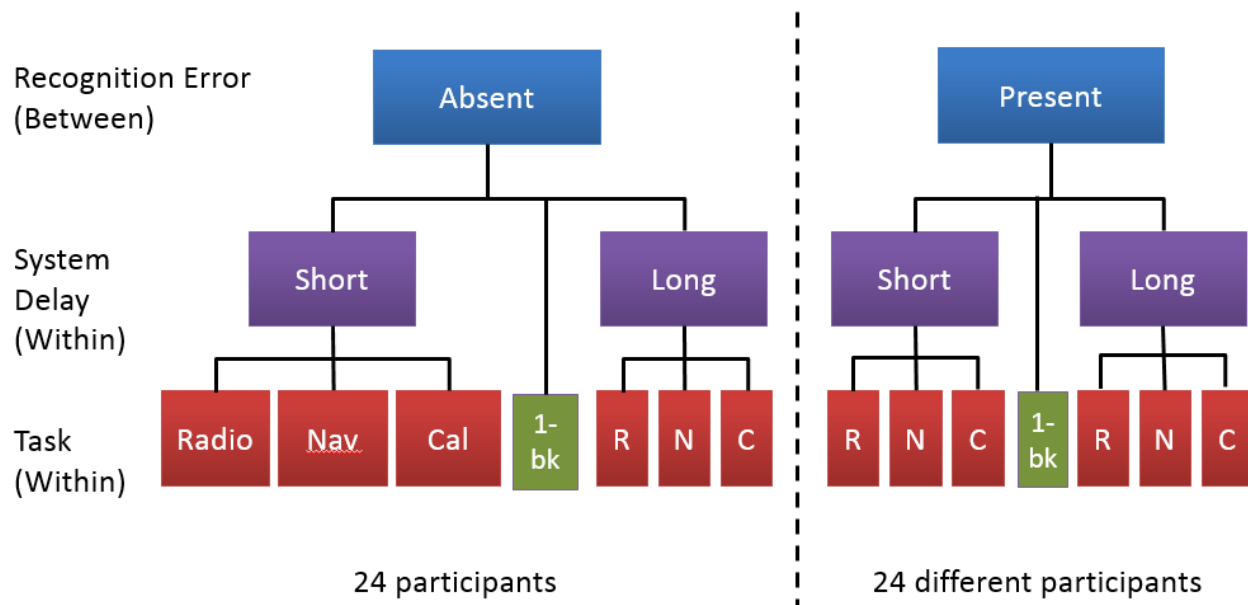


Figure 15. Experimental design

Task (3 levels): The three verbal tasks were Radio channel selection, Navigation, and Calendar. The order of presentation for these tasks was randomized within the time delay blocks.

Delay (2 levels—Short or Long): Time delay was defined as the time required for the voice recognition system to respond to the user. All participants encountered both short and long delays. The condition was blocked such that all tasks in the “Short Delay” condition were presented together and all tasks in the “Long Delay” were presented together with the 1-back task presented in between the two levels (Figure 16).

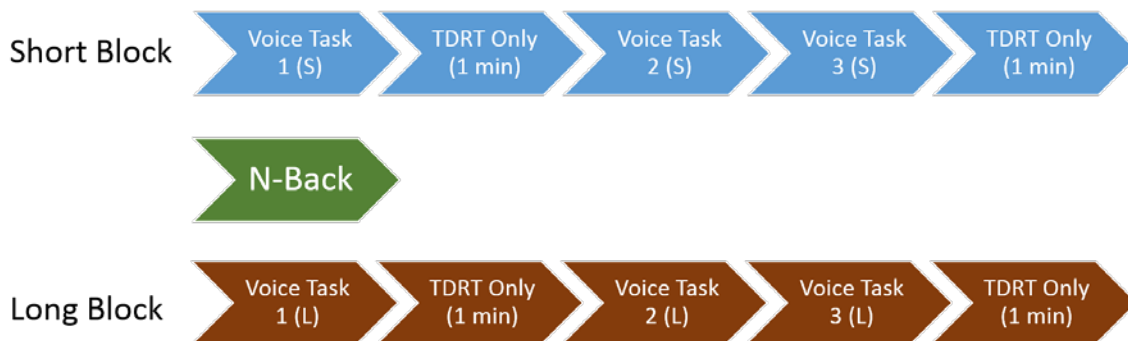


Figure 16. Voice Tasks were presented in blocks of short and long time delays

Based on Helder et al. (2010), the between-speaker-intervals, or intervals of silence in conversation in the English language, typically range from 200 to 1000 milliseconds. This interval was classified as short time delay, which mimics the response time of natural speech. The long delay provided feedback eight seconds after the wizard pressed a button in response to the participant. The 8-second delay was based on observation of the voice tasks in Study 1 (contextual interviews), which had relatively longer task durations. For both the short and long

delay conditions, the wizard always responded immediately to the participant, but in the long delay condition the system was designed to provide that feedback to the participant eight seconds after the wizard’s button was pressed.

Recognition Error (2 levels - Present or Absent): Voice recognition systems are imperfect and there may be safety implications with recognition errors. Ambient noise, acoustic similarity of commands, and length of spoken word can all undermine recognition accuracy (Gellatly, 1997). In this study two levels of recognition error were used: In the “Present” condition, the error rate was 66 percent, so that 2/3 of the tasks contained system recognition errors. The “Absent” condition did not include any system recognition errors. “Recognition Error” was a between-subject condition with 24 participants in the PRESENT condition, and another 24 participants in the ABSENT condition.

Order (2 levels – TDRT First or TDRT Second): Order is the sequence of tasks that the participant has to perform which are detailed in Table 8. Hence, if the participant performed the SIM/CDT trials first with the voice tasks and the TDRT second with the voice tasks, Order was denoted as “TDRT Second”; otherwise, it was “TDRT First.”

There were three combinations tested: TDRT only, TDRT+Voice Tasks, TDRT+Voice Tasks+SIM. In the TDRT+Voice+SIM condition, participants multi-tasked between the TDRT, issuing voice commands, and driving the car in the simulator. There were two different presentation orders (Table 8) and participants were randomly assigned to either Order 1 or Order 2. Each order had the same six data collection runs but in a different order.

The SIM only condition consisted of a 3-minute drive only. Participants were provided an opportunity to rest after run #3. During the break, the participant also filled out a demographic survey before beginning the second half of the experiment. At the conclusion of the experiment, the participant was provided compensation for their time.

Table 8: Experimental Run Order

<b>ORDER</b>		
<b>Run</b>	<b>Order 1: TDRT second</b>	<b>Order 2: TDRT first</b>
1	SIM only	TDRT only
2	SIM + TDRT + Voice Tasks	TDRT + Voice Tasks
3	SIM only	TDRT only
	<b>Break</b>	<b>Break</b>
4	TDRT only	SIM only
5	TDRT + Voice Tasks	SIM + TDRT + Voice Tasks
6	TDRT only	SIM only

### 3.2.6 Procedure

Once a participant had been screened and recruited, the research team arranged an appointment. Upon arrival at the laboratory, the experimenter confirmed that the participant had a valid driver’s license. Then the experimenter verbally reviewed the Informed Consent Form and participant signed a copy of this form. Each participant received instructions about the voice tasks, TDRT, and simulator task and had an opportunity to ask the researcher any questions that he or she may have about the procedures.

This was followed by a practice session that included three components:

1. Familiarization with the simulator,
2. Use of the simulator with the TDRT, and
3. Use of the simulator with the TDRT and the in-vehicle voice tasks.

Participants practiced the in-vehicle tasks included using voice control to select a radio channel, navigate to an address, schedule a calendar appointment, and perform a single digit memory recall with the 1-back task. When the participant felt comfortable multi-tasking between driving, responding to tactile feedback with the TDRT, and issuing a voice commands, the participant proceeded with the main portion of the experiment. The experiment itself was divided into two 20-30 minute sessions with a break in between each session.

For all the voice tasks (except 1-back), an automated male voice provided the commands for the participant to input into the voice recognition system. The male voice was meant to simulate instructions given by a passenger to tune to a radio station, enter a destination, or schedule an instruction. After the male prompt played, a female prompt played, which mimicked a VCS prompt, inquiring which destination or appointment the participant would like to enter. Immediately after the female voice, a chime sounded that indicated that the participant could begin responding. For the radio task, there was no female prompt, but the chime sounded immediately following the male prompt.

For participants' responses to be logged as correct, they needed to reproduce all parts of the command and only after that a chime sounded. If participants did not correctly reproduce the command or the response was given before the chime, the male voice prompt was repeated. This prompt could be repeated up to three times before moving on to the next task.

Confirmation that the WOZ heard the participant's voice command was provided visually and auditorily. A female voice mimicking the VCS provided the auditory confirmation and a visual confirmation was displayed on the 7" monitor. Depending on the delay condition, the response appeared immediately following the participants' response (Short Delay) or 8 seconds later (Long Delay). In the Recognition Error Present condition, if the participant correctly identified that the selection was not correct participants repeated the trial now with Recognition Error Absent.

### **3.2.7 Dependent Measures**

There were three categories of dependent measures examined: VCS task duration, TDRT performance, and eye glance behavior. TDRT performance indicates the cognitive demands of the task and eye glance behavior indicates the visual demands of the task.

VCS task completion time was measured by the time from when the task began (male prompt starts playing), until the time when the participant uttered the last response. Each test condition consisted of multiple trials of each task: 6 trials for the Radio task, 3 trials for the Calendar task, and 3 trials for the navigation tasks. In addition, a participant could attempt 3 times to complete each trial. For the tasks in which the system behaved ideally (e.g., Recognition Error Absent and Short Delay), the total task duration was expected to be less than the tasks that include a recognition error. The mean duration of trials within each condition was used in the analysis (e.g., 6 Radio, 3 Calendar, and 3 Navigation trials). The mean duration excluded trials where the

“wizard” entered a response incorrectly or trials in which the TDRT equipment malfunctioned. The total task duration for the 1-back tasks was the duration of the single 1-back task that was administered per participant.

The TDRT performance measures were based on the ISO standard (ISO NP/WD 17488, 2012). The reaction time to the tactile stimuli and the miss rate were calculated. The TDRT Reaction was calculated according to the ISO draft standards. Reaction times are only analyzed for hits. A hit on the TDRT trials were indicated by responses that occurred at least 100 ms after, but no more than 2500 ms after the participant receives each tactile stimuli. The reaction times are then averaged across one trial, and then across trials for each task in each condition. Figure 17 shows the expected lognormal distribution of TDRT response times that range from 100 ms to 2500 ms. The miss rate is the proportion of misses aggregated across trials for each task in each condition.

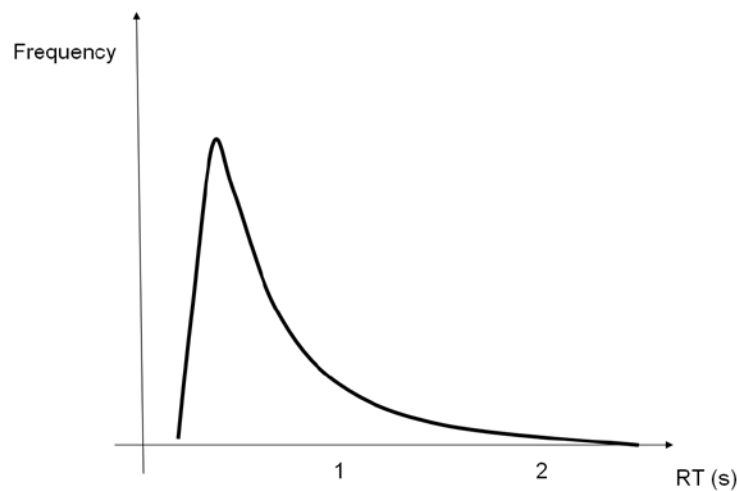


Figure 17. Model RT distribution for checking data quality (ISO NP/WD 17488, 2012)

Eye-glance behavior is measured according to the Visual-Manual guidelines. Video recordings of drivers were manually coded to identify the duration of off-road glances. These eye glances were analyzed using the visual-manual conformance criteria outlined in the most recent April 2013 version of the NHTSA guidelines.

1. Criterion 1: Percentage of long eyes-off-road glances: For at least, 21 of 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 greater than 2.0 seconds. For each participant, the percentage of long EOR for each trial is calculated as:

$$\% \text{ long EOR in trial } i = \frac{\text{No. of EOR} \geq 2 \text{ seconds in trial } i}{\text{Total No. of EOR in trial } i} \times 100$$

For example, if there are four glances in a single trial and two of them were longer than 2 seconds, then the percentage long EOR in this task would be 50 percent. Performance on a single trial conforms to criterion 1 when percentage long EOR is less than 15 percent.

Because each test condition was repeated at least 3 times for each participant, each test condition is examined using (1) the mean percent long EOR as well as (2) the maximum percent long EOR among trials within a participant. That is, the task conforms (or passes) Criterion 1 if mean or maximum percent long EOR is less than 15 percent for that task.

2. Criterion 2: Mean glance duration: For at least 21 of the 24 test participants, the mean duration of all eye glances away from the forward scene is **less than or equal to 2.0 seconds**. For each participant, MGD for a single trial is defined as:

$$MGD \text{ in trial } i = \frac{\sum_{j=1}^n EOR \text{ Duration}}{n}$$

where EOR duration is the  $j$ th EOR in the trial, and  $n$  is the total number of EOR in the trial. Similar to criteria 1, each task condition was examined with MGD averaged from the mean and maximum values of among trials. The test condition is considered not to conform if 15 percent of the study participants have MGD greater than 2.0 seconds.

3. Criterion 3: Total eyes-off-road time : For at least 21 of 24 test participants, the sum of durations of each individual participant's eye glances away from the forward scene is **less than or equal to 12.0 seconds**. The mean TEORT of the trials was used to evaluate whether a participant adhered to the 12 second criterion, given a specific experiment condition. Mean and maximum values among trials were calculated for the conditions. Hence, a test condition was considered in conformance if no more 15 percent participants had a mean TEORT greater than 12 seconds.

For the voice control system to conform to the visual-manual guidelines, 21 out of 24 participants must meet the criteria ( $24 \times 85\% = 20.4 \approx 21$  participants). Because there are at least three trials per task, a task may or may not conform to the criterion depending on how the trials are aggregated. For example, using the mean duration from three eye glance observations may minimize the impact of outliers, which allow easy conformance to the criteria. Choosing the maximum eye glance duration from the three trials makes the result highly sensitive to outliers, which also makes it more difficult to conform to the visual-manual guidelines. In this study, we focus on the maximum value among all trials within a test condition for the analysis of the eye glance criterion, but we provide the mean for comparison.

### 3.3 DATA REDUCTION

The reduced dataset included time-stamped data of the wizard inputs, eye glance video, and TDRT responses. The dependent measures were calculated per trial for each participant. The final data set excluded any tasks that included a wizard error (i.e., experimenter pressed the wrong interface button) or TDRT error (i.e., equipment malfunction). Because these exclusions resulted in an unbalanced number of trials per task per condition, the mean and maximum values of the trials for the three eye glance criteria were calculated to evaluate conformance with the NHTSA Visual-Manual Driver Distraction Guidelines.

A 2 (Recognition Error) x 2 (Delay) x 2 (Order) x 3 (Task) mixed factorial design was used to examine the dependent variables described earlier. The R statistical package (version 3.1.1) was used with the *lmer* function. The results were analyzed for the TDRT + Voice Task + SIM



combination session. The model included two within-subjects variables and two between-subjects variables. The within subject variables included the three Tasks (Radio, Navigation, Calendar) and two levels of Delay (Short and Long). The two between-subject variables were the Recognition Error (Present, Absent) and Order of trials as outlined in Table 8 (TDRT first, TDRT second).

According to the draft ISO standards for Detection Response Tasks (ISO/NP WD 17488, 2012) responses between 0.1 and 2.5 seconds were included for reaction time calculation. The TDRT Reaction Time variable was log transformed before analysis to fit the assumptions of homogeneous variance. The TDRT miss rates per task were calculated as the mean miss rate across trials.

### 3.4 RESULTS

#### 3.4.1 Demographics

There were 48 participants recruited in Seattle with an equal number of males and females from the four age groups. The mean age was 40.46 ( $SD=16.67$ ) for females and 39.44 ( $SD=15.97$ ) for males and Table 9 shows the mean age for each group. The age ranges from 19 to 73.

Table 9: Demographics of Participants From Study 2

<b>Age Group</b>	<b>Mean</b>	<b>Std. Dev.</b>
18-24	20.50	0.83
25-39	31.42	3.89
40-54	46.41	5.01
55-75	61.78	5.38

#### 3.4.2 VCS Task Performance: Total Task Duration

The total task duration was defined as the mean durations of trials for each voice task. Hence, the Radio task was averaged over six trials and the Navigation and Calendar tasks were each averaged over three trials. Task duration was expected to depend on the task type, recognition error, and system delay.

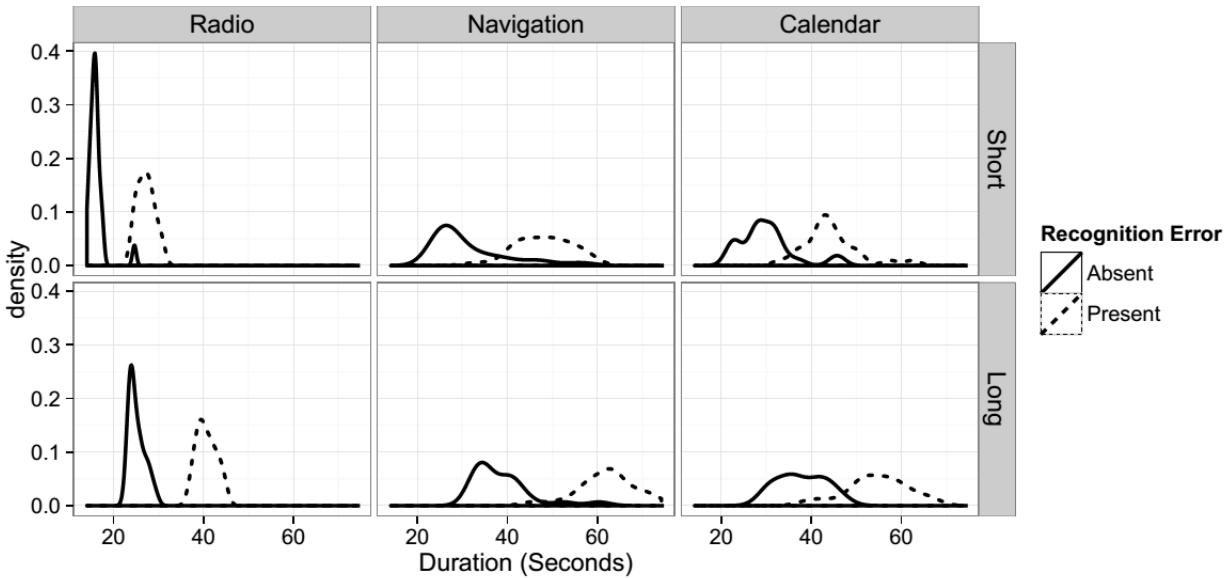


Figure 18. Distribution of total task durations across experimental conditions from Study 2.

Task type affected the task duration ( $F(2, 88) = 355.01, p < 0.001$ ) as shown in Figure 18. As expected, the more complex tasks took more time to complete, but the navigation and calendar tasks did not differ substantially. Recognition Error ( $F(1, 44) = 261.76, p < 0.001$ ) and Delay ( $F(1, 132) = 332.91, p < 0.001$ ) also influenced task duration. The long Delay condition, for instance, delayed the response of the voice control system by eight seconds to a command issued by the participant. If a participant in the “Recognition Error Present” condition correctly identified the system’s recognition error, then they had to redo the entire trial, thus increasing the total task duration. Errors and delays not only increased the mean task duration, but they also increased the variance of the task duration, making extremely long task durations particularly likely. More than 50 percent of the navigation tasks that were performed with recognition errors and long delays took over 60 seconds to perform. Two-way interactions of Task x Recognition Error ( $F(2, 88) = 11.45, p < 0.001$ ) and Recognition Error x Delay ( $F(1, 132) = 16.08, p < 0.001$ ) were observed. (Figure 35 and Table 20). Recognition Error Present conditions led to greater task durations than Recognition Error Absent for all tasks. The Navigation task resulted in the highest task durations followed by the Calendar, and then the Radio tasks (Figure 35). Long Delays increased task duration. Table 10 summarizes the mean, standard deviation, and range of the different task and experiment conditions.

