Assessing Cognitive Workload of In-Vehicle Voice Control Systems

Chun-Cheng Chang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Linda Ng Boyle, Chair

Thomas Furness

Peter Johnson

Program Authorized to Offer Degree:
Industrial & Systems Engineering

University of Washington

Abstract

Assessing Cognitive Workload of
In-Vehicle Voice Control Systems

Chun-Cheng Chang

Chair of the Supervisory Committee:
Title of Chair Linda Ng Boyle
Department of Chair

In-Vehicle Information Systems (IVIS) are becoming more accessible to drivers and contain more complex communication features. Voice control systems (VCS's) promise to be less distracting than visual-manual interfaces, but they still impose cognitive workload as drivers divert attention to using the voice interface instead of being attentive about safe driving. The goal of this dissertation is determine if the cognitive distraction induced by VCS's can be reliably measured using the Tactile Detection Response (TDRT) protocol. A contextual interview and two driving simulator studies were conducted. Findings from driving simulator studies showed that the TDRT was sensitive to changes in cognitive workload for VCS subtasks such as listening, speaking, visual search, confirmation, and waiting. The TDRT was less reliable in inferring the cogntive workload of VCS interactions containing errors (e.g. recognition error and system timeout). VCS's are multi-faceted interactions that require various cognitive processes complete. Understanding the cogntive workload of each cognitive process provides better guidance on how to improve the designs of VCS interactions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ADRT: Auditory Detection Response Task

DRT: Detection Response Task

EOR: Eyes-Off-Road

EORT: Eyes-Off-Road Time

HDRT: Head-Mounted Detection Response Task

ISO: International Organization for Standardization

IVIS: In-Vehicle Information System

MLP: Mean Lane Position

MGD: Mean Glance Duration

N-BACK: A delayed digit recall task in which participant listen and repeat to a simple auditory stimuli

NADS: National Advance Driving Simulator

NHTSA: National Highway Traffic & Safety Administration

PDT: Peripheral Detection Task

POI: Point of Interest

RDRT: Remote Detection Response Task

RT: Response Time

SDLP: Standard Deviation of Lane Position

SIM: Simulator Study

TDRT OR TDT: Tactile Detection Response Task or Tactile Detection Task

TEORT: Total Eyes-Off-Road Time

VCS: Voice Control System

VDRT: Visual Detection Response Task

WOZ: Wizard of Oz

DEDICATION

This thesis is dedicated to both my parents who have been a constant source of support during the challenges of graduate school and life.

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Linda Ng Boyle for the continuous support of my Ph.D study and related research, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I would like to thank Prof. Tom Furness whose bold enthusiasm towards research and academics made me believe that I could make a difference and inspired me to pursue a Ph.D. I also would like to thank Prof. Peter Johnson and Prof. Sean Munson, for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

This research was supported by NHTSA (contract DTNH22-11-D-00237), therefore I would like to thank Ritchie Huang and Jim Jenness for this research opportunity, along with project collaborators John D. Lee, Vindhya Venkatraman, and Madeline Gibson for their contributions in the experimental designs.

I would like to give special thanks to my labmates Hao Huang, Erika Miller, Edith Guo, Xingwei Wu, Yuqing Wu, Yiyun Peng, and Ben Ries. They helped with laying out the groundwork for my research, improved my work with valuable feedback, and most importantly they supported me with lots of laughter.

Finally I would like to thank all the undergraduate research assistants who helped me run the various driving studies over the years. They are Daniel Ullom, Kayla Fukuda, Sean Yagi, Henry Chang, Nick Di Fulvio, Kristen Higashi, Alice Wu, Stephanie Wang, Yoel Tekle,

Warren Sprecher, and Sean Hoon. This dissertation would not be possible without their contributions.

Chapter 1

INTRODUCTION

In-vehicle voice interfaces are becoming more prevalent as nearly all vehicles made in the U.S after 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 manufactures offer these features with voice control to minimize distraction. Voice-based systems promise less distraction than systems that demand visual-manual interaction, but speech-recognition errors, complex interactions, and response delays might all draw drivers attention away from the road.