Table 10: Summary of total task duration (in sec) from Study 2

Task	Recognition Error	Delay	Mean	Standard Deviation	Range	
					Minimum	Maximum
Radio	Absent	Short	16.18	1.95	14.10	24.63
		Long	25.22	1.69	23.34	29.27
	Present	Short	27.16	1.87	24.31	30.84
		Long	40.63	2.13	37.25	44.51
Navigation	Absent	Short	31.40	8.24	24.45	56.27
		Long	38.77	6.46	32.35	60.32
	Present	Short	48.00	6.12	33.34	57.70
		Long	61.78	6.30	45.33	74.17
Calendar	Absent	Short	30.35	6.45	22.31	46.11
		Long	38.67	6.16	29.85	55.66
	Present	Short	44.43	6.35	33.06	62.44
		Long	54.56	6.83	38.88	66.37

### 3.4.3 TDRT performance measures

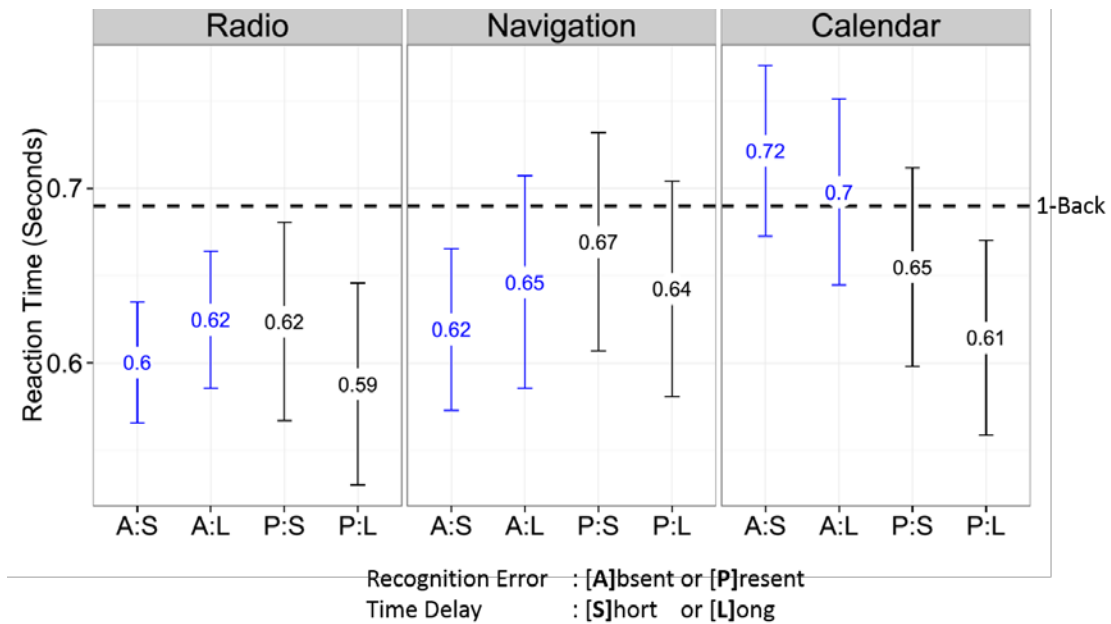


Figure 19. Mean TDRT reaction time across Task and experiment conditions from Study 2. The dashed line represents the mean TDRT reaction time during the 1-Back baseline task.

The mean TDRT reaction times were calculated from the averages of the trials for each condition, which was then compared to the 1-back task (Figure 19). The 1-back task was averaged across all participants (mean = 0.69 sec) and used as a baseline comparison to the VCS tasks. Task type affected TDRT reaction time ( $F(2, 89) = 10.22, p < 0.001$ ), with more complex tasks of Navigation and Calendar having a higher mean TDRT reaction times than the Radio task. A two-way interaction of Recognition Error x Task ( $F(2, 89) = 6.44, p < 0.01$ ) shows that the error and delay affect the Calendar Task differently than the Navigation and Radio task. More specifically, the TDRT reaction time for the Recognition Error Absent condition was greater than the Recognition Error Present condition for the calendar task, but the reverse is true for the navigation task. No other significant findings were observed.

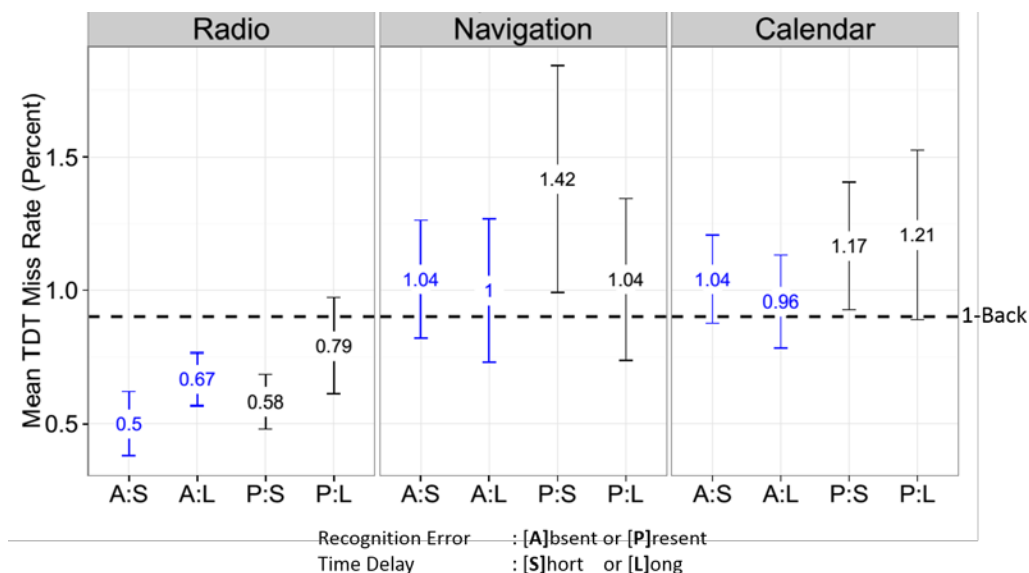


Figure 20. TDRT miss rates across task and experiment conditions from Study 2. The dashed line represents the mean TDRT miss rates observed during the 1-Back task.

The overall miss rate for the TDRT task is very low—less than two percent for all conditions. The results were well below the 30 percent miss rate (70% hit rate) threshold advised by the ISO standard. Figure 20 compares the mean TDRT miss rate percentages across Task, Recognition Error, and Delay combinations. The dashed line indicates the mean TDRT miss rates averaged across all participants for the 1-Back task. The mean miss rates for Navigation and Calendar tasks were found to be greater than the Radio and the 1-back task.

### 3.4.4 Criterion 1: Percentage of long eyes-off-road glances ( $\geq 2$ sec)

According to the NHTSA Visual-Manual Guidelines, an individual glance longer than or equal to 2.0 seconds is referred to as long EOR.

Table 11 shows conformance of the tasks to criterion 1. All tasks, under all conditions, conform to the criterion when calculated using the mean of the trials. All but one task conforms to criterion, when using the maximum of the trials. Surprisingly, the simplest task failed: Radio task with Recognition Error Absent and long Delay did not conform to criterion 1 with 21 percent participants having more than 15 percent long EOR. It should be noted that the Radio task was

repeated six times rather than three (as with the calendar and navigation task). Because there were six trials of the radio task, the long delays would cumulatively result in a higher maximum percent long EOR as well as a greater likelihood of a long EOR.

Table 11: Conformance with Criterion 1: Percentage of long EOR glances

Task	Recognition Error	Delay	Number Participant (out of 24) and % who do not comply			
			Mean		Max	
			Count	Percentage	Count	Percentage
Radio	Absent	Short	1	4%	2	8%
		Long	0	0%	5	21%
	Present	Short	0	0%	1	4%
		Long	1	4%	2	8%
Navigation	Absent	Short	1	4%	1	4%
		Long	1	4%	1	4%
	Present	Short	1	4%	1	4%
		Long	0	0%	1	4%
Calendar	Absent	Short	0	0%	2	8%
		Long	0	0%	0	0%
	Present	Short	1	4%	1	4%
		Long	0	0%	0	0%

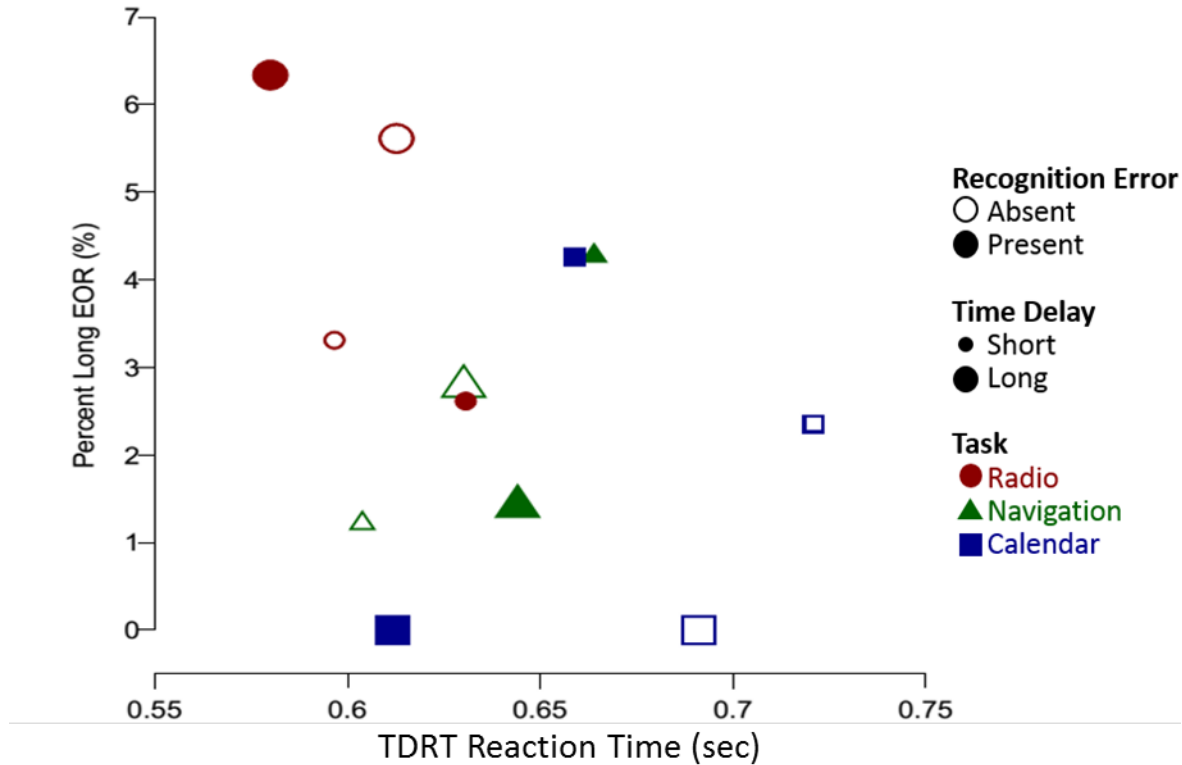


Figure 21. Mean TDRT reaction times and percent long EOR from maximum of trials in Study 2.

Figure 21 compares the TDRT reaction time with percent long EOR. The percent long EOR is plotted using the maximum of all trial within a test condition. The maximum distinguishes the experiment conditions better than the mean (Table 11). The scatterplot shows that the Radio task resulted in faster reaction to the TDRT while Navigation and Calendar had greater variability. Long Delay in the Radio task and Short Delay in the Calendar task led to greater percent EORs. The TDRT reaction time was shorter for the Navigation tasks when compared to the calendar tasks with no recognition error. However, no main or interaction effect was significant for this criteria with the following F-values for the maximum of trials: Task ( $F(2, 88) = 1.74, p = 0.18$ ), Recognition Error ( $F(1, 44) = 0.11, p = 0.75$ ), and Delay ( $F(1, 132) = 0.94, p = 0.33$ ).

### 3.4.5 Criterion 2: Mean glance duration

The acceptance criterion from the visual-manual guidelines is that 85 percent of participants' mean glance durations (MGD) should be less than 2.0 seconds. Table 12 shows that all the tasks conform to criterion 2, considering either the mean or the maximum of trials, because no task induced a MGD that was greater than 2.0 seconds for more than 15 percent of the participants.

Table 12: Conformance with Criterion 2: Mean glance duration

Task	Recognition Error	Delay	Number Participant (out of 24) and % who do not comply			
			Mean		Max	
Radio	Absent	Short	0	0%	1	4%
		Long	0	0%	0	0%
	Present	Short	0	0%	0	0%
		Long	1	4%	1	4%
Navigation	Absent	Short	1	4%	1	4%
		Long	0	0%	1	4%
	Present	Short	1	4%	1	4%
		Long	0	0%	1	4%
Calendar	Absent	Short	0	0%	0	0%
		Long	0	0%	0	0%
	Present	Short	1	4%	1	4%
		Long	0	4%	0	0%

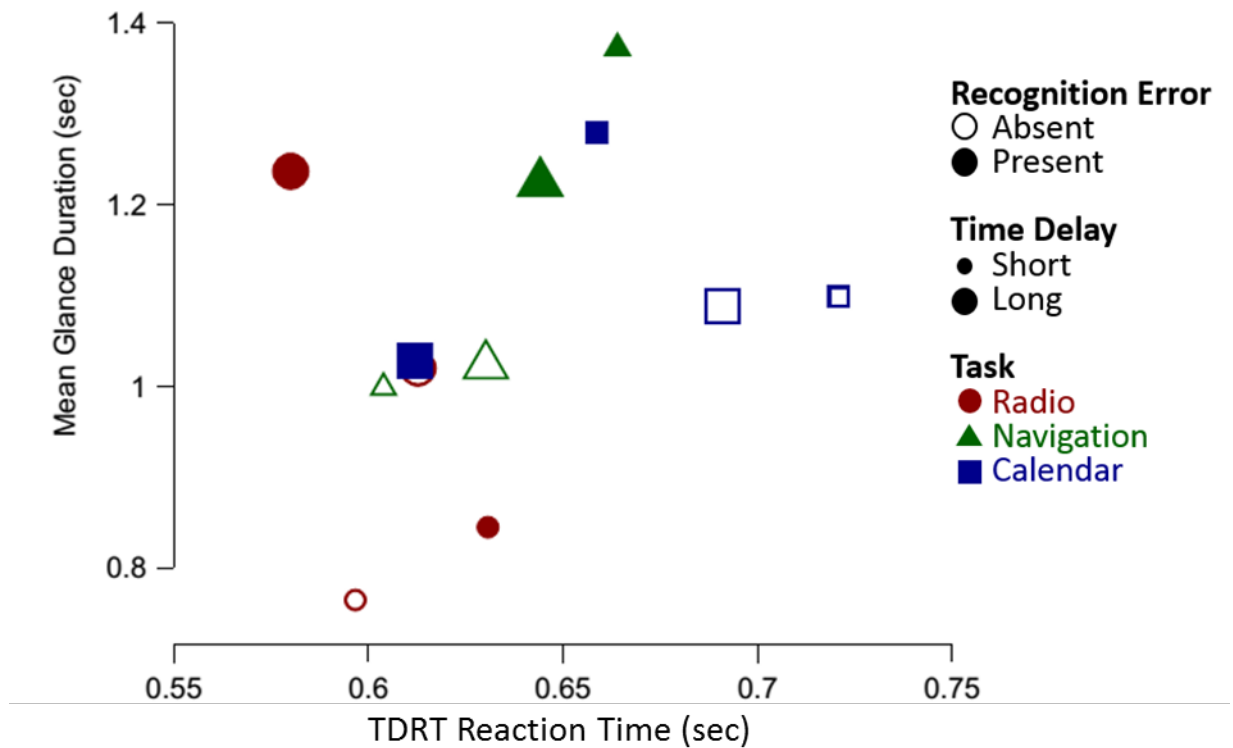


Figure 22. Scatterplot of mean TDRT reaction times and mean glance duration from maximum of trials in Study 2.

Figure 22 shows a scatterplot of the TDRT reaction times plotted against MGD calculated from the maximum of all trial for a test condition. The Recognition Error Present condition induces

longer MGD for the Navigation task than the corresponding Recognition Error Absent Condition. Long Delays in system response for the Radio task resulted in double the MGD than the corresponding Short Delay conditions. However, there was no significant main or interaction effects for the variables of interest for the maximum of trials: Task ( $F(2, 176) = 1.79, p = 0.17$ ), Recognition Error ( $F(1, 44) = 0.18, p = 0.67$ ), and Delay ( $F(1, 44) = 0.02, p = 0.88$ ).

### 3.4.6 Criterion 3: Total eyes-off-road time

The acceptance criterion for the visual-manual guidelines states that for 85 percent of the participants the sum of individual glance durations should be less than or equal to 12 seconds.

Table 13 shows that all experiment conditions conform to criterion 3 when using the mean among trials. All but one condition conforms when using the maximum of trials. The Radio task with Recognition Error Present and long Delay led to summation of glances for that were greater than 12 seconds for 17 percent of the participants. The Navigation Task under similar conditions lead to 13 percent of participants having longer than 12 second TEORT, and just passed the criterion.

Table 13: Conformance with Criterion 3: Total Eyes-Off-Road Time

Task	Recognition Error	Delay	Number Participant (out of 24) and % who do not comply			
			Mean		Max	
			Count	%	Count	%
Radio	Absent	Short	0	0%	0	0%
		Long	0	0%	1	4%
	Present	Short	0	0%	0	0%
		Long	1	4%	4	17%
Navigation	Absent	Short	0	0%	0	0%
		Long	0	0%	1	4%
	Present	Short	2	8%	2	8%
		Long	0	0%	3	13%
Calendar	Absent	Short	0	0%	0	0%
		Long	0	0%	1	4%
	Present	Short	2	8%	2	8%
		Long	1	4%	1	4%



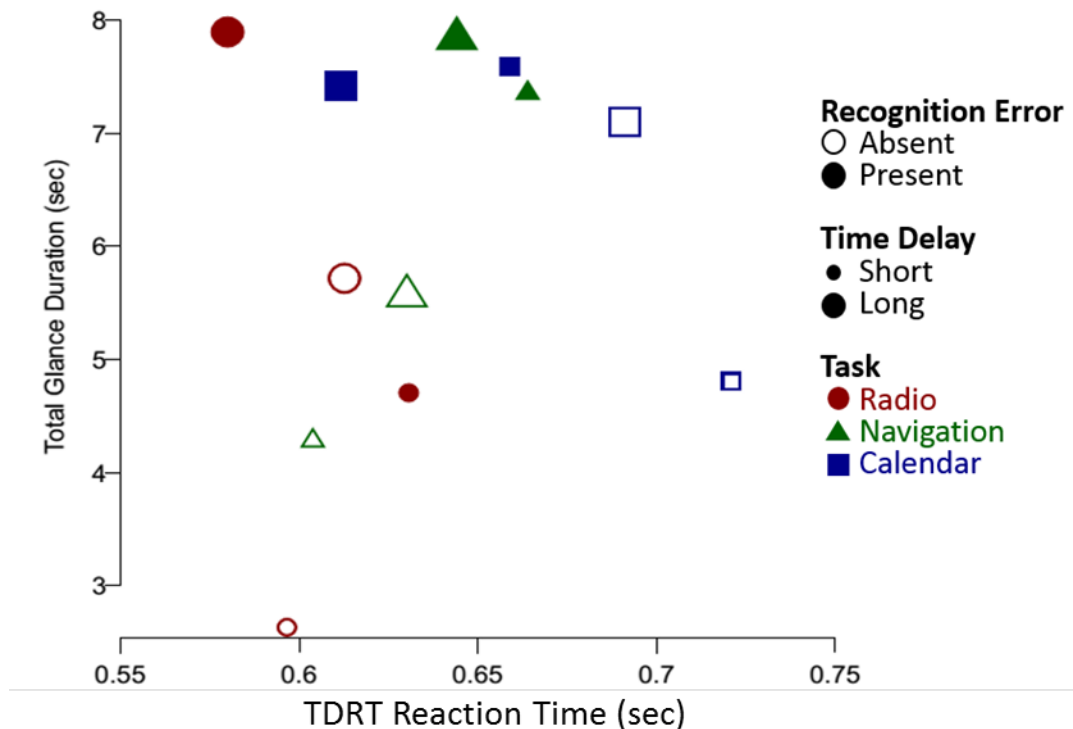


Figure 23. Scatterplot of mean TDRT reaction times and total eyes-off-road time from maximum of trials in Study 2.

Figure 23 shows a scatterplot of TDRT reaction time with TEORT plotted from the maximum values of the trials. With the exception of the Radio task with Short Delay, all tasks with Recognition Error Present resulted in greater TEORT, while there was greater variability in mean TDRT reaction times. Delay had an effect on TEORT ( $F(1, 220) = 11.57, p < 0.001$ ). Almost all tasks in the long Delay condition led to greater TEORT; this trend was not observed only with Calendar Task in the Recognition Error Present condition. A three-way interaction was observed with Task, Delay, and Order ( $F(2,220) = 4.57, p < 0.01$ ), which was mainly driven by the effect of the TDRT order, so that when the TDRT occurred second the TEORT doubled compared to when TDRT occurred first in the case of the Radio Task with long Delay (Figure 37).

### 3.5 DISCUSSION

Total task duration increased with task complexity as did the experimental high levels of recognition error and system delay. The effect of task complexity on duration can be explained in terms of memory recall demands and the number of steps required to complete the task. For example, the radio task required the participant to recall one chunk of information (name of radio station). The navigation task required the participant to recall three chunks of information (street number, street name, street type). The calendar task required the participant to recall four chunks of information (contact name, location, time, and day). The standard deviations for the navigation and calendar tasks were much greater than the radio task indicating more variability between participants in completing the complex tasks. This variability also increased with recognition error and system delay. The combined effects of task complexity, recognition error,

and system delay on both the mean duration and the variability of duration led to long task durations that frequently exceeded 60 seconds. These long task durations are consistent with the observations in Study 1, where people performed typical tasks, with their own systems, in actual driving situations.

The results of Study 2 also indicate that the NHTSA Visual-Manual Guidelines are relevant to VCS evaluation. Similar to the voice control systems observed in Study 1, the VCS used in this study included visual feedback that was redundant to the auditory information provided. Participants looked away from the road to the visual display to confirm that they completed the task correctly and in response to recognition errors and delays. As a consequence, recognition errors resulted in greater mean glance durations (Figure 22) and total eyes-off-road times (Figure 23). Although, the results show VCS can draw drivers' eyes away from the road, all task conditions conformed to the three visual-manual criteria when using the mean of trials.

Conformance with the NHTSA Visual-Manual Guidelines distraction criteria does not ensure that tasks performed with VCS will not distract because VCS tasks might also place a substantial cognitive load on drivers. The tactile detection response Task measures cognitive load associated with the use of VCS—conditions that are cognitively demanding lead to long TDRT reaction times. Reaction times increase with task complexity, suggesting complex tasks were more cognitively demanding than the simple radio channel selection task. Contrary to expectations, the TDRT reaction time increased with presence of recognition error and long system delay. These results suggest that recognition errors and system delays reduce cognitive load. One explanation for this counterintuitive outcome is that the errors and delays introduced an opportunity for a greater degree of self-pacing, which can help modulate cognitive demands. The TDRT miss rates (Figure 20) shows that although participants responded faster to the TDRT stimuli in with more recognition errors and longer system delays, they missed more TDRT stimuli in these conditions.

The TDRT measures suggest that the presence of recognition error or long system delay might reduce cognitive load, but the visual-manual criteria show that corresponding eyes-off-road glances increase. The divergent outcomes of the TDRT and the visual-manual distraction metrics suggest that these metrics measure different dimensions of distraction associated with VCS use, and that both are needed for a comprehensive assessment.

## 4 STUDY 3 – COLLISION DETECTION TASK WHEN USING VCS

Study 2 showed that a driving simulator and the TDRT can be used to assess the visual-manual and cognitive demands of a VCS; however, many developers do not have easy access to a driving simulator. A lower cost evaluation tool would be appealing, but only if it produces valid assessments of the visual-manual and cognitive demands of VCS.

### 4.1 BACKGROUND

The purpose of this study was to assess low cost alternative methods to evaluate VCS, such as a Collision Detection Task. The study was conducted in Madison, WI and follows the same recruitment strategy, voice control tasks, and tactile detection response task as used in Study 2 (the driving simulator study).

Driving simulators are frequently used to understand how VCS interaction might interfere with driving but they can also be costly and time consuming for product evaluation. Surrogate driving tasks allow precise control of cognitive load and precise performance measurement. A summary of some potential surrogate tasks is provided here, followed by a justification for the Collision Detection Task.

- Visual Detection Task-- This task requires participants to respond to a visual stimulus like a LED light reflected off a windshield by pressing a microswitch attached to the index finger.
- Peripheral Detection Task-- This is one of several Visual Detection Tasks, that requires a participant to respond of an LED signal located in the peripheral field of view (Martens & van Winsum, 1999).
- Variation of the Visual Detection Task-- This task uses LED stimuli located in the central view of the participant (Engström & Mårdh, 2007; Victor, Engstrom, Harbluk, 2008; Harbluk, Burns, Hernandez, Tams, Glazduri, 2013).
- Useful Field of View-- The UFoV task measures “the total visual field area in which useful information can be acquired without eye and head movements (i.e., within one eye fixation)” (Ball, Beard, Roenker, Miller, & Griggs, 1988, pg. 2210).
- Enhanced Peripheral Detection Task-- -This task was developed by Hsieh, Young, and Seaman (2012) and the Enhanced Peripheral Detection Task I (EPDT-I) is composed of a single visual event detection task and a video of a real-world driving scene, which makes it simple and easy to run in the lab or on the road (Angell, Young, Hankey, & Dingus, 2002).
- Balloon Analogue Risk Task-- -This task was first introduced by Lejuez, Read, Kahler, Richards, Ramsey, Stuart, and Brown (2002) to assess risky behavior. It requires a certain degree of monitoring and vigilance regarding the status of a balloon and decision making.
- Multiple Objects Tracking-- -This task strives to understand effects of sustained attention on a central field.
- Lane Change Test-- -This task combines the advantages of classic reaction time measures with those of driving simulation, to create a simple, cost-effective, yet reliable and valid method (Mattes & Hallén, 2009).

This study used the CDT (Vaux, Ni, Rizzo, Uc, & Andersen, 2010, Andersen & Kim, 2001) because it presents a more comprehensive and theoretically grounded array of safety-critical driving demands than the other surrogate tasks. The driving demands presented by the CDT include perceiving the current state of the road elements, monitoring for dynamic changes in three dimensions, and making predictions of a near future state of an element, thus assessing the three levels of situation awareness: perception, integration, and prediction. Unlike the DRT that presents a clear signal and demands an immediate response, the CDT requires participants to identify when a response is needed, which is more representative of actual driving situations where threats are not indicated by discrete signals, but emerge out of a field of potential threats. The CDT also requires that participants modulate their attention to the display in a way that mimics the way drivers must divide their gaze between the road and distracting task, but unlike the Balloon Analog Risk Task, the dynamics of the underlying the CDT events are similar to those encountered on the road. More generally, the CDT performance is sensitive to the demands that VCS interaction places on short-term working memory and the demands that more conventional visual-manual tasks place on moderating the duration of glances away from the forward roadway (Young & Angell, 2003). The CDT performance is also linked to driving safety outcomes. Vaux and colleagues (2010) found that poorer performance on the UFOV tasks is associated with poor CDT performance and that both were sensitive to age-related cognitive decline associated with diminish capacity for safe driving in older adults.

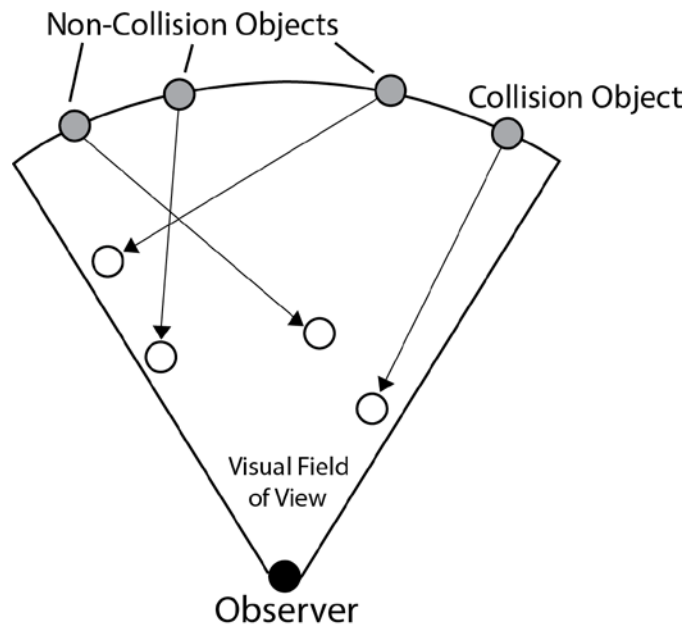


Figure 24. An example of location of the objects and their trajectories (adapted from Vaux et al., 2010)

Figure 24 shows a top-down view of the CDT used in this study. All objects originated on the horizontal eyesight plane at approximately 9 s away from the participant, and then the objects moved towards the participant. Although all objects moved towards the participant, only one was on a collision path, and this object was designated as the target (Figure 24). The participant's task was to detect these targets that were on a collision path and indicate the target object by

touching the object on the screen. If correctly detected (“Hit”), the object faded away from the screen. “False alarms” were those objects detected erroneously by the participant that are not on a collision path. “Misses” were those objects on a collision path with the driver, but were not detected before it “collides” with the participant. An undetected target object approached the participant to the point of collision, with the object filling the screen completely at the point of collision.

For this study, there were four moving objects on the screen at any given time. Non-collision objects did not collide with the participant and merely continued along a path away from the participant. Objects were replaced so that at any time four objects were on the screen. The target objects were generated at intervals of 5 to 9 seconds from the previous one. The object velocity of the target objects was selected based on a pilot study.

The scene had a dimension of 2,000 x 1,000 units (a unit is approximately one eye height or ~1.5m). The roadway extended four units horizontally and extended directly in front of the driver to the length of the simulated space, 2,000 units. The objects were spheres of one unit radius (adapted from Vaux et al., 2010). The spheres were shaded using a Gouraud shading model so that the shape was easily discerned (Figure 25). The angular size of the object varied with distance to mimic the effect of distance on the visual angle subtended by objects in a real driving scene.

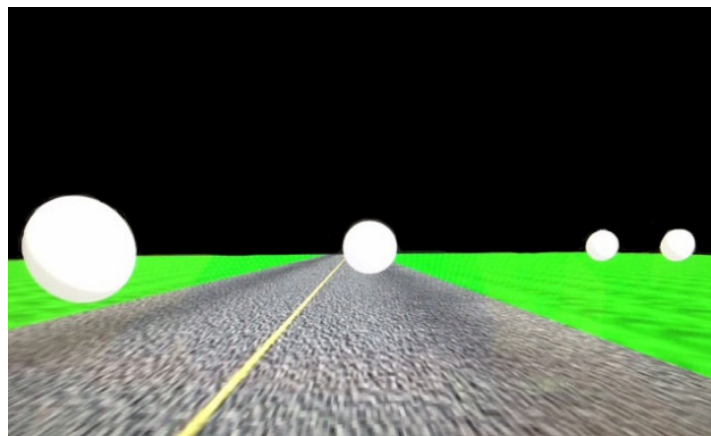


Figure 25. Screenshot of the CDT

## **4.2 CALIBRATION STUDY TO SELECT COLLISION DETECTION TASK PARAMETERS**

The CDT has many parameters that could be tuned to enhance its sensitivity in estimating the distraction potential of non-driving tasks: extent of the scene, size, shape, and number of objects, speed of the participant, speed of the objects, frequency of target objects, and the exposure periods of the target and non-target objects.

Object velocity has been shown to affect the reaction time and sensitivity significantly (Andersen & Kim, 2001). A pilot study at Wisconsin considered three levels of object velocity of 20 mph, 40mph, and 60 mph corresponding to speeds in residential, city, and highway areas. The velocity of the spheres was the same as the velocity of the driver. The n-back task was used to assess the

sensitivity of the CDT to distraction. No task (baseline), 0-back task, and 1-back tasks defined the three levels of cognitive distraction.

Object velocity and the n-back task were combined in a 3 by 3 within subject experimental design. The conditions were presented in a blocked fashion (n-back was blocked) to create 9 experimental blocks for each participant, and 6 trials in each block. Three blocks of practice trials were administered for each participant with the varying levels of n-back and speed. The total experiment time per participant was about 1.5 hours, including initial consenting and training processes. Eighteen participants took part in the pilot study and were students from the University of Wisconsin-Madison. Participants ranged from 18 to 30 years old, with a mean age of 22. Compensation was at the rate of \$10 per hour. Performance on the CDT was defined by detection performance and reaction time.

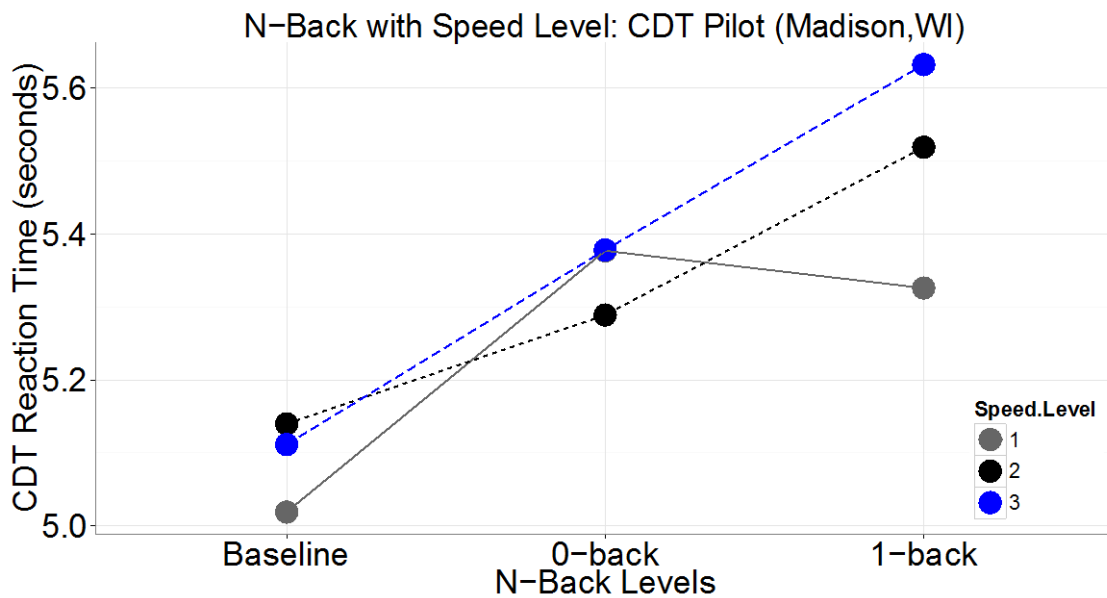


Figure 26. Reaction time to detecting targets in the CDT across three levels of speed and two levels of the n-back and a baseline (only CDT). Speed levels 1, 2 and 3 are 20mph, 40mph, and 60 mph respectively.

As expected, increasingly difficult levels of the n-back task led to longer reaction time with the CDT. The speed that was most sensitive to distraction associated with the n-back task was adopted for the main study. As shown in Figure 26, the speed level of 60 mph was the most discerning condition with the greatest reaction time for the 1-back task, and this speed was selected for the main study.

### 4.3 METHOD

The CDT study used the same experimental design as Study 2: It assessed task complexity, recognition accuracy, and system delay. That is, the CDT study included the same three VCS tasks with same levels of recognition error and delay as in Study 2. With respect to the order of presentation, the same two orders were used, but the CDT was used instead of the simulator task (Table 14). The CDT study also used the same 1-back task, TDRT apparatus, age groups, and overall experimental protocol.

Table 14: Experimental Run Order

<b>ORDER</b>		
<b>Run</b>	<b>Order 1: TDRT second</b>	<b>Order 2: TDRT first</b>
1	CDT only	TDRT only
2	CDT + TDRT + Voice Tasks	TDRT + Voice Tasks
3	CDT only	TDRT only
	<b>Break</b>	<b>Break</b>
4	TDRT only	CDT only
5	TDRT + Voice Tasks	CDT + TDRT + Voice Tasks
6	TDRT only	CDT only

## 4.4 RESULTS

### 4.4.1 Demographics

The mean age was 38.90 ( $SD=15.54$ ) for females and 39.06 ( $SD=14.57$ ) for males. The ages ranged from 19 to 69 and Table 15 shows the mean for each age group.

Table 15: Demographics of Participants in Madison Study

<b>Age Group</b>	<b>Mean</b>	<b>Std. Dev.</b>
18-24	22.67	1.82
25-39	28.91	3.85
40-54	46.90	3.81
55-75	61.25	4.85

### 4.4.2 VCS task performance: Total task duration

Similar to Study 2, total task duration was defined as the mean of six trials for the Radio Task and mean of three trials each for the Navigation and Calendar Tasks. Figure 27 shows the distributions of the task durations across the participants.

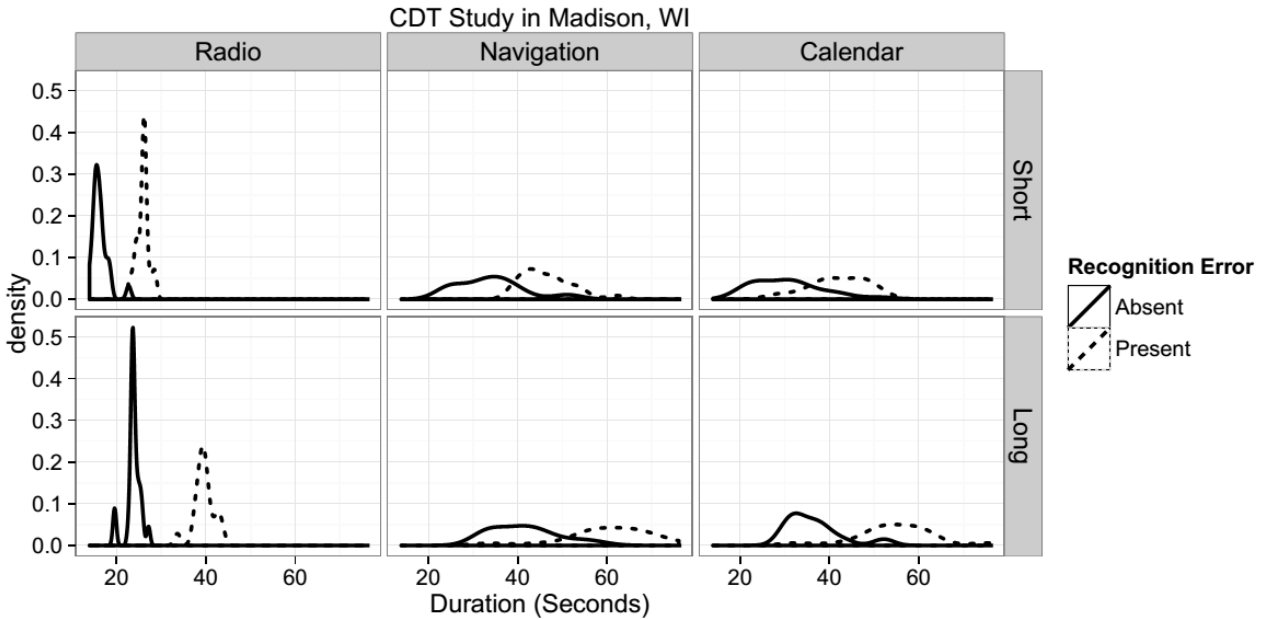


Figure 27. Total task duration for Study 3

The independent variables of Task ( $F(2, 81) = 221.82, p < 0.001$ ), Recognition Error ( $F(1, 40) = 200.92, p < 0.001$ ) and Delay ( $F(1, 40) = 243.95, p < 0.001$ ) all increased Total Task Duration substantially and in a fashion very similar to Study 2. Beyond the main effects, two-way interactions (Figure 38) was found between Recognition Error x Delay ( $F(1, 40) = 23.88, p < 0.001$ ) and Order x Delay ( $F(1, 40) = 7.07, p = 0.01$ ). Long Delay coupled with Recognition Error Present conditions resulted in longer task durations. An interaction between Recognition Error and Delay was also observed in Study 2. Three-way interaction between Recognition Error x Task x Order ( $F(2, 81) = 4.02, p = 0.02$ ) was observed, which reflects the longer task duration in the TDRT First order for the Calendar Task with Recognition Error Present (Figure 39). A four-way interaction between Recognition Error x Task x Order x Delay ( $F(2, 79) = 3.88, p = 0.02$ ) was observed that was the result of the more complex tasks of Navigation and Calendar (Figure 40). Table 16 summarizes the mean, standard deviation and range of the total task duration for Study 3. Overall, the task durations observed in Study 3 closely match those observed in Study 2—the difference between the overall mean task duration for the two studies is only 0.16 seconds. Figure 28 shows the correspondence between the two studies, with perfect correspondence indicated by the diagonal line.