Voice control systems (VCS's) have the potential to reduce visual-manual distraction by keeping drivers' eyes on the road and hands on the steering wheel (Putze & Schultz, 2012). This apparent ease of a voice interaction makes it possible to offer drivers more complex features that would not be safely possible with a visual-manual interface. However, in practice, VCS's may require additional button presses and glances toward the in-vehicle display. Even without the button presses and off-road glances, interactions with VCS's may also place additional demands on drivers' attention such that driving performance and safety are compromised.

VCS demands have also been directly linked to driver performance. Lee, Caven, Haake, and Brown (2000) found that drivers' reaction time to a lead vehicle braking event increased by 30% when using a speech based interface compared to a baseline condition. Visual at-

tention decrements have also been reported with the use of VCS. A study conducted by Engstrom, Johansson, and Ostlund (2005) observed that participants' gazes were concentrated in the center of the road during the auditory task, limiting their attentional resources for normal scanning.

A technical working group of the International Organization for Standardization (ISO) has prepared a document on how to best measure distraction caused by secondary tasks such as voice control interaction, and they have selected the Detection Response Task as the method to assess the cognitive load of a secondary task (ISO/DIS 17488, 2015). In a Detection Response Task (DRT), drivers respond to simple frequently occurring targets such light or a vibration. Response time and detection accuracy are the primary metrics. This means longer response time to a target or stimuli, infers higher levels of cognitive distraction for the driver. The ISO committee developed the DRT protocol based on studies that used cognitive tests such as counting and n-back (a working memory number recall task) as a surrogate for a VCS (ISO/DIS 17488, 2015). Few studies have yet to apply the DRT in conjunction with voice control systems. The purpose of this dissertation is to determine if the DRT can be used to assess the cognitive workload of voice control systems. In addition, this dissertation also explores the cognitive workload created by VCS errors along with the cognitive workload of different VCS cognitive processes such as listening and speaking.

Chapter 2 contains literature reviews of driver distractions, voice control systems, and cognitive workload measurements. Chapter 3 describes a VCS contextual interview study that identifies non-optimal user/system interactions that occur for experienced VCS users. Chapter 4 describes a driving simulator study that measures the cognitive workload of VCS interactions using the ISO DRT protocol. Chapter 5 presents a way to characterize and measure cognitive workload for different VCS cognitive processes such as listening and speaking. Chapter 6 describes a second driving simulator study that quantifies the amount of listening, speaking, and reading a driver can engage in before task performances worsens. Chapter 7

summarizes the overall findings in this dissertation and discusses the contributions and possible future research topics.

Chapter 2
BACKGROUND

This chapter will go over the concept of driver distraction, and how a voice control system can be used to minimize some of the effects of driver distraction. Despite the benefits of allowing drivers to keep eyes on the forward road scene, VCS's still impose cognitive distractions. Literature on in-vehicle cognitive distraction will be presented and gaps in literature will be identified.

2.1 Driver Distraction

Driving is a very complex task as it requires concurrent execution of various cognitive, physical, and psycho-motor skills, and yet its not unusual to observe drivers performing non-driving related activities (K. Young & Regan, 2007). The term *driver distraction* implies that drivers do things that are not primarily relevant to the driving task (driving safely) and this reduces the available attention that would otherwise be needed for driving safely (Patten, Kircher, Östlund, Nilsson, & Svenson, 2006). Distraction activities may include a conversation with another passenger, applying makeup, drinking a beverage, or interacting with the In-Vehicle Information System (IVIS).

When drivers are distracted, they often exhibit compensatory behavior. A lot of drivers attempt to reduce workload by either decreasing speed (Alm & Nilsson, 1995; Burns, Parkes, Burton, Smith, & Burch, 2002; Haigney, Taylor, & Westerman, 2000) or increasing distance from a lead vehicle (Jamson, Westerman, Hockey, & Carsten, 2004; Strayer & Drew, 2004). Drivers also adapt to increased distraction by accepting temporary degradation in certain driving tasks like checking the side mirrors and instruments less frequently (Harbluk, Noy,

& Eizenman, 2002; Brookhuis, de Vries, & de Waard, 1991).