Table 16: Summary of Total Task duration (sec) for Study 3

Task	Recognition Error	Delay	Mean	Standard Deviation	Range	
					Minimum	Maximum
Radio	Absent	Short	16.16	1.77	13.92	22.70
		Long	23.72	1.54	19.46	27.13
	Present	Short	25.85	1.30	22.95	28.74
		Long	39.68	2.16	33.59	43.71
Navigation	Absent	Short	34.66	8.34	23.89	55.03
		Long	41.76	7.70	32.51	58.15
	Present	Short	47.09	5.75	39.22	62.03
		Long	61.24	10.19	33.53	81.55
Calendar	Absent	Short	31.12	8.07	21.16	51.29
		Long	36.85	6.84	29.55	52.33
	Present	Short	42.01	6.06	27.94	51.50
		Long	55.08	9.04	32.27	75.11

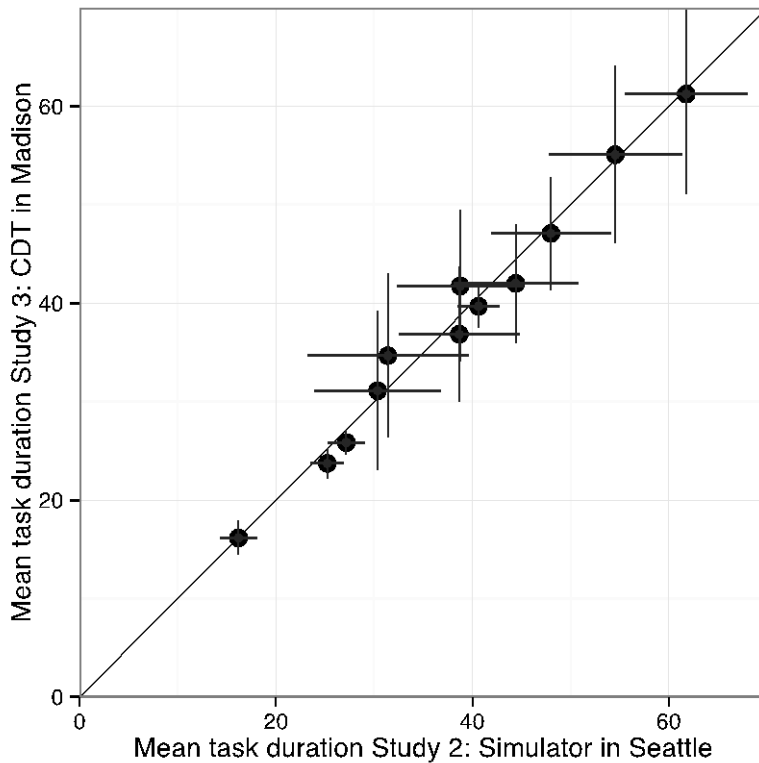


Figure 28. Mean task duration and standard deviation for Study 2 and Study 3.

### 4.4.3 TDRT performance measures

ISO draft standards for the TDRT were followed and the reaction time was calculated for responses between 0.1 and 2.5 seconds. Failures to respond or responses greater than 2.5 seconds were counted as a miss. Responses less than 0.1 second after the trigger were considered invalid and discarded.

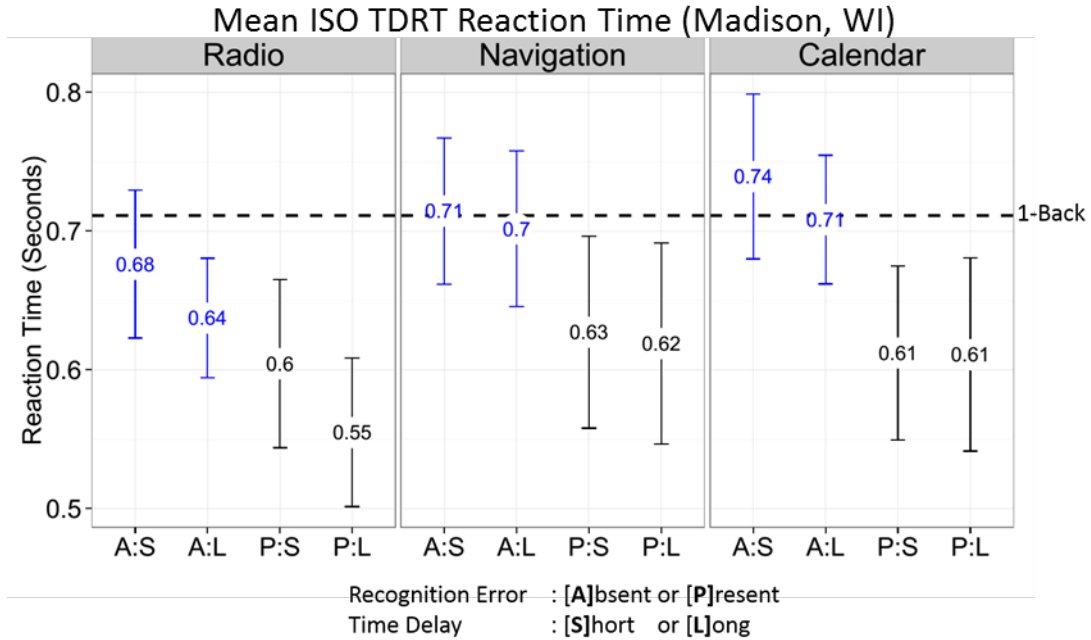


Figure 29. Mean TDRT Reaction time (dashed line is 1-back mean)

Figure 29 shows Recognition Error and Delay plotted against the mean TDRT Reaction time for each task across the three tasks. Similar to Study 2, the reaction times to triggers during the 1-back task is averaged across all participants (mean = 0.71 sec) and indicated as a baseline comparison. Task type resulted in differences in reaction time to the TDRT ( $F(2, 80) = 4.81, p = 0.01$ ). In all conditions, the reaction time was similar to, or lower than the 1-back task.

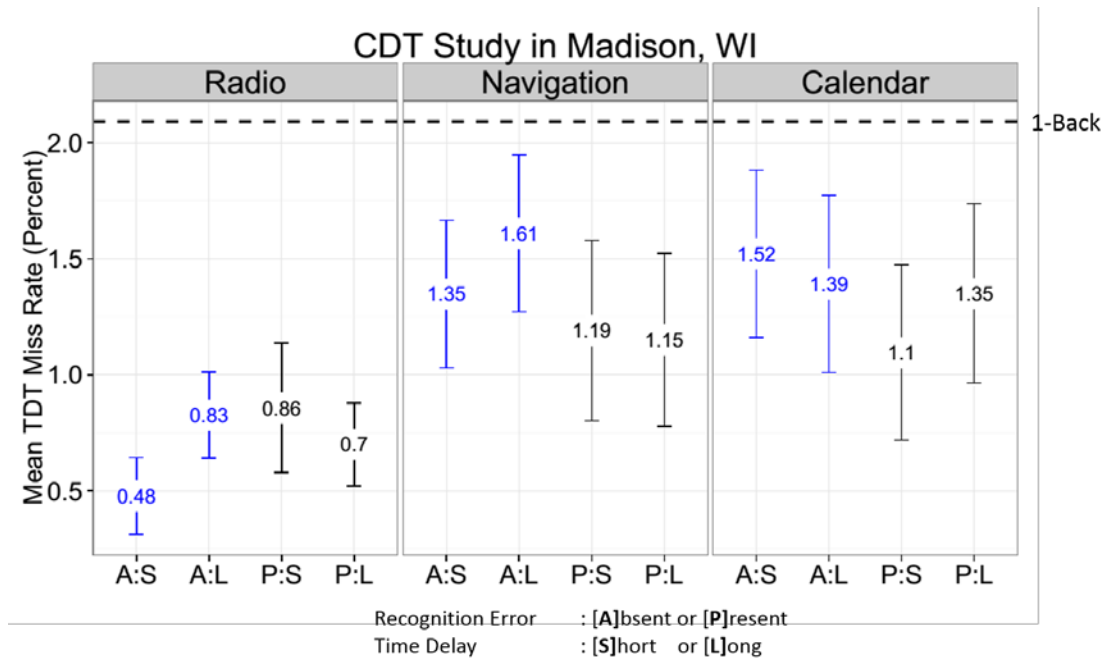


Figure 30. TDRT miss rate

Figure 30 shows the TDRT miss rates and similar to Study 2, the miss rates were well within the 30 percent threshold suggested by the ISO draft standards. In addition, all the task conditions resulted in miss rates less than the 1-back task. The Radio task induced fewer misses than the Navigation or Calendar tasks. Both the absolute value and the general pattern of results for the TDRT reaction time and miss rates were similar to those observed in Study 2.

#### 4.4.4 Criterion 1: Percentage of longlong EOR

Similar to Study 2, the mean and the maximum of the trials were calculated for the three tasks and the recognition error and system delay conditions. Table 17 shows that almost all conditions conform to criterion 1, where the task conforms if less than 15 percent of the participants have percent long EOR less than 2.0 seconds for the task conditions. Only the Navigation Task with Recognition Error Present and Short Delay fails to conform to the criterion for both mean and maximum values. The Calendar task under Recognition Error Present and long Delay condition also fails to conform when the maximum values of the trials is used.

Table 17: Conformance with Criterion 1: Percentage of long EOR glances

Task	Recognition Error	Delay	Number Participant (out of 24) and % who do not comply			
			Mean		Max	
Radio	Absent	Short	0	0%	0	0%
		Long	0	0%	0	0%
	Present	Short	0	0%	2	8%
		Long	0	0%	2	8%
Navigation	Absent	Short	0	0%	0	0%
		Long	0	0%	1	4%
	Present	Short	5	21%	9	38%
		Long	0	0%	1	4%
Calendar	Absent	Short	0	0%	0	0%
		Long	0	0%	0	0%
	Present	Short	1	4%	1	4%
		Long	1	4%	4	17%

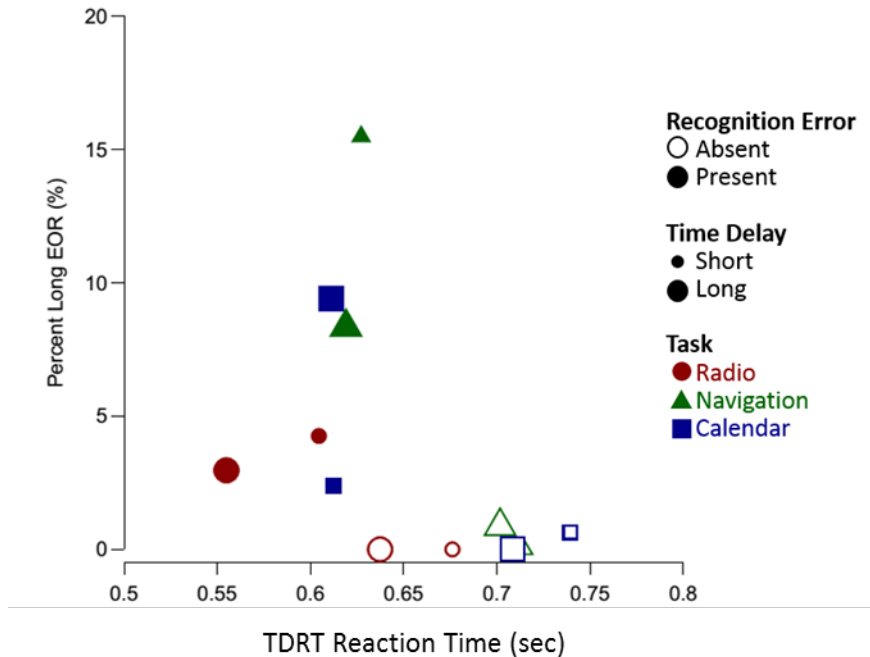


Figure 31. Mean TDRT reaction times and percent long EOR from maximum of trials in Study 3

Figure 31 compares the TDRT reaction time with percentage of long EOR. The percentage of long EOR is plotted for the maximum percentage of long EOR of trials, because using the maximum accounts for the extreme glance behaviors and also distinguishes the conditions better than the mean (Table 17). The scatterplot indicates that the Recognition Error Present condition results in shorter reaction times to the TDRT. However, these conditions also resulted in larger percent of EOR glances. On the other hand, all tasks under the Recognition Error Absent

condition resulted in moderate to slow responses to the TDRT, but very few long glances. Considering the maximum values among trials, Task ( $F(2, 159) = 2.86, p = 0.06$ ) as well as Delay ( $F(1, 80) = 0.06, p = 0.81$ ) failed to affect TEORT at a statistically significant level. Hence, only the Recognition Error affected criterion 1 ( $F(1, 80) = 14.84, p < 0.001$ ).

#### 4.4.5 Criterion 2: Mean glance duration

Table 18 shows that all but one task conformed to criterion 2 for both mean and maximum of trials. The Navigation task under Recognition Error Present and Short Delay condition failed to conform to the criterion, when the maximum among the trials is considered.

Table 18: Conformance with Criterion 2: Mean glance duration

Task	Recognition Error	Delay	Number Participant (out of 24) and % who do not comply			
			Mean		Max	
			Count	%	Count	%
Radio	Absent	Short	0	0%	0	0%
		Long	0	0%	0	0%
	Present	Short	0	0%	0	0%
		Long	0	0%	0	0%
Navigation	Absent	Short	0	0%	0	0%
		Long	0	0%	0	0%
	Present	Short	0	0%	4	17%
		Long	0	0%	0	0%
Calendar	Absent	Short	0	0%	0	0%
		Long	0	0%	0	0%
	Present	Short	0	0%	0	0%
		Long	0	0%	1	4%

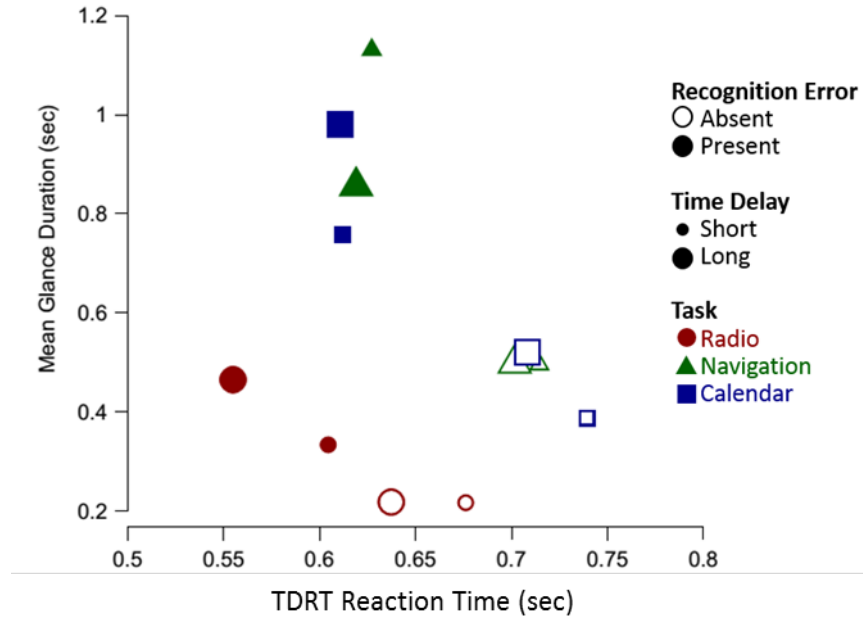


Figure 32. TDRT vs (max) mean glance duration

Figure 32 compares the mean TDRT reaction time to the MGD. Here, Recognition Error Present conditions led to larger mean glance durations than corresponding Recognition Error Absent conditions. In addition, as with criterion 1, the conditions with Recognition Error Present resulted in shorter TDRT reaction times. Another notable result is that the Radio task results in much shorter MGD than Navigation or Calendar Tasks. Hence Recognition Error ( $F(1, 40) = 9.86, p < 0.01$ ), Task ( $F(2, 76) = 40.02, p < 0.001$ ), and an interaction between them ( $F(2, 70) = 6.75, p < 0.01$ ) affect MGD (Figure 41, Figure 42). A two-way interaction between Task x Delay ( $F(2, 71) = 6.75, p < 0.01$ ) is observed, such that long Delay for the Calendar task increases MGD. Overall, Figure 32 shows that TDRT reaction time is shorter with the recognition error, but mean glance duration is longer.

#### 4.4.6 Criterion 3: Total eyes-off-road time

Table 19 shows that all tasks conform with criterion 3 and the total eyes-off-road time was shorter than 12.0 seconds for all participants across all test conditions for both the mean and maximum of trials.

Table 19: Conformance with Criterion 3: Total eyes-off-road time

Task	Recognition Error	Delay	Number Participant (out of 24) and % who do not comply			
			Mean		Max	
			Count	%	Count	%
Radio	Absent	Short	0	0%	0	0%
		Long	0	0%	0	0%
	Present	Short	0	0%	0	0%
		Long	0	0%	0	0%
Navigation	Absent	Short	0	0%	0	0%
		Long	0	0%	0	0%
	Present	Short	0	0%	0	0%
		Long	0	0%	0	0%
Calendar	Absent	Short	0	0%	0	0%
		Long	0	0%	0	0%
	Present	Short	0	0%	0	0%
		Long	0	0%	1	4%

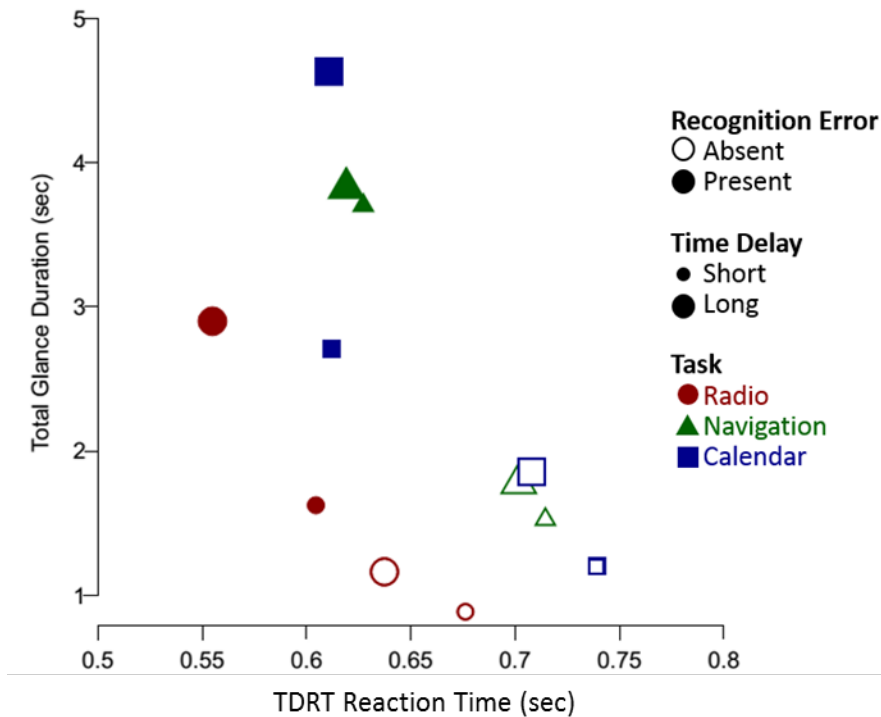


Figure 33. TDRT versus (max) total eyes-off-road time

Consistent with results for criteria 1 and 2, Recognition Error Present leads to greater TEORT and shorter TDRT reaction time (Figure 33). Short Delays result in less TEORT for the corresponding task and error conditions, however delays also result in slower responses to the TDRT. The more complex tasks of Navigation and Calendar result in longer TEORT than the Radio Task, as well as slower responses to TDRT. Hence Task ( $F(2, 158) = 14.01, p < 0.001$ ) Recognition Error ( $F(1, 40) = 12.96, p < 0.001$ ), and Delay ( $F(1, 40) = 11.35, p < 0.1$ ) increase

TEORT. As with the other criteria, Figure 33 shows that TDRT reaction time is shorter with the recognition error, but mean glance duration is longer.

## 4.5 DISCUSSION

Similar to Study 2, task duration increased with task complexity, recognition error, and system delay. As with Study 2, greater variability was observed for the more complex tasks of Navigation and Calendar. Task duration across all conditions corresponded very closely to that observed in Study 2, suggesting that the CDT placed demands on drivers that were generally similar to those of a driving simulator and that the CDT might be a useful surrogate driving task.

The TDRT reaction time was shorter with recognition errors and with longer system delays, which is consistent with Study 2 results. The TDRT miss rates (Figure 30) did not differentiate between recognition errors and system delays, however the more complex tasks resulted in greater miss rates than the Radio task. The consistent pattern of results between Study 2 and Study 3 further demonstrates that the CDT might be a useful surrogate driving task.

Considering the results in terms of the visual-manual guidelines, demonstrated that VCS can produce substantial visual demands, particularly when recognition errors occur. Presence of recognition errors led to a greater percent long EOR, mean glance durations, and TEORT for all tasks. Most tasks conformed to the visual-manual criteria, but several conditions from the navigation and calendar tasks did not conform to criteria 1 and 2. In Study 2, the radio channel selection task failed to conform to criteria 1 and 3. The general tendency of conformance with the visual-manual criteria in Study 1 parallels the results from Study 2.

The patterns of results obtained with the CDT protocol are generally similar to those obtained with the driving simulator protocol, but the CDT had greater sensitivity in differentiating between conditions. One reason for this is that it includes active and latent hazards that the driver must respond to that are absent in the simulator protocol used in Study 2. Because of this, the CDT produces greater attentional demands compared to the simulator protocol that involves car following on a straight road without any traffic, hazards or other events that demand attention. This difference is reflected in the glance behavior of the participants. In Study 2 the mean total off-road glance duration ranged from 2.8 to 8.0 seconds across the experimental conditions compared a range of 1.0 to 4.8 seconds in Study 3. Participants in Study 3 looked away from the road less, had shorter mean glance durations, and had less eyes-off-road time compared to those in Study 2.



## 5 GENERAL DISCUSSION

The results of the three studies are considered together in addressing the main questions that motivated this study.

### **1. Are commonly experienced VCS usability issues such as task complexity, interaction errors, and response delays potentially distracting for drivers?**

*Task Complexity:* The varying levels of task complexity impose different demands. In Study 1, Entertainment tasks (which included radio tuning) had shorter total task durations and smaller error rates than navigation tasks. Navigation tasks were the most frequently demonstrated and produced one of the highest overall error rates (Table 5). Similar results were observed in Studies 2 and 3. The navigation task resulted in the longest task duration, and the radio task had the shortest. Particularly for Study 3, the findings showed that the navigation task led to larger TDRT reaction times and greater number of glances away from the road than the radio task.

Increasing task complexity can lead to longer task durations and increases in cognitive load.. Reducing task complexity is important even with VCS because increased complexity is associated with increased cognitive load, as measured by TDRT reaction time, and because more steps make recognition errors more likely. If each step of an interaction has a 10 percent chance of a recognition error then a seven-step interaction has more than a 50 percent chance of including at least one recognition error.

*Recognition Accuracy:* The recognition accuracy of the VCS may inform the amount of visual feedback to provide. For Studies 2 and 3, all but one task conformed to the visual-manual criteria when there were no recognition errors, indicating that participants did not heavily depend on the visual information to confirm that they completed the task correctly. However, as indicated by Study 1 errors occurred in almost 50 percent of VCS interactions. Of these interactions, Clarification and Wrong Task errors were the most observed. These errors were defined as VCS “misrecognition” and resulted in either requiring users to restate their last command or the VCS executing the wrong task. In Study 2 and 3, the recognition errors were most similar to the Wrong Task errors identified in Study 1. Participants exposed to the recognition errors in both Study 2 and 3, had longer mean glance duration and more TEORT even though these tasks were designed such that no visual-manual interaction was needed. While the visual display caused more eyes-off-road time, it also seemed to provide a benefit in that glances to the display resolved uncertainty regarding systems state, leading to lower TDRT reaction time. Visual feedback may help to alleviate cognitive load when drivers experience recognition errors.

*Design Consideration:* Judicious use of visual feedback might reduce the negative effects of that recognition errors can have on cognitive load and system interaction. The benefit of such visual feedback needs to be carefully balanced against the cost of off-road glances.

*System Delay and Pacing:* Frequent system timeouts were observed in Study 1 where users were not able to give commands in the time allotted by their VCS. Users were also frustrated that they could not interrupt long prompts or messages. The laboratory task enforced no delay or an eight-second delay issuing the confirmation message once the participant gave a correct command. Studies 2 and 3 found that the delay reduced cognitive load such that mean TDRT reaction times were generally faster when participants experienced the eight-second delay. This suggests that

incorporating a longer delay in the VCS design enables better self-pacing. Unlike most visual-manual interactions, VCS actively prompt and demand responses from drivers. If drivers have no means of pacing the interaction they might not be able to adapt their interaction to devote more attention to the road when the roadway demands increase. This is particularly true for complex tasks that extend for 30 seconds to a minute. The driver might initiate these long tasks during a low-demand situation, but the driving situation might evolve into a high demand situation by the end of the task.

*Design Consideration:* The findings from the three studies suggest a possible need for more user-paced interaction. For example, mechanisms for pacing the interaction might include VCS interactions that are adaptive based on the driver's previous behavior, road situation, or in response to commands such as "Pause".

## **2. Are the NHTSA Phase I visual-manual criteria relevant to VCS?**

Even with VCS that do not require visual-manual interactions, drivers look away from the road. The findings of Study 2 and 3 showed that different VCS characteristics affect the number of glances toward the in-vehicle display. Even though no visual-manual interactions were required, the participants still looked away from the road to confirm that they completed the task correctly or in response to recognition errors. Study 1 found that many voice-based systems currently on the market also contain visual information that drivers consult to confirm their commands. The laboratory results, in combination with the contextual interview study, clearly demonstrate that drivers are likely to look away from the road during voice-based tasks, suggesting that the visual-manual guidelines apply to voice-based systems.

When considering the mean of trials for Studies 2 and 3, only one condition did not conform to the NHTSA Visual-Manual Guidelines (Navigation/Recognition Error Present/Short Delay). Compared to text reading and text entry tasks, no text entry tasks regardless of character length conformed with criteria 1 and 2, and only short and medium text entry tasks (4 and 6 character words respectively) conformed to criterion 3 (Boyle et. al, 2012). These results may suggest that VCS interaction imposes much less visual demand on drivers; however, voice interactions, particularly those involving recognition error, can draw drivers' eyes off the road and impose a cognitive load that is not reflected in the glance behavior. Conforming with the visual-manual guidelines therefore represents a necessary, but not a sufficient condition in assessing voice-based systems.

## **3. As part of a VCS evaluation protocol, is the TDRT sensitive to task complexity, system delay, and recognition errors?**

Different types of tasks with varying complexity have a strong effect on task duration, glance behavior, and cognitive demand, as measured by TDRT reaction time. Recognition accuracy and system delay increased visual demand and reduced cognitive demand. These results have important implications for evaluating voice-based systems, suggesting that TDRT measures of cognitive demand are a useful complement to measures of visual demand. TDRT is sensitive to different features of VCS compared to glance behavior. This sensitivity is evident even though the duration of each task performance was considerably less than the three or one minute exposure investigated by Ranney et. al (2014).

#### **4. Is a low-cost evaluation method for VCS, such as the Collision Detection Task (CDT), a reasonable method for assessing VCS?**

The results suggest that the CDT is a promising surrogate for driving, particularly for examining the combined cognitive and visual-manual demands of VCS. The CDT was created to replicate some of the demands associated with driving that the cognitive demands of VCS interaction might interfere with. The VCS requires consistent attention of drivers to address the intermittent demands of active and latent hazards. This is particularly critical for voice task evaluation because others have found that drivers' gazes tend to focus on the center of the road (Engstrom et al., 2005) and pay less attention to peripheral areas (Harms et al., 2003) when engaged in VCS interaction. The CDT involves more visual scanning than the simulator protocol, which only requires drivers to maintain lane position and constant speed. Although similar results were obtained in the simulator and CDT protocol, the CDT had greater sensitivity in differentiating between conditions. Even though trends were observed in Study 2 (simulator), the only statistically significant outcome was observed for Criterion 3, whereas there were significant differences observed for all three criteria in Study 3. Similar to the results of Ranney et al. (2014), who compared outcomes from a driving simulator to a non-driving situation, these results suggest relatively consistent outcomes for the TDRT across various data collection venues.

#### **5. What are the implications of the present findings for developing a protocol for assessing VCS?**

The findings of this study have broader considerations for the distraction potential of vehicle systems. The three studies described in this report considered the distraction potential of VCS from very different perspectives. Each perspective offers different insights and imposes different demands. The contextual interviews used in Study 1 reveal how drivers actually use products that are in the marketplace and indicate potential challenges that might be addressed by further refinement of VCS. Other advantages of contextual inquiry methods are that they require little specialized equipment, and that they are applicable to systems that have already been deployed in production vehicles. Study 2 and Study 3 replicate and extend a summative evaluation protocol based on the visual-manual guidelines. This summative evaluation represents a final check to assess the distraction potential of a vehicle system. Studies 2 and 3 require time consuming human subjects recruiting and testing. Study 2 also requires costly equipment. Because of this only a small subset of the full range of "testable" tasks can be evaluated. Together these three approaches provide complementary means of detecting distraction potential, but they may not be sufficient.

Evaluation protocols such as those used in Studies 1, 2 and 3 could be considered as part of a broader system evaluation, with Study 1 corresponding to surveillance of products in the marketplace and Studies 2 and 3 corresponding to summative testing. The contextual interviews in Study 1 identified the importance of recognition error and systems delay as considerations in the laboratory evaluations in Study 2 and 3.

#### **TDRT design and metrics**

The unexpectedly low cognitive demand associated with VCS errors suggests that a more refined metric is needed to capture momentary increases in cognitive demand. Currently TDRT data are

averaged over the entire task, but response to errors might induce a short, transient demand that gets obscured when averaged over the period of the entire task. Beyond a more refined analysis of the data, a different structure of the TDRT might provide an index of cognitive load that is more closely related to potential safety consequences. Because the TDRT repeats so frequently drivers may develop a response set that does not represent the process associated with detecting and responding to events on the road. Increasing the uncertainty associated with the TDRT stimuli by increasing the mean and standard deviation of the interstimulus interval might make the resulting reactions times much more sensitive to the distraction potential of VCS. Because the TDRT stimuli occur randomly they might not occur during challenging aspects of the task, such as when errors occur. This deficiency could be addressed by linking the onset of the stimuli to specific aspects of the interaction, such as the onset of VCS errors.

### **Calibration of CDT and TDRT with brake reaction time**

Currently, measures of cognitive demand are not closely linked to safety-relevant vehicle control measures, such as brake reaction time. This missing link makes it difficult to interpret the potential safety consequence of a 500 ms longer or shorter TDRT reaction time. To address this issue, drivers could be exposed to a range of tasks on a test track and then brake reaction times could be measured. The TDRT reaction time for the same tasks could then be recorded in the simulator or in with CDT. These data could support a calibration process that could relate increases in TDRT reaction times to brake response times. This would provide a first step towards relating measures of cognitive load to safety-related driving outcomes.

### **Contextual interviews to identify use patterns with actual systems**

As demonstrated in this study, contextual interviews provide an invaluable window into how people actually engage VCS while driving. Such studies identify types of tasks attempted, types of errors, failure modes, points of frustration, users' mental models, and user acceptance of current systems. Findings can then be used to ensure real world relevance in the design of controlled laboratory studies. The value of such contextual interviews will increase as the variety of devices proliferates.

### **Comprehensive assessment of task duration and selection of representative tasks**

Contextual interviews, simulator, or CDT investigations offer a window into a small subset of the many possible tasks a driver might perform with a vehicle information system. Selecting this small subset of "representative tasks" from the many possible tasks represents an important challenge that might be addressed with a model-based approach. Such an approach could map the menu or network structure of the vehicle information system. This map of the system indicates the number of steps and associated time required to traverse the menu structure of the system, such as the time from the home screen to the final step of selecting a radio station. The distribution of time estimates for every possible task can then identify tasks that might be dangerously long. This distribution can also justify a selection of tasks for more evaluation in the laboratory. The complementary strengths of modeling, contextual interviews, and laboratory evaluation suggest the need for a broad evaluation protocol that extends beyond a simple summative evaluation.

## 6 CONCLUSIONS

The findings suggest that all VCS tasks as studied in this project conformed to Phase I Visual-Manual guidelines, demonstrating the substantial benefit of VCS relative to visual-manual interaction. Interaction errors and system delays with VCS are common and a normal part of current users' experience. Hence, it is important to consider these challenges in evaluating systems. In both the on-road and laboratory studies, VCS users often look away from the forward roadway during user-system interaction errors. In fact, the typical interactions with multimodal VCS often include looking at a visual display and require manual inputs. Hence, the criteria based on the NHTSA Visual-Manual Distraction Guidelines are appropriate for evaluation of VCS, but are not sufficient given the cognitive demands of voice interaction. The studies showed that increasing VCS error rate or increasing system delays is associated with increased glances away from the forward roadway, and decreased TDRT reaction time, which is a measure of cognitive load. In other words, glance measures and TDRT appear to assess different aspects of distraction. The CDT protocol appears to be a viable assessment method for driver distraction, yielding results similar to the NHTSA driving simulator protocol. Findings of the contextual enquiry and laboratory evaluations were complementary and suggest that, along with other measures, task duration is an evaluation metric that may be particularly well suited to VCS assessment.

## REFERENCES

- Andersen, G. J., & Kim, R. D. (2001). Perceptual information and attentional constraints in visual search of collision events. *Journal of experimental psychology. Human perception and performance*, 27(5), 1039–56.
- Angell, L. S., Young, R. A., Hankey, J. M., & Dingus, T. A. (2002). *An evaluation of alternative methods for assessing driver workload in the early development of in-vehicle information systems*. Warrendale, PA: Society of Automotive Engineers.
- Apple, Inc. (2013). Apple iPhone 5 Siri, [www.apple.com/ios/siri/](http://www.apple.com/ios/siri/) Accessed on Jan 20, 2013.
- Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision*, 7(13), 1–10. doi:10.1167/7.13.14.Introduction
- Ball, K. K., Beard, B. L., Roenker, D. L., Miller, R. L., & Griggs, D. S. (1988). Age and visual search: expanding the useful field of view. *Journal of the Optical Society of America. A, Optics and image science*, 5(12), 2210–9.
- Ball, K., Owsley, C., Sloane, M. E., Roenker, D. L., & Bruni, J. R. (1993). Visual attention problems as a predictor of vehicle crashes in older drivers. *Investigative ophthalmology & visual science*, 34(11), 3110–23.
- Barón, A., & Green, P. (2006). *Safety and usability of speech interfaces for in-vehicle tasks while driving: A brief literature review*. Ann Arbor: University of Michigan Transportation Research Institute.
- Bengler, K., Kohlmann, M., & Lange, C. (2012). Assessment of cognitive workload of in-vehicle systems using a visual peripheral and tactile detection task setting. *Work: A Journal of Prevention, Assessment and Rehabilitation*, 41(0).
- Beyer, H. & Holtzblatt, K. (1997). *Contextual Design: Defining customer-centered systems*. San Francisco: Morgan Kaufmann Publishers Inc.
- Blanco, M., Biever, W. J., Gallagher, J. P., & Dingus, T. A. (2006, September). The impact of secondary task cognitive processing demand on driving performance. *Accident Analysis & Prevention*, 38(5), 895-906.
- Boril, H., Boyraz, P., & Hansen, J. (2012). Towards Multimodal Driver's Stress Detection. In J. Hansen, P. Boyraz, K. Takeda, & H. Abut, *Digital Signal Processing for In-Vehicle Systems and Safety*. Heidelberg: Springer.
- Boyle, L. N., Lee, J. D., Peng, Y., Ghazizadeh, M., Wu, Y., Miller, E., & Jenness, J. (2013). *Text reading and text input assessment in support of the NHTSA visual-manual driver distraction guidelines* (Report No. DOT HS 811 820). Washington, DC: National Highway Traffic Safety Administration.
- BusinessWire (2013). Conexant Introduces New Far-Field Voice Input Processor SoC for Smart TVs Accessed on Jan 25, 2013, at [www.businesswire.com/news/home/20130108005330/en/Conexant-Introduces-Far-Field-Voice-Input-Processor-SoC](http://www.businesswire.com/news/home/20130108005330/en/Conexant-Introduces-Far-Field-Voice-Input-Processor-SoC).