Despite the fact that drivers are able to adapt their driving behavior to meet the increased demands of the non-driving task, under certain conditions the adaptive behaviors can breakdown resulting in significant degradation in driving performance (K. Young & Regan, 2007). The degree or magnitude of distraction is a function of complexity of the task, current driving demands, and the skill and experience of the driver. A non-driving secondary task may distract the driver in one situation but not in another (K. Young & Regan, 2007).

There are numerous ways to define, characterize, and categorize driver distraction or inattention. Van Elslande and Fouquet (2007) defined driver inattention in terms of "human functional failures" which may lead to crashes (e.g. "failures to diagnose the situation" or "failures to predict the situation"). Treat (1980) created a driver distraction taxonomy based on why drivers were delayed in their recognition of situations requiring adjustment of speed or path of travel for safe completion of the driving task. Victor, Harkbluk, and Engström (2009) identified five key factors that contribute to crashes due to driver inattention which include saliency, visual eccentricity, shutter vision, cognitive factors, and expectancy.

Based on Wicken's Multiple Resource Theory (Wickens, Gordon, Lee, & Liu, 2004), driver distraction has been categorized in terms of the modality of sensory input which include: visual, manual, and cognitive distraction (Ranney, Harbluk, & Noy, 2005; Liang & Lee, 2010; Engström & Markkula, 2007). Visual distractions are situations where drivers take their eyes off the road. Manual distractions are movements of motor or muscular components, like reaching for a beverage in the cup holder while driving. Finally cognitive distractions cause the driver to take their mind off the task of driving. A voice control system for instance is able to minimize the damaging effects associated with visual distractions, but it still induces cognitive distraction. Driver distraction activities are not constrained to a single distraction type as texting on a phone and driving simultaneously imposes heavy load of visual, manual, and cognitive distraction.

There are many ways to measure driver distractions. Driver distraction exposure (frequency and duration of distraction) can be measured via cross-sectional surveys, roadside studies, and even naturalistic driving studies (McEvoy & Stevenson, 2009). The Lane Change Task (LCT) measures the degradation in driving performance as the result of simultaneously performing a distracting secondary task (Mattes & Hallen, 2009). If a driving simulator is not accessible, the Collision Detection Task is an inexpensive alternative that emulates demands associated with detecting a potential collision situation while driving (Vaux, Ni, Rizzo, Uc, & Andersen, 2010). Performance on this surrogate-driving task provides an assessment of cognitive distraction, as well as visual-manual distraction.

Since driving is primarily a visual-manual task, eyes-off-road glance data has proven to be an adept metric to measure driver distraction (Peng, Boyle, & Hallmark, 2013; Sodhi et al., 2002; Harbluk et al., 2002; NHTSA, 2012). Direct eye-glance data however is difficult to collect and time consuming to analyze. The visual occlusion is an effective surrogate method that simulates eyes-off-road glance behavior (Foley, 2009). The occlusion method consists of systematically obscuring the driver's vision and then removing the obscuration. This can be accomplished by turning a display on and off or by physically blocking the driver's vision intermittently with a screen or similar device such as eye goggles that contain a shutter device, which opens and shuts (Foley, 2009).

2.2 Voice Control Systems

As technology in vehicles advance, the integration of systems with different modalities becomes more prevalent. The use of voice control becomes an important factor in mitigating distraction of complex systems. A Voice Control System or VCS is an interface that uses speech as an input to access the In-Vehicle Information System (IVIS). Current voice control systems have a broad range of functions and features which include:

- Climate control

- Music retrieval (e.g. tuning radio, playing CD track)
- Calling
- Sending SMS messages
- Address entry for navigation systems
- Real time information retrieval (e.g. weather, traffic)

Theoretically, a voice control system requires no physical interface other than microphones for speech inputs and speakers for acoustic outputs. In current practice however, voice control systems generally include hard controls such as a button for voice control activation, and a visual display for information provision.