- Calhoun, G., Draper, M., Ruff, H., Fontejon, J., & Guilfoos, B. (2003). Evaluation of tactile alerts for control station operation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 47(20), 2118-2122.
- Carter, C., & Graham, R. (2000). Experimental Comparison of manual and voice controls for the operation of in-vehicle systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 44(20), 286-289.
- Cassavaugh, N. D., & Kramer, A. F. (2009). Transfer of computer-based training to simulated driving in older adults. *Applied Ergonomics*, 40(5), 943-952.
- Chisholm, S., Caird, J., & Lockhart, J. (2008). The effects of practice with MP3 players on driving performance. *Accident Analysis & Prevention*, 40(2), 704-713.
- Crundall, D., Underwood, G., & Chapman, P. (1999). Driving experience and the functional field of view. *Perception*, 28(9), 1075–1087. doi:10.1068/p2894
- Detection Response Task (DRT). (2011, April 18). Retrieved from Lehrstuhl für Ergonomie - TU München at [www.lfe.mw.tum.de/en/research/labs/drt](http://www.lfe.mw.tum.de/en/research/labs/drt)
- Dijksterhuis, C., Stuiver, A., Mulder, B., Brookhuis, K. S., & de Waard, D. (2012). An adaptive driver support system: user experiences and driving performance in a simulator. *Human Factors*, 54(5), 772-785.
- Donmez, B., Boyle, L., & Lee, J. D. (2006). The impact of distraction mitigation strategies on driving performance. *Human factors*, 48(4), 785-804.
- Dybkjaer, L., Bernsen, N.O., & Minker, W. (2004). Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication*, 43, 33-54.
- Engström, J. (2009). The Tactile Detection Task as a method for assessing drivers' cognitive load. Warrendale, PA: Society of Automotive Engineers.
- Engström, J., Åberg, N., Johansson, E., & Hammarbäck, J. (2005). *Comparison between visual and tactile signal detection tasks applied to the safety assessment of in-vehicle information systems*. Paper presented at the Proceedings of the Third International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design.
- Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. Transportation Research Part F. *Traffic Psychology and Behaviour*, 8(2), 97-120.
- Engström, J., & Mårdh, S. (2007). SafeTE Final Report. (Report No. 2007:36). Gothenburg, Sweden: Swedish Road Agency.
- Engström, J. (2010). The tactile detection task as a method for assessing drivers' cognitive load. In G. L. Rupp, ed., *Performance Metrics for Assessing Driver Distraction: The Quest for Improved Road Safety*. Warrendale, PA: Society of Automotive Engineers.
- Entner, R. (2011). International Comparisons: The handset replacement cycle. Retrieved December 12, 2012, from [www.mobilefuture.org/page/handset-replacement-cycle.pdf](http://www.mobilefuture.org/page/handset-replacement-cycle.pdf)
- Fehd, H., & Seiffert, A. (2008). Eye movements during multiple object tracking: Where do participants look? doi:10.1016/j.cognition.2007.11.008.Eye

- Fitchard K. (2012) BMW taps Nuance for in-car speech recognition, Accessed on Jan 25, 2013 on <http://gigaom.com/2012/07/09/bmw-taps-nuance-for-in-car-speech-recognition/>.
- Fisher, D. L., Pradhan, A. K., Pollatsek, A. & Knodler, M. A. Jr. (2007). Empirical evaluation of hazard anticipation behaviors in the field and on a driving simulator using an eye tracker. *Transportation Research Record*, 2018, 80-86.
- Fisk, G. D., Owsley, C., & Mennemeier, M. (2002). Vision, attention, and self-reported driving behaviors in community-dwelling stroke survivors. *Archives of Physical Medicine and Rehabilitation*, 83(4), 469–477. doi:10.1053/apmr.2002.31179
- Fodor, B., Scheler, D., & Fingscheidt, t. (2012). a novel way to start speech dialogs in cars by talk-and-push (TAP). In J. Hansen, P. Boyraz, K. Takeda, & H. Abut, *Digital Signal Processing for In-Vehicle Systems and Safety*. Heidelberg: Springer.
- Fraser, N. M., & Gilbert, G. N. (1991). Simulating speech systems. *Computer Speech & Language*, 5(1), 81-99.
- Garay-Vega, L., Pradhan, A., Weinberg, G., Schmidt-Nielsen, B., Harsham, B., Shen, Y., . . . Fisher, D. (2010). Evaluation of different speech and touch interfaces to in-vehicle music retrieval systems. *Accident Analysis & Prevention*, 42(3), 913-920.
- Gellatly, A. W., & Dingus, T. A. (1998). Speech recognition and automotive applications: using speech to perform in-vehicle tasks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 42(17), 1247-1251.
- Gordon, M. A. (2007). Evaluating the Balloon Analogue Risk Task (BART) as a Predictor of Risk Taking in Adolescent and Adult Male Drivers. Hamilton, New Zealand: University of Waikato.
- Hamish Jamson, A., & Merat, N. (2005). Surrogate in-vehicle information systems and driver behaviour: Effects of visual and cognitive load in simulated rural driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 8(2), 79–96. doi:10.1016/j.trf.2005.04.002
- Harbluk, J. L., Burns, P. C., Hernandez, S., Tam, J., & Glazduri, V. (2013). Detection Response Tasks: Using Remote, Headmounted and Tactile Signals to Assess Cognitive Demand While Driving. In *7th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, (pp. 78-83), Bolton Landing, NY.
- Harbluk, Joanne, Noy, Ian, Trbovich, Patricia, Eizenman, Moshe (2007), An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance, *Accident Analysis & Prevention*, 39.
- Hardiess, G., & Mallot, H. (2010). Task-dependent representation of moving objects within working memory in obstacle avoidance. *Strabismus*. Retrieved from <http://informahealthcare.com/doi/abs/10.3109/09273972.2010.502958>
- Harms, L., & Patten, C. (2003). Peripheral detection as a measure of driver distraction. A study of memory-based versus system-based navigation in a built-up area. *Transportation Research Part F: Traffic Psychology and Behaviour*, 6(1), 23-36.
- Hart, S. G. & Staveland, L. E. (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.) *Human*



- mental workload*. Amsterdam: North Holland Press.
- Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX); 20 Years Later. In Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting, 904-908. Santa Monica: HFES.
- Horrey, W. J., & Wickens, C. D. (2004). *The impact of cell phone conversations on driving: A meta-analytic approach*. Warren, MI: General Motors Corporation.
- Hsieh, L., Young, R., & Seaman, S. (2012). Development of the Enhanced Peripheral Detection Task: A Surrogate Test for Driver Distraction. *SAE International Journal of Passenger Cars-Electronic and Electrical Systems*, 5(1), 317-325.
- Hua, Z., & Ng, W. L. (2010). *Speech recognition interface design for in-vehicle system*. Paper presented at the Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications.
- IMS Research (2013, February 7). Over half of all new vehicles will feature voice recognition in 2019. (Press Release).. Retrieved March 11, 2013, from [www.imsresearch.com/press-release/Over\\_Half\\_of\\_All\\_New\\_Vehicles\\_Will\\_Feature\\_Voice\\_Recognition\\_in\\_2019](http://www.imsresearch.com/press-release/Over_Half_of_All_New_Vehicles_Will_Feature_Voice_Recognition_in_2019)
- ISO/NP WD 17488: Road vehicles -Transport information and control systems - Man machine interface, 2012-02. Geneva: International Organization for Standardization.
- Jamson, A. H., Westerman, S. J., Hockey, G. R. J., & Carsten, O. M. J. (2004). Speech-based e-mail and driver behavior: Effects of an in-vehicle message system interface. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(4), 625–639. doi:10.1518/hfes.46.4.625.56814
- Jamson, A. H., & Merat, N. (2005). Surrogate in-vehicle information systems and driver behaviour: Effects of visual and cognitive load in simulated rural driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 8(2), 79-96.
- J.D. Power and Associates (2013). 2012 U.S. navigation usage and satisfaction study. Press release. Retrieved February 5, 2013 from: [www.jdpower.com/content/press-release/O4ycHMo/2012-u-s-navigation-usage-and-satisfaction-study.htm](http://www.jdpower.com/content/press-release/O4ycHMo/2012-u-s-navigation-usage-and-satisfaction-study.htm)
- Johnson, K. B., Syroid, N. D., Drews, F. A., Ogden, L. L., Strayer, D. L., Pace, N. L., . . . Westenskow, D. R. (2008). Part Task and variable priority training in first-year anesthesia resident education: A combined didactic and simulation-based approach to improve management of adverse airway and respiratory events. *Anesthesiology*, 108(5), 831-840.
- Johnson, T. (2012). Technical Overview of NHTSA’s Proposed Driver Distraction Guidelines. Presentation at ITS World Congress Special Interest Session 86. Available at: [www.ftw.at/veranstaltungen/SIS86\\_5\\_Johnson.pdf](http://www.ftw.at/veranstaltungen/SIS86_5_Johnson.pdf)
- Keane, B., & Pylyshyn, Z. (2006). Is motion extrapolation employed in multiple object tracking? Tracking as a low-level, non-predictive function. *Cognitive psychology*. Retrieved from [www.sciencedirect.com/science/article/pii/S0010028505001027](http://www.sciencedirect.com/science/article/pii/S0010028505001027)
- Kieras, D. E., Meyer, D. E., Ballas, J. A., & Lauber, E. J. (2000). Modern computational perspectives on executive mental processes and cognitive control: Where to from here.

*Control of cognitive processes: Attention and performance XVIII*, 681-712.

- Kubose, T. T., Bock, K., Dell, G. S., Garnsey, S. M., Kramer, A. F., & Mayhugh, J. (2006). The Effects of Speech Production and Speech. *Applied Cognitive Psychology*, 20, 43-63.
- Kun, A., Paek, T., & Medenica, Z. (2007). The Effect of speech interface accuracy on driving performance. Antwerp: Interspeech.
- Labiale, G. (1990). In-car road information: Comparisons of auditory and visual presentations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 34(9), 623–627. doi:10.1177/154193129003400909
- Lee, J. D., Caven, B., Haake, S., & Brown, T. L. (2001). Speech-based interaction with in-vehicle computers: The effect of speech-based e-mail on drivers' attention to the roadway. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 43(4), 631-640.
- Lee, J. D., Caven, B., Haake, S., & Brown, T. L. (2002). *Are conversations with your car distracting? Understanding the promises and pitfalls of speech-based interfaces*. IEEE Vehicular Technology Society News (pp. 2000–01–C012). Detroit: SAE.
- Leong, Cheng-I et al. (2012). *Safety Oriented Voice-based Interface for Vehicle's AV Systems: Talking Car Systems, Advances in Social and Organizational Factors*. Philadelphia Taylor & Francis Group.
- Lejuez, C., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., . . . Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, 8(2), 75.
- Lo, V.E., & Green, P.A. (2012). *Development and evaluation of automotive speech interfaces: Useful information from the human factors and related literature*. Retrieved February 5, 2013 from: <http://downloads.hindawi.com/journals/ijvt/aip/924170.pdf>
- Martens, M., & van Winsum, W. (1999). *The effect of speech versus tactile driver support messages on workload, driver behavior and user acceptance*. (TNO-report TM-99-C043). Soesterber, Netherlands: TNO Human Factors Research Institute.
- Mattes, S., & Hallén, A. (2009). Surrogate Distraction Measurement Techniques: The lane change test. In M. A. Regan, J. D. Lee & K. L. Young (Eds.), *Driver Distraction: Theory, Effects, and Mitigation* (pp. 107-121). Boca Raton, FL: Taylor & Francis Group.
- McCallum, M. C., Campbell, J. L., Rickman, J. B., Brown, J. L., & Wiese, E. (2004). Speech Recognition and In-Vehicle Telematics Devices: Potential Reductions in Driver Distraction. *International Journal of Speech Technology*, 7(1), 25-33.
- McCowan, I., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., et al. (2005). *On The Use of Information Retrieval Measures for Speech Recognition Evaluation*. Martigny, Switzerland: IDAP Research Institute.
- McEvoy, S. P., & Stevenson, M. R. (2008). Measuring exposure to driver distraction. In M. A. Regan, J. D. Lee & K. L. Young (Eds.), *Driver Distraction: Theory, Effects, and Mitigation* (pp. 73-83). Boca Raton, FL: Taylor & Francis Group.

- McKenna, F. P., & Crick, J. L. (1994). Hazard Perception In Drivers: A Methodology For Testing and Training. *TRL Contractor Report*(313).
- Mehler, B. B., Reimer, B., & Dusek, J. A. (2011). Delayed Digit Recall Task (n-back). Cambridge, MA: MIT AgeLab.
- Merat, N. (2003). Loading drivers to their limit: The effect of increasing secondary task on driving. Proceedings of the Second International Driving Symposium on Human Factors in Driving Assessment, Training and Vehicle Design, 13-18.
- Merat, N., Johansson, E., Engström, J., Chiu, E., & Nathan, F. (2006). *Specification of a secondary task to be used in Safety Assessment of IVIS*. AIDE Deliverable 2.2.3. European Commission, IST-1-507674-IP.
- Merat, N., & Jamson, A. (2007). Multisensory signal detection: How does driving and ivis management affect performance. Paper presented at the Proceedings of the 4th International Driving Symposium on Human Factors in Driver Assesemnt, Training and Vehicle Design.
- Miura, T. (1987). Behavior oriented vision: Functional field of view and processing resources. In O. R. J.K. & L.-S. A. (Eds.), *Eye Movements: From Physiology to Cognition*. Amsterdam: Elsevier North-Holland.
- Miura, T. (1990). Active function of eye movement and useful field of view in a realistic setting. In R. Groner, G. d'Ydewalle & R. Parham (Eds.), *From eye to mind: Information acquisition in perception, search, and reading*. Amsterdam: Elsevier North-Holland.
- Mishra, T., Ljolje, A., & Gilbert, M. (2011). Predicting Human Perceived Accuracy of ASR Systems. *12th Annual Conference of the International Speech Communication Association*. Florence, Italy: Interspeech.
- Moldenhauer, M. & McCrickard, S. (2003), Effect of Information Modality on Geographic Cognition in Car Navigation Systems, In M. Rauterberg, et al., (Eds.) *Human-Computer-Interaction – INTERACT'03*. Fairfax, VA: IOS Press.
- Morgan, J. F., & Hancock, P. A. (2009). Estimations in Driving. In C. Castro (Ed.), *Human Factors of Visual and Cognitive Performance in Driving* (pp. 51–62). FL: CRC Press.
- Mourant, R. R., & Rockwell, T. H. (1970). Mapping eye-movement patterns to the visual scene in driving: an exploratory study. *Human factors*, 12(1), 81–7.
- Muherer, E., Reinprecht, K., & Vollrath, M. (2012). Driving With a Partially Autonomous Foward Collision Warning System: How Do Drivers React? *Human Factors*, 54(5), 698-708.
- Nass, C., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. Cambridge, MA: MIT press.
- Nass, C., Jonsson, I.-M., Harris, H., Reaves, B., Endo, J., Brave, S., & Takayama, L. (2005). Improving automotive safety by pairing driver emotion and car voice emotion. Paper presented at the CHI'05 extended abstracts on Human factors in computing systems, Portland, OR.
- Navarathna, R., Deana, D., Sridharana, S., & Lucey, P. (2013). Multiple cameras for audio-

- visual speech recognition in an automotive environment. *Computer Speech & Language*, 27(4), pp. 911–927.
- NASA (2015). NASA-TLX Paper/Pencil Version. Available at <http://humansystems.arc.nasa.gov/groups/tlx/>
- National Highway Traffic Safety Administration. (2013). Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices (Docket No. NHTSA-2010-0053). Washington, DC: National Highway Traffic Safety Administration. Available at: [www.nhtsa.gov/staticfiles/nti/distracted\\_driving/pdf/distracted\\_guidelines-FR\\_04232013.pdf](http://www.nhtsa.gov/staticfiles/nti/distracted_driving/pdf/distracted_guidelines-FR_04232013.pdf)
- Neurauter, M. L., Hankey, J. M., Schalk, T. B., & Wallace, G. (2012). Outbound texting: Comparison of speech-based approach and handheld touch-screen equivalent. *Transportation Research Record*(2321), 23-30.
- Noy, I., & Harbluk, J. (2002). The Impact of Cognitive Distraction on Driver Visual Behaviour and Vehicle Control. Ottawa: Transport Canada.
- NRC Committee on Electronic Vehicle Controls and Unintended Acceleration. (2012). *The Safety Promise and Challenge of Automotive Electronics: Insights from Unintended Acceleration*. Washington D.C.: Transportation Research Board, National Research Council.
- Olafsson, S. (2012). Voice input processing for automotive speech recognition systems, EE Times Design. Accessed on Jan 25, 2013 from [www.eetimes.com/design/automotive-design/4394227/Voice-input-processing-for-automotive-speech-recognition-systems](http://www.eetimes.com/design/automotive-design/4394227/Voice-input-processing-for-automotive-speech-recognition-systems).
- Owsley, C., Ball, K., Sloabe, M. ., Roenker, D. L., & Bruni, J. R. (1991). Visual/cognitive correlates of vehicle accidents in older drivers. *Psychology and Aging*, 6(3), 403–415.
- Parasuraman, R., & Nestor, P. G. (1991). Attention and driving skills in aging and Alzheimer's disease. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 33(5), 539–557. doi:10.1177/001872089103300506
- Patten, C. J. D., Kircher, A., Ostlund, J., & Nilsson, L. (2004). Using mobile telephones: cognitive workload and attention resource allocation. *Accident; Analysis and Prevention*, 36(3), 341–50.
- Putze, F., & Schultz, T. (2012). Cognitive Dialog Systems for Dynamic Environments: Progress and Challenges. In J. Hansen, P. Boyraz, K. Takeda, & H. Abut, *Digital Signal Processing for In-Vehicle Systems and Safety*. Heidelberg: Springer.
- Ranney, T. A. (1994). Models of driving behavior: a review of their evolution. *Accident Analysis and Prevention*, 26(6), 733–750. Retrieved from [www.ncbi.nlm.nih.gov/pubmed/7857489](http://www.ncbi.nlm.nih.gov/pubmed/7857489)
- Ranney, T. A., Baldwin, G., Vasko, S. M., & Mazzae, E. N. (2009). Measuring Distraction Potential of Operating In-Vehicle Devices. (Report No. DOT HS 811 231). Washington, DC: National Highway Traffic Safety Administration.
- 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.
- Ranney, T. A., Baldwin, G. H. S., Smith, L. A., Mazzae, E. N., & Pierce, R. S. (2014, November). *Detection response task evaluation for driver distraction measurement application*. (Report No. DOT HS 812 077). Washington, DC: National Highway Traffic Safety Administration.
- Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of Experimental Psychology: Applied; Journal of Experimental Psychology: Applied*, 9(2), 119.
- Reimer, B. (2009). Impact of cognitive task complexity on drivers' visual tunneling. *Transportation Research Record: Journal of the Transportation Research Board*, 2138(-1), 13-19.
- Reimer, B., & Mehler, B. (2013). *The Effects of a Production Level "Voice-Command" Interface on Driver Behavior: Summary Findings on Reported Workload, Physiology, Visual Attention, and Driving Performance*. Cambridge, MA. doi:2013-18A
- Rizzo, M., Stierman, L., Skaar, N., Dawson, J. D., Anderson, S. W., & Vecera, S. P. (2004). Effects of a Controlled Auditory—Verbal Distraction Task on Older Driver Vehicle Control. *Transportation Research Record: Journal of the Transportation Research Board*, 1865(-1), 1-6.
- Romoser, M. R. E., Pollatsek, A., Fisher, D. L., & Williams, C. C. (2013). Comparing the glance patterns of older versus younger experienced drivers: Scanning for hazards while approaching and entering the intersection. *Transportation Research Part F*, 16, 104-116.
- Romoser, M. and Fisher, D. L. (2009). The effect of active versus passive training strategies on improving older drivers' scanning for hazards while negotiating intersections. *Human Factors*, 51, 652-668.
- SAE International (2015). *Guidelines for Speech Input and Audible Output in a Driver Vehicle Interface*. *Surface Vehicle Information Report J2988*. Warrendale, PA: Author. Retrieved June 5, 2015 from [www.sae.org](http://www.sae.org).
- Society of Automotive Engineers (2004). *Navigation and route guidance function accessibility while driving (SAE recommended practice 2364)*. Warrendale, PA: Author.
- Sanbonmatsu, D. M., Strayer, D. L., Medeiros-Ward, N., & Watson, J. M. (2013). Who Multi-Tasks and Why? Multi-Tasking Ability, Perceived Multi-Tasking Ability, Impulsivity, and Sensation Seeking. *PloS one*, 8(1), e54402.
- Scott, J. J., & Gray, R. (2008). A Comparison of Tactile, Visual, and Auditory Warnings for Rear-End Collision Prevention in Simulated Driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(2), 264-275.
- Smith, R.K., T. Luke, A.P. Parkes, P.C. Burns, T.C. Landsdown (2005). A study of driver visual behavior while talking with passengers, and on mobile phones, *Human Factors in Design, Safety, and Management* Maastricht, Netherlands: Shaker Publishing,, pp. 11–22.

- Sodhi, M., Reimer, B., & Llamazares, I. (2002). Glance analysis of driver eye movements to evaluate distraction. *Behavior Research Methods, Instruments, & Computers*, 34(4), 529-538.
- Strayer, D. L., & Johnston, W. A. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological Science*, 12(6), 462-466.
- Taylor, T., Roman, L., McFeaters, K., Romoser, M., Borowsky, A., Merritt, D., Pollatsek, A., Lee, J., & Fisher, D. L. (in review). The Effect of Cell Phone Conversations on Latent Hazard Anticipation. *Human Factors*.
- Tijerina, L., Parmer, E., & Goodman, M. J. (1998). Driver Workload Assessment of Route Guidance System Destination Entry While Driving: A Test Track Study. *Proceedings of the 5th ITS World Congress*. Seoul, Korea.
- Tsimhoni, O., Smith, D., & Green, P. (2002). Destination Entry while Driving: Speech Recognition versus a Touch-Screen Keyboard. Technical Report UMTRI-2001-241. Ann Arbor: University of Michigan Transportation Research Institute.
- Tsimhoni, O., Smith, D., & Green, P. (2004, 01). Address Entry While Driving: Speech Recognition Versus a Touch-Screen Keyboard. Ann Arbor: University of Michigan Transportation Research Institute
- Van Winsum, W., Martens, M., & Herland, L. (1999). The Effects of Speech Versus Tactile Driver Support Messages on Workload Driver Behaviour and User Acceptance: TNO-Report, TM-99-C043. Soesterber, Netherlands TNO Human Factors Research Institute.
- Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of memory and language*, 28(2), 127-154.
- van Winsum, W., Martens, M., & Herland, L. (1999). The Effects of Speech Versus Tactile Driver Support Messages on Workload Driver Behaviour and User Acceptance: TNO-Report, TM-99-C043. Soesterber, Netherlands: TNO Human Factors Research Institute.
- Vaux, L. M., Ni, R., Rizzo, M., Uc, E. Y., & Andersen, G. J. (2010). Detection of imminent collisions by drivers with Alzheimer's disease and Parkinson's disease: a preliminary study. *Accident Analysis and Prevention*, 42(3), 852-8.
- Victor, T. W. (2005). Keeping eye and mind on the road. Uppsala, Sweden: Uppsala University,.
- Victor, T. W., Engström, J., & Harbluk, J. L. (2008). Distraction Assessment Methods Based on Visual Behavior and Event Detection. In M. A. Regan, J. D. Lee & K. L. Young (Eds.), *Driver Distraction: Theory, Effects, and Mitigation* (pp. 135-165). Boca Raton, FL: Taylor & Francis Group.
- Victor, T. W., Harbluk, J. L., & Engström, J. A. (2005). Sensitivity of eye-movement measures to in-vehicle task difficulty. *Transportation Research Part F: Traffic Psychology and Behaviour*, 8(2), 167-190.
- Vlakveld, W., Romoser, M. R. E., Mehranian, H., Diete, F., and Fisher, D. L. (2011). Does the experience of crashes and near crashes in a simulator-based training program enhance novice driver's visual search for latent hazards? *Transportation Research Record*, 2265, 153-160.
- Wickens, C. D., Lee, J., Liu, Y. D., & Gordon-Becker, S. (2003). *Introduction to human factors*

- engineering*. Upper Saddle River, NJ: Pearson Education.
- Wiese, E. E., & Lee, J. D. (2007). Attention grounding: a new approach to IVIS implementation. *Theoretical Issues in Ergonomics Science*, 8(3), 255–276.  
doi:10.1080/14639220601129269
- Wood, J., Chaparro, A., Hickson, L., Thyer, N., Carter, P., Hancock, J., . . . & Ybarzabal, F. (2006). The effect of auditory and visual distracters on the useful field of view: Implications for the driving task. *Investigative Ophthalmology & Visual Science*, 47(10), 4646-4650.
- Yager, C. (2013). *An Evaluation of the Effectiveness of Voice-to-Text Programs at Reducing Incidence of Distracted Driving*. (Report No. SWUTC/13/600451-00011-1. 2013).  
College Station, Texas: Southwest Region University Transportation Center.
- Young, R. A., & Angell, L. S. (2003). The Dimensions of Driver Performance during Secondary Manual Tasks. In *Driving Assessment: International driving symposium on human factors in driver assessment, training, and vehicle design*. Park City, Utah.
- Young, R. A., Hsieh, L., & Seaman, S. (2013). The Tactile Detection Response Task: Preliminary Validation for Measuring the Attentional Effects of Cognitive Load. *7th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*. Bolton Landing, NY.

## 7 APPENDIX A

### 7.1 STUDY 1: CONTEXTUAL INTERVIEWS

#### 7.1.1 Voice Control Tasks Experience Questionnaire

Do you use your voice control system to:	Yes/No		Frequency						Driving Conditions				
	Circle One:		Circle One:						Circle All That Apply:				
			1.	2.	3.	4.	5.	6.	A.	B.	C.	D.	E.
Change system settings (e.g., brightness, contrast, personal information)	Y	N	1	2	3	4	5	6	A	B	C	D	E
Adjust interior climate control	Y	N	1	2	3	4	5	6	A	B	C	D	E
Get vehicle health report	Y	N	1	2	3	4	5	6	A	B	C	D	E
Tune to a radio station	Y	N	1	2	3	4	5	6	A	B	C	D	E
Listen to and control music on an mp3 player (e.g., play, shuffle, browse music, skip songs, pause)	Y	N	1	2	3	4	5	6	A	B	C	D	E



Listen to and control audiobooks on BlueTooth® Device (e.g., play, switch books, browse books, pause)	Y	N	1	2	3	4	5	6	A	B	C	D	E
Search for directions to a known address	Y	N	1	2	3	4	5	6	A	B	C	D	E
Search for directions to a point of interest (e.g., restaurants, gas stations, ATMs, parking garages, etc.)	Y	N	1	2	3	4	5	6	A	B	C	D	E
Get estimated arrival time (ETA) to a destination	Y	N	1	2	3	4	5	6	A	B	C	D	E
Modify current route (e.g., take a toll free route, avoid highways, add a new destination along current route)	Y	N	1	2	3	4	5	6	A	B	C	D	E
Place a phone call to a family member/friend	Y	N	1	2	3	4	5	6	A	B	C	D	E
Call and listen to your voicemail	Y	N	1	2	3	4	5	6	A	B	C	D	E
Add a new phone number to your phonebook/contact list	Y	N	1	2	3	4	5	6	A	B	C	D	E
Create and send a text message	Y	N	1	2	3	4	5	6	A	B	C	D	E
Listen to a text message	Y	N	1	2	3	4	5	6	A	B	C	D	E
Listen to your e-mail	Y	N	1	2	3	4	5	6	A	B	C	D	E

Check the news (e.g., top headlines, sports scores, stock quotes, etc.)	Y	N	1	2	3	4	5	6	A	B	C	D	E
Check the weather forecast	Y	N	1	2	3	4	5	6	A	B	C	D	E
Check the traffic report	Y	N	1	2	3	4	5	6	A	B	C	D	E
Check gas prices	Y	N	1	2	3	4	5	6	A	B	C	D	E
Plan for a future activity (e.g., search movie theater listings and/or purchase tickets, search for and/or book a dinner reservation, etc.)	Y	N	1	2	3	4	5	6	A	B	C	D	E
Browse the web	Y	N	1	2	3	4	5	6	A	B	C	D	E
Check your social media accounts (e.g., listen to your Twitter feed)	Y	N	1	2	3	4	5	6	A	B	C	D	E
Update social media accounts (e.g., compose and send a Tweet, favorite a Tweet, etc.)	Y	N	1	2	3	4	5	6	A	B	C	D	E

## 7.2 STUDY 2: DRIVING SIMULATOR STUDY IN SEATTLE, WA

### 7.2.1 Total task duration

Table 20: Total task duration

	NumDF	DenDF	F.value	Pr(>F)
<b>Recognition Error</b>	1	43.99	261.76	<0.001
<b>Task</b>	2	87.95	355.01	<0.001
<b>Order</b>	1	43.99	0.02	0.9
<b>Delay</b>	1	132	332.91	<0.001
<b>Recognition Error * Task</b>	2	87.95	11.45	<0.001
<b>Recognition Error * Order</b>	1	43.99	0.05	0.83
<b>Task * Order</b>	2	87.95	0.42	0.66
<b>Recognition Error * Delay</b>	1	132	16.08	<0.001
<b>Task * Delay</b>	2	132	1.45	0.24
<b>Order * Delay</b>	1	132	0	0.99
<b>Recognition Error * Task * Order</b>	2	87.95	0.51	0.6
<b>Recognition Error * Task * Delay</b>	2	132	1.23	0.3
<b>Recognition Error * Order * Delay</b>	1	132	1.96	0.16
<b>Task * Order * Delay</b>	2	132	0.06	0.94
<b>Recognition Error * Task * Order * Delay</b>	2	132	0.75	0.47

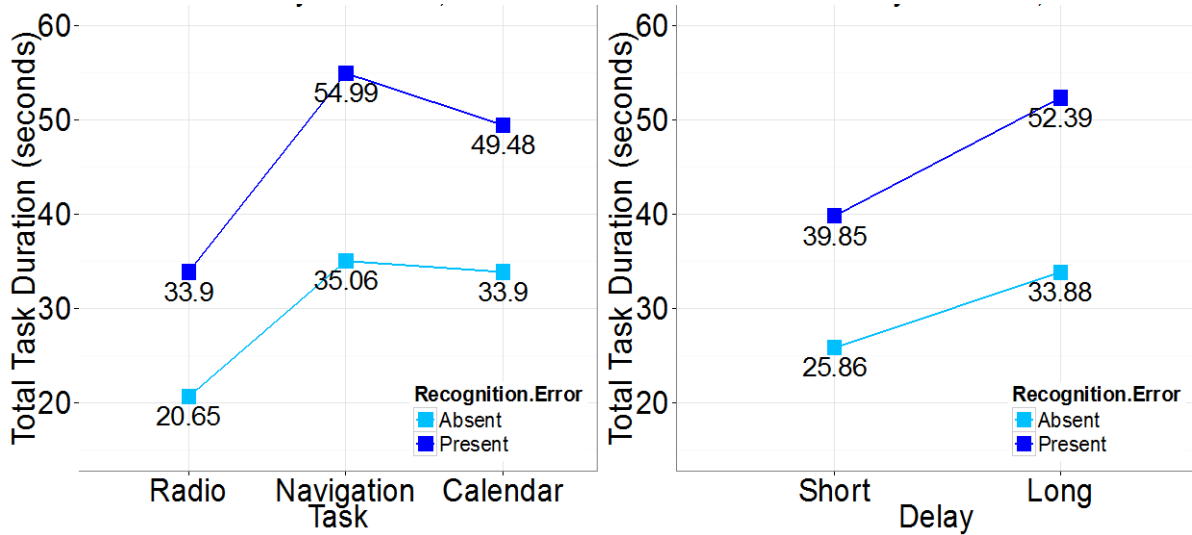


Figure 34. Two-way interactions of Recognition Error x Task and Recognition Error x Delay for total task duration.

## 7.2.2 Tactile detection response task performance measure: Reaction time

Table 21: TDRT Reaction time

	NumDF	DenDF	F.value	Pr(>F)
<b>Recognition Error</b>	1	43.8	0.02	0.88
<b>Task</b>	2	88.32	10.22	<0.001
<b>Order</b>	1	43.8	1.37	0.25
<b>Delay</b>	1	42.86	0.08	0.78
<b>Recognition Error * Task</b>	2	88.32	6.44	<0.01
<b>Recognition Error * Order</b>	1	43.8	0.2	0.66
<b>Task * Order</b>	2	88.32	0.28	0.76
<b>Recognition Error * Delay</b>	1	42.86	0.96	0.33
<b>Task * Delay</b>	2	88.19	0.8	0.45
<b>Order * Delay</b>	1	42.86	0.49	0.49
<b>Recognition Error * Task * Order</b>	2	88.32	0.61	0.55
<b>Recognition Error * Task * Delay</b>	2	88.19	0.38	0.69
<b>Recognition Error * Order * Delay</b>	1	42.86	1.06	0.31
<b>Task * Order * Delay</b>	2	88.19	0.03	0.97
<b>Recognition Error * Task * Order * Delay</b>	2	88.19	0.26	0.77

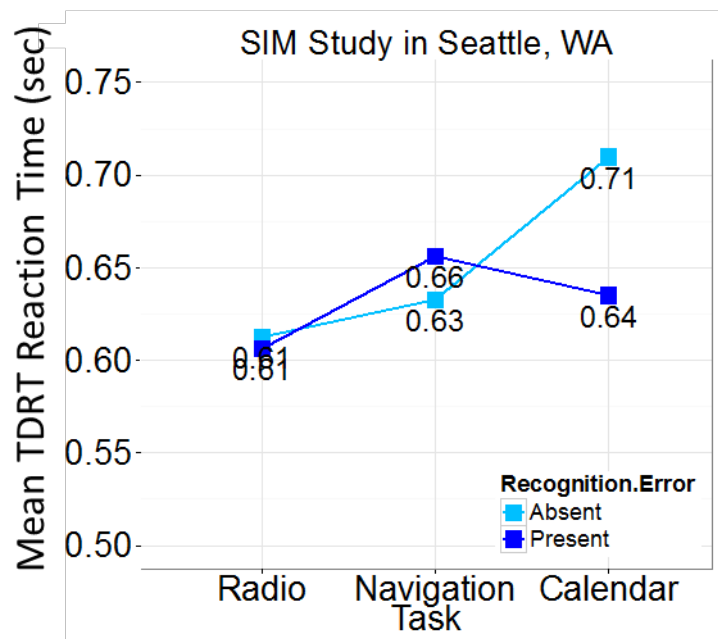


Figure 35. Two-way interaction of Recognition Error x Task for mean TDRT reaction time.

### 7.2.3 Criterion 1: Percentage of long eyes-off-road glances ( $\geq 2$ seconds)

Table 22: Percentage of long EOR from mean of trials

	NumDF	DenDF	F.value	Pr(>F)
<b>Recognition Error</b>	1	43.99	0.11	0.74
<b>Task</b>	2	220.00	0.75	0.47
<b>Order</b>	1	43.99	0.34	0.56
<b>Delay</b>	1	220.00	1.44	0.23
<b>Recognition Error * Task</b>	2	220.00	0.05	0.95
<b>Recognition Error * Order</b>	1	43.99	1.95	0.17
<b>Task * Order</b>	2	220.00	0.05	0.95
<b>Recognition Error * Delay</b>	1	220.00	0.35	0.56
<b>Task * Delay</b>	2	220.00	1.12	0.33
<b>Order * Delay</b>	1	220.00	0.09	0.77
<b>Recognition Error * Task * Order</b>	2	220.00	1.33	0.27
<b>Recognition Error * Task * Delay</b>	2	220.00	1.21	0.30
<b>Recognition Error * Order * Delay</b>	1	220.00	1.57	0.21
<b>Task * Order * Delay</b>	2	220.00	2.33	0.10
<b>Recognition Error * Task * Order * Delay</b>	2	220.00	0.29	0.75

Table 23: Percentage of long EOR from maximum of trials

	NumDF	DenDF	F.value	Pr(>F)
<b>Recognition Error</b>	1	43.99	0.11	0.75
<b>Task</b>	2	87.93	1.74	0.18
<b>Order</b>	1	43.99	0.15	0.70
<b>Delay</b>	1	132.00	0.94	0.33
<b>Recognition Error * Task</b>	2	87.93	0.26	0.78
<b>Recognition Error * Order</b>	1	43.99	2.13	0.15
<b>Task * Order</b>	2	87.93	0.64	0.53
<b>Recognition Error * Delay</b>	1	132.00	0.00	0.98
<b>Task * Delay</b>	2	132.00	1.05	0.35
<b>Order * Delay</b>	1	132.00	0.97	0.33
<b>Recognition Error * Task * Order</b>	2	87.93	1.33	0.27
<b>Recognition Error * Task * Delay</b>	2	132.00	0.24	0.79
<b>Recognition Error * Order * Delay</b>	1	132.00	2.62	0.11
<b>Task * Order * Delay</b>	2	132.00	3.08	0.05
<b>Recognition Error * Task * Order * Delay</b>	2	132.00	0.29	0.75

## 7.2.4 Criterion 2: Mean glance duration

Table 24: Mean glance duration from mean of trials

	NumDF	DenDF	F.value	Pr(>F)
<b>Recognition Error</b>	1	44.00	0.14	0.71
<b>Task</b>	2	219.94	2.62	0.07
<b>Order</b>	1	44.00	0.03	0.86
<b>Delay</b>	1	219.94	0.18	0.67
<b>Recognition Error * Task</b>	2	219.94	0.31	0.73
<b>Recognition Error * Order</b>	1	44.00	1.00	0.32
<b>Task * Order</b>	2	219.94	0.17	0.84
<b>Recognition Error * Delay</b>	1	219.94	0.41	0.52
<b>Task * Delay</b>	2	219.94	1.29	0.28
<b>Order * Delay</b>	1	219.94	4.99	0.03
<b>Recognition Error * Task * Order</b>	2	219.94	1.98	0.14
<b>Recognition Error * Task * Delay</b>	2	219.94	0.76	0.48
<b>Recognition Error * Order * Delay</b>	1	219.94	2.22	0.14
<b>Task * Order * Delay</b>	2	219.94	1.16	0.32
<b>Recognition Error * Task * Order * Delay</b>	2	219.94	1.03	0.36

Table 25: Mean glance duration from maximum of trials

	NumDF	DenDF	F.value	Pr(>F)
<b>Recognition Error</b>	1	44.00	0.18	0.67
<b>Task</b>	2	175.97	1.79	0.17
<b>Order</b>	1	44.00	0.01	0.93
<b>Delay</b>	1	43.96	0.02	0.88
<b>Recognition Error * Task</b>	2	175.97	0.39	0.67
<b>Recognition Error * Order</b>	1	44.00	1.21	0.27
<b>Task * Order</b>	2	175.97	0.19	0.82
<b>Recognition * Error * Delay</b>	1	43.96	0.21	0.65
<b>Task * Delay</b>	2	175.97	1.42	0.24
<b>Order * Delay</b>	1	43.96	1.52	0.22
<b>Recognition Error * Task * Order</b>	2	175.97	1.38	0.25
<b>Recognition Error * Task * Delay</b>	2	175.97	0.80	0.45
<b>Recognition Error * Order * Delay</b>	1	43.96	2.34	0.13
<b>Task * Order * Delay</b>	2	175.97	2.21	0.11
<b>Recognition Error * Task * Order * Delay</b>	2	175.97	0.94	0.39

### 7.2.5 Criterion 3: Total eyes-off-road time

Table 26: Total eyes-off-road time from mean of trials

	NumDF	DenDF	F.value	Pr(>F)
<b>Recognition Error</b>	1	44.00	1.76	0.19
<b>Task</b>	<b>2</b>	<b>175.99</b>	<b>10.86</b>	<b>&lt;0.001</b>
<b>Order</b>	1	44.00	0.00	0.96
<b>Delay</b>	<b>1</b>	<b>43.99</b>	<b>8.19</b>	<b>&lt;0.01</b>
<b>Recognition Error * Task</b>	2	175.99	0.41	0.66
<b>Recognition Error * Order</b>	1	44.00	0.01	0.92
<b>Task * Order</b>	2	175.99	0.36	0.70
<b>Recognition Error * Delay</b>	1	43.99	0.23	0.63
<b>Task * Delay</b>	2	175.99	1.44	0.24
<b>Order * Delay</b>	1	43.99	0.09	0.77
<b>Recognition Error * Task * Order</b>	2	175.99	0.50	0.61
<b>Recognition Error * Task * Delay</b>	2	175.99	1.11	0.33
<b>Recognition Error * Order * Delay</b>	1	43.99	0.05	0.83
<b>Task * Order * Delay</b>	2	175.99	1.76	0.17
<b>Recognition Error * Task * Order * Delay</b>	2	175.99	0.56	0.57

Table 27: Total eyes-off-road time from maximum of trials

	NumDF	DenDF	F.value	Pr(>F)
<b>Recognition Error</b>	1	44.00	1.50	0.23
<b>Task</b>	2	219.99	0.28	0.76
<b>Order</b>	1	44.00	0.07	0.79
<b>Delay</b>	<b>1</b>	<b>219.99</b>	<b>11.57</b>	<b>&lt;0.001</b>
<b>Recognition Error * Task</b>	2	219.99	0.01	0.99
<b>Recognition Error * Order</b>	1	44.00	0.14	0.71
<b>Task * Order</b>	2	219.99	1.41	0.25
<b>Recognition Error * Delay</b>	1	219.99	0.06	0.80
<b>Task * Delay</b>	2	219.99	3.48	0.03
<b>Order * Delay</b>	1	219.99	1.01	0.32
<b>Recognition Error * Task * Order</b>	2	219.99	0.47	0.63
<b>Recognition Error * Task * Delay</b>	2	219.99	0.83	0.44
<b>Recognition Error * Order * Delay</b>	1	219.99	0.02	0.90
<b>Task * Order * Delay</b>	<b>2</b>	<b>219.99</b>	<b>4.57</b>	<b>0.01</b>
<b>Recognition Error * Task * Order * Delay</b>	2	219.99	0.62	0.54

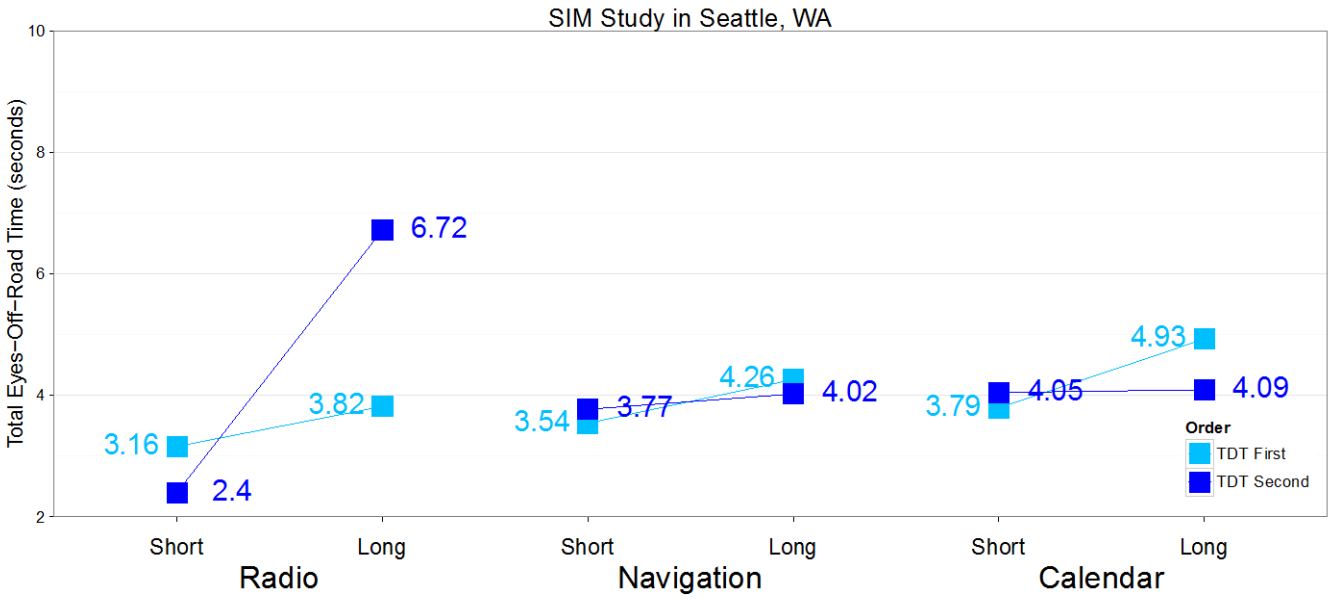


Figure 36. Three-way interaction of Task x Delay x Order for total eyes-off-road time from maximum of trials.



### 7.3 STUDY 3: CDT STUDY IN MADISON, WI

#### 7.3.1 Total task duration

Table 28: Total task duration

	NumDF	DenDF	F.value	Pr(>F)
<b>Recognition Error</b>	1	39.96	200.92	<0.001
<b>Task</b>	2	80.59	221.82	<0.001
<b>Order</b>	1	39.96	0.18	0.67
<b>Delay</b>	1	39.16	243.95	<0.001
<b>Recognition Error * Task</b>	2	80.59	1.29	0.28
<b>Recognition Error * Order</b>	1	39.96	4.02	0.05
<b>Task * Order</b>	2	80.59	0.12	0.89
<b>Recognition Error * Delay</b>	1	39.16	23.88	<0.001
<b>Task * Delay</b>	2	78.64	0.53	0.59
<b>Order * Delay</b>	1	39.16	7.07	0.01
<b>Recognition Error * Task * Order</b>	2	80.59	4.02	0.02
<b>Recognition Error * Task * Delay</b>	2	78.64	0.03	0.97
<b>Recognition Error * Order * Delay</b>	1	39.16	0.42	0.52
<b>Task * Order * Delay</b>	2	78.64	2.3	0.1
<b>Recognition Error * Task * Order * Delay</b>	2	78.64	3.88	0.02

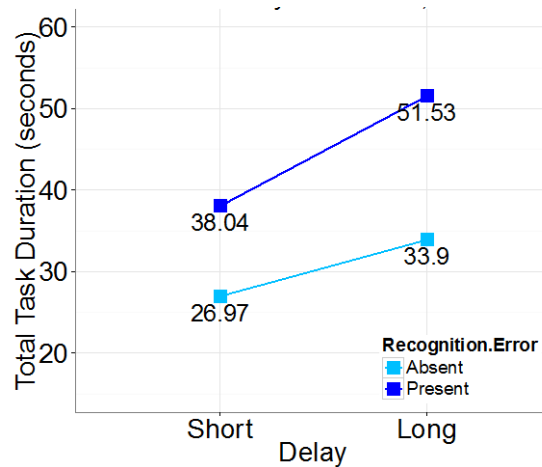


Figure 37. Two-way interaction of Recognition Error x Delay for total task duration.

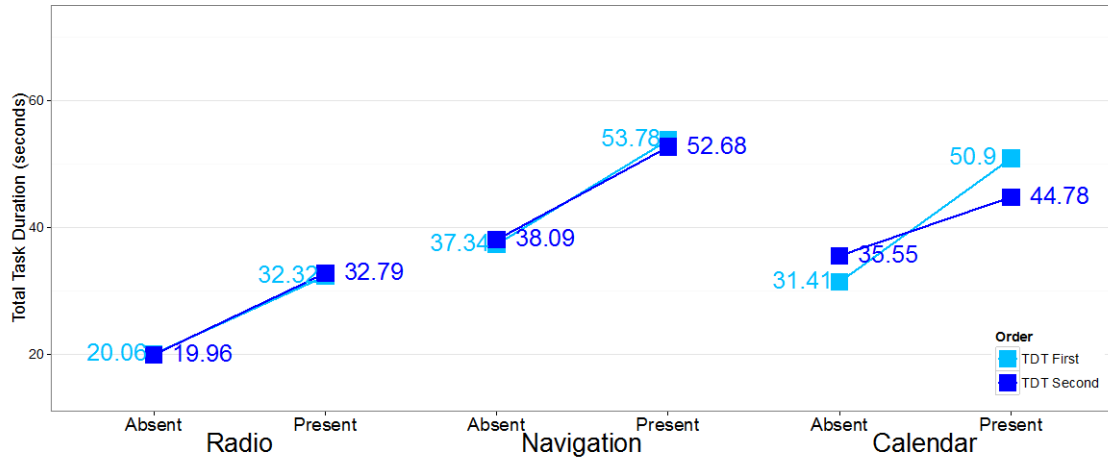


Figure 38. Three-way interaction of Recognition Error x Task x Order for total task duration.

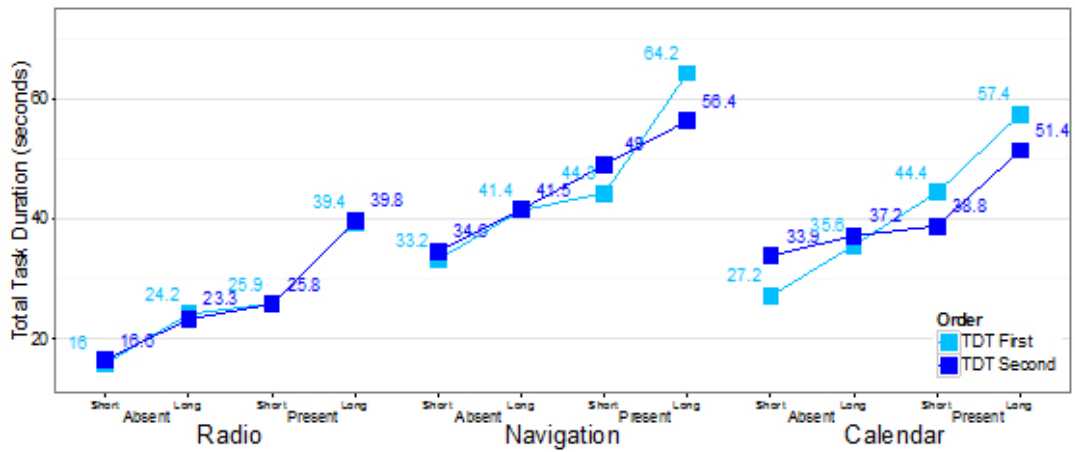


Figure 39. Four-way interaction between Recognition Error x Task x Delay x Order for total task duration.

### 7.3.2 Tactile Detection Response Task Performance Measure: Reaction Time

Table 29: TDRT Reaction Time

	<b>NumDF</b>	<b>DenDF</b>	<b>F.value</b>	<b>Pr(&gt;F)</b>
<b>Recognition Error</b>	1	40.02	2.12	0.15
<b>Task</b>	2	80.13	4.81	0.01
<b>Order</b>	1	40.02	1.06	0.31
<b>Delay</b>	1	118.52	3.65	0.06
<b>Recognition Error * Task</b>	2	80.13	0.59	0.56
<b>Recognition Error * Order</b>	1	40.02	3.01	0.09
<b>Task * Order</b>	2	80.13	0.38	0.69
<b>Recognition Error * Delay</b>	1	118.52	0.02	0.9
<b>Task * Delay</b>	2	118.77	0.62	0.54
<b>Order * Delay</b>	1	118.52	0.76	0.39
<b>Recognition Error * Task * Order</b>	2	80.13	0.07	0.93
<b>Recognition Error * Task * Delay</b>	2	118.77	0.23	0.8
<b>Recognition Error * Order * Delay</b>	1	118.52	0.14	0.71
<b>Task * Order * Delay</b>	2	118.77	0.16	0.85
<b>Recognition Error * Task * Order * Delay</b>	2	118.77	1.28	0.28

### 7.3.3 Criterion 1: Percentage of long eyes-off-road glances ( $\geq 2$ seconds)

Table 30: Percentage of long EOR from mean of trials

	NumDF	DenDF	F.value	Pr(>F)
<b>Recognition Error</b>	<b>1</b>	<b>79.80</b>	<b>12.76</b>	<b>&lt;0.001</b>
<b>Task</b>	<b>2</b>	<b>158.53</b>	<b>4.55</b>	<b>0.01</b>
<b>Order</b>	1	79.80	0.01	0.94
<b>Delay</b>	1	79.80	0.22	0.64
<b>Recognition Error * Task</b>	<b>2</b>	<b>158.53</b>	<b>4.04</b>	<b>0.02</b>
<b>Recognition Error * Order</b>	1	79.80	0.09	0.76
<b>Task * Order</b>	2	158.53	2.14	0.12
<b>Recognition Error * Delay</b>	1	79.80	0.26	0.61
<b>Task * Delay</b>	2	158.53	1.18	0.31
<b>Order * Delay</b>	1	79.80	1.75	0.19
<b>Recognition Error * Task * Order</b>	2	158.53	2.21	0.11
<b>Recognition Error * Task * Delay</b>	2	158.53	1.74	0.18
<b>Recognition Error * Order * Delay</b>	1	79.80	1.85	0.18
<b>Task * Order * Delay</b>	2	158.53	0.31	0.74
<b>Recognition Error * Task * Order * Delay</b>	2	158.53	0.53	0.59

Table 31: Percentage of long EOR from maximum of trials

	NumDF	DenDF	F.value	Pr(>F)
<b>Recognition Error</b>	<b>1</b>	<b>79.56</b>	<b>14.84</b>	<b>&lt;0.001</b>
<b>Task</b>	2	158.36	2.86	0.06
<b>Order</b>	1	79.56	0.00	0.99
<b>Delay</b>	1	79.56	0.06	0.81
<b>Recognition Error * Task</b>	2	158.36	2.43	0.09
<b>Recognition Error * Order</b>	1	79.56	0.06	0.80
<b>Task * Order</b>	2	158.36	2.37	0.10
<b>Recognition Error * Delay</b>	1	79.56	0.08	0.77
<b>Task * Delay</b>	2	158.36	1.31	0.27
<b>Order * Delay</b>	1	79.56	3.14	0.08
<b>Recognition Error * Task * Order</b>	2	158.36	2.55	0.08
<b>Recognition Error * Task * Delay</b>	2	158.36	1.98	0.14
<b>Recognition Error * Order * Delay</b>	1	79.56	3.30	0.07
<b>Task * Order * Delay</b>	2	158.36	0.28	0.76
<b>Recognition Error * Task * Order * Delay</b>	2	158.36	0.51	0.60

### 7.3.4 Mean glance duration

Table 32: Mean glance duration from mean of trials

	NumDF	DenDF	F.value	Pr(>F)
<b>Recognition Error</b>	<b>1</b>	<b>39.75</b>	<b>9.31</b>	<b>&lt;0.01</b>
<b>Task</b>	<b>2</b>	<b>158.11</b>	<b>43.29</b>	<b>&lt;0.001</b>
<b>Order</b>	1	39.75	0.03	0.86
<b>Delay</b>	1	39.76	0.14	0.71
<b>Recognition Error * Task</b>	<b>2</b>	<b>158.11</b>	<b>4.72</b>	<b>0.01</b>
<b>Recognition Error * Order</b>	1	39.75	0.06	0.82
<b>Task * Order</b>	2	158.11	1.49	0.23
<b>Recognition Error * Delay</b>	1	39.76	0.64	0.43
<b>Task * Delay</b>	<b>2</b>	<b>158.11</b>	<b>7.71</b>	<b>&lt;0.001</b>
<b>Order * Delay</b>	1	39.76	0.05	0.83
<b>Recognition Error * Task * Order</b>	2	158.11	1.81	0.17
<b>Recognition Error * Task * Delay</b>	2	158.11	3.50	0.03
<b>Recognition Error * Order * Delay</b>	1	39.76	0.05	0.82
<b>Task * Order * Delay</b>	2	158.11	0.74	0.48
<b>Recognition Error * Task * Order * Delay</b>	2	158.11	1.28	0.28

Table 33: Mean glance duration from maximum of trials

	NumDF	DenDF	F.value	Pr(>F)
<b>Recognition Error</b>	<b>1</b>	<b>39.77</b>	<b>9.86</b>	<b>&lt;0.01</b>
<b>Task</b>	<b>2</b>	<b>75.73</b>	<b>40.02</b>	<b>&lt;0.001</b>
<b>Order</b>	1	39.77	0.15	0.70
<b>Delay</b>	1	38.52	0.52	0.47
<b>Recognition Error * Task</b>	<b>2</b>	<b>75.73</b>	<b>5.13</b>	<b>&lt;0.01</b>
<b>Recognition Error * Order</b>	1	39.77	0.04	0.84
<b>Task * Order</b>	2	75.73	0.87	0.42
<b>Recognition Error * Delay</b>	1	38.52	0.15	0.70
<b>Task * Delay</b>	<b>2</b>	<b>70.45</b>	<b>6.75</b>	<b>&lt;0.01</b>
<b>Order * Delay</b>	1	38.52	0.11	0.74
<b>Recognition Error * Task * Order</b>	2	75.73	2.39	0.10
<b>Recognition Error * Task * Delay</b>	2	70.45	3.31	0.04
<b>Recognition Error * Order * Delay</b>	1	38.52	0.00	0.97
<b>Task * Order * Delay</b>	2	70.45	1.64	0.20
<b>Recognition Error * Task * Order * Delay</b>	2	70.45	1.84	0.17

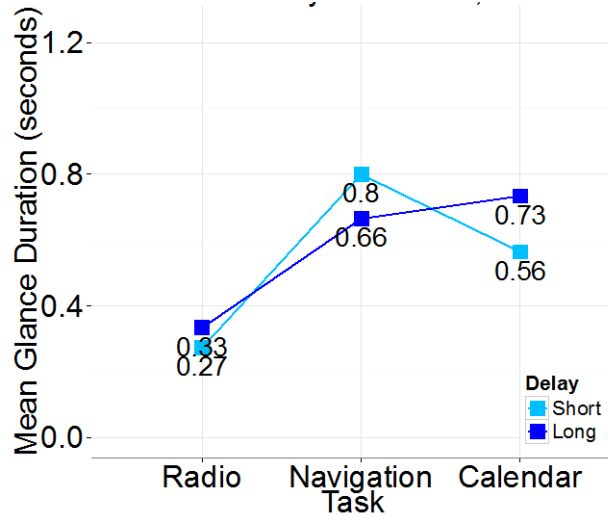


Figure 40. Two-way interaction of Task x Delay for mean glance duration from maximum of trials.

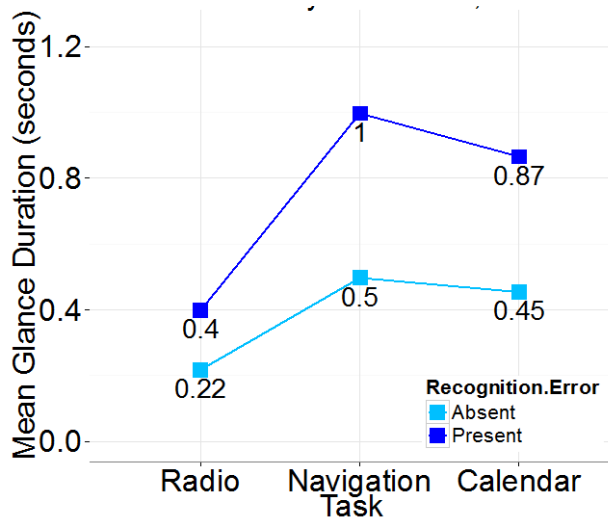


Figure 41. Two-way interaction of Recognition Error x Task for mean glance duration from maximum of trials.

### 7.3.5 Total eyes-off-road time

Table 34: Total eyes-off-road time from mean of trials

	NumDF	DenDF	F.value	Pr(>F)
<b>Recognition Error</b>	<b>1</b>	<b>39.94</b>	<b>9.15</b>	<b>&lt;0.001</b>
<b>Task</b>	<b>2</b>	<b>75.80</b>	<b>35.52</b>	<b>&lt;0.001</b>
<b>Order</b>	1	39.94	0.45	0.51
<b>Delay</b>	<b>1</b>	<b>39.51</b>	<b>10.01</b>	<b>&lt;0.01</b>
<b>Recognition Error * Task</b>	<b>2</b>	<b>75.80</b>	<b>7.16</b>	<b>0.001</b>
<b>Recognition Error * Order</b>	1	39.94	0.23	0.64
<b>Task * Order</b>	2	75.80	0.10	0.91
<b>Recognition Error * Delay</b>	1	39.51	2.86	0.10
<b>Task * Delay</b>	<b>2</b>	<b>72.65</b>	<b>5.71</b>	<b>&lt;0.01</b>
<b>Order * Delay</b>	1	39.51	0.15	0.67
<b>Recognition Error * Task * Order</b>	2	75.80	0.97	0.38
<b>Recognition Error * Task * Delay</b>	2	72.65	2.14	0.13
<b>Recognition Error * Order * Delay</b>	1	39.51	0.00	0.95
<b>Task * Order * Delay</b>	2	72.65	2.18	0.12
<b>Recognition Error * Task * Order * Delay</b>	2	72.65	3.31	0.04

Table 35: Total eyes-off-road time from maximum of trials

	NumDF	DenDF	F.value	Pr(>F)
<b>Recognition Error</b>	<b>1</b>	<b>39.87</b>	<b>12.96</b>	<b>&lt;0.001</b>
<b>Task</b>	<b>2</b>	<b>157.73</b>	<b>14.01</b>	<b>&lt;0.001</b>
<b>Order</b>	1	39.87	1.19	0.28
<b>Delay</b>	<b>1</b>	<b>40.15</b>	<b>11.35</b>	<b>&lt;0.01</b>
<b>Recognition Error * Task</b>	2	157.73	2.95	0.06
<b>Recognition Error * Order</b>	1	39.87	0.00	0.95
<b>Task * Order</b>	2	157.73	0.33	0.72
<b>Recognition Error * Delay</b>	1	40.15	2.40	0.13
<b>Task * Delay</b>	2	157.73	3.73	0.03
<b>Order * Delay</b>	1	40.15	0.12	0.73
<b>Recognition Error * Task * Order</b>	2	157.73	2.16	0.12
<b>Recognition Error * Task * Delay</b>	2	157.73	1.74	0.17
<b>Recognition Error * Order * Delay</b>	1	40.15	0.17	0.67
<b>Task * Order * Delay</b>	2	157.73	0.99	0.37
<b>Recognition Error * Task * Order * Delay</b>	2	157.73	0.38	0.68

DOT HS 812 314  
October 2016



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