Studies have shown that voice control interfaces are less distracting than their visual-manual counterpart. Voice control system (VCS) use improved lane keeping and resulted in fewer collisions when compared to visual manual controls when operating in-car entertainment system (Carter & Graham, 2000). Dialing a phone using voice had 22% fewer lane keeping errors and 56% fewer glances away from the road scene when compared to manual dialing (Jenness, Lattanzio, O'Toole, Taylor, & Pax, 2002). When it comes to music retrieval, voice control system (VCS) reduced the total time drivers spent with their eyes off the forward roadway, as well as both the total number of glances away from the forward roadway (Garay-Vega et al., 2010).

While VCS's have clear advantages over visual-manual interfaces, VCS's are not completely distraction free. Using speech to retrieve emails for instance increased reaction time to a lead vehicle braking event by 30% or 310 ms (Lee, Caven, Haake, & Brown, 2001). Voice control was also associated with longer glance duration on the road scene which is a sign of driver distraction, as drivers often fixate on a single point while driving instead of scanning a wider visual field including the periphery. As in-vehicle task demands grow, the driver is often prompted to shorten intermittent glances back to the road scene (perhaps to

reduce working memory load), potentially missing safety-relevant objects and events (Tijerina, Parmer, & Goodman, 1998). A study conducted by Harbluk et al. (2002) confirmed the findings of Tijerina et al. (1998) as hands-free interface such as voice control systems caused drivers to make fewer saccades, spend less time looking at rear mirror, and spend more time looking centrally (Harbluk et al., 2002). Although studies have suggested that VCS's are visually less demanding than visual-manual interfaces, they still impose distraction.

2.3 Cognitive Workload

VCS's minimize visual and manual distractions, but they still evoke cognitive workload on the driver. Cognition is defined as the mechanisms by which people perceive, think, and remember (Wickens et al., 2004). Workload is defined as the amount of information processing available per unit time (Patten et al., 2006). Therefore, cognitive workload in this dissertation is defined as the sum of perception, thought, and memory demands placed upon the individual in a given moment in time. It is important to note that cognitive workload is not only a function of task complexity, but the skills, abilities, and motivation of the individual performing the task also determines the degree or magnitude of cognitive workload.

In this dissertation, the term *cognitive distraction* has the same definition *cognitive workload*, except the term *cognitive distraction* is a subset of *cognitive workload* as it only applies to the domain of distracted driving. The term 'driver distraction' implies that drivers do things that is not relevant to the primary task of driving safely (Patten, Kircher, Östlund, & Nilsson, 2004). Therefore, cognitive distraction is the demand placed on a user by a task and the user's cognitive resources that is not related to driving safely. Throughout this dissertation, *cognitive workload* and *cognitive distraction* will be used interchangeably.

2.4 Measuring Cognitive Workload

Cognitive workload is often measured in terms of self-reporting, physiology, and performance. Self-reporting consists of surveys and questionnaires. Physiology measures are able to capture information about cognitive workload via physiological features such as heart rate and skin conductance. Finally performance measures are metrics associated with the primary and secondary tasks such as lane keeping or break reaction time.

2.4.1 Subjective Measures

Self-reporting measure involve evaluation methods where the study participants report their self-perceived mental workload to complete a task. The Rating Scale Mental Effort (RSME) is a unidimensional survey used to measure subjective mental workload. The RSME consists of a line with a length of 150 mm marked with nine anchor points, each accompanied by a descriptive label indicating a degree of effort (Widyanti, Johnson, & de Waard, 2013). The NASA-TLX is a self-reported survey based on a multidimensional scale. Study participants are asked to complete a survey that is based on a 21 point scale about the mental demand, physical demand, temporal demand, performance, effort, and frustration to complete the given task (Hart & Staveland, 1988).

The issue with self-reporting measures such as the NASA TLX is that they are commonly administered after the completion of a task, since asking the participant to complete the survey in the middle of the experiment can interrupt the task itself. Having participants do a workload survey at the completion of a task greatly limits the number of observations or samples that can be collected. If the survey is administered after the task, people may forget the amount of workload they were feeling during a particular segment of the task if the delay is excessive (Miller, 2001).

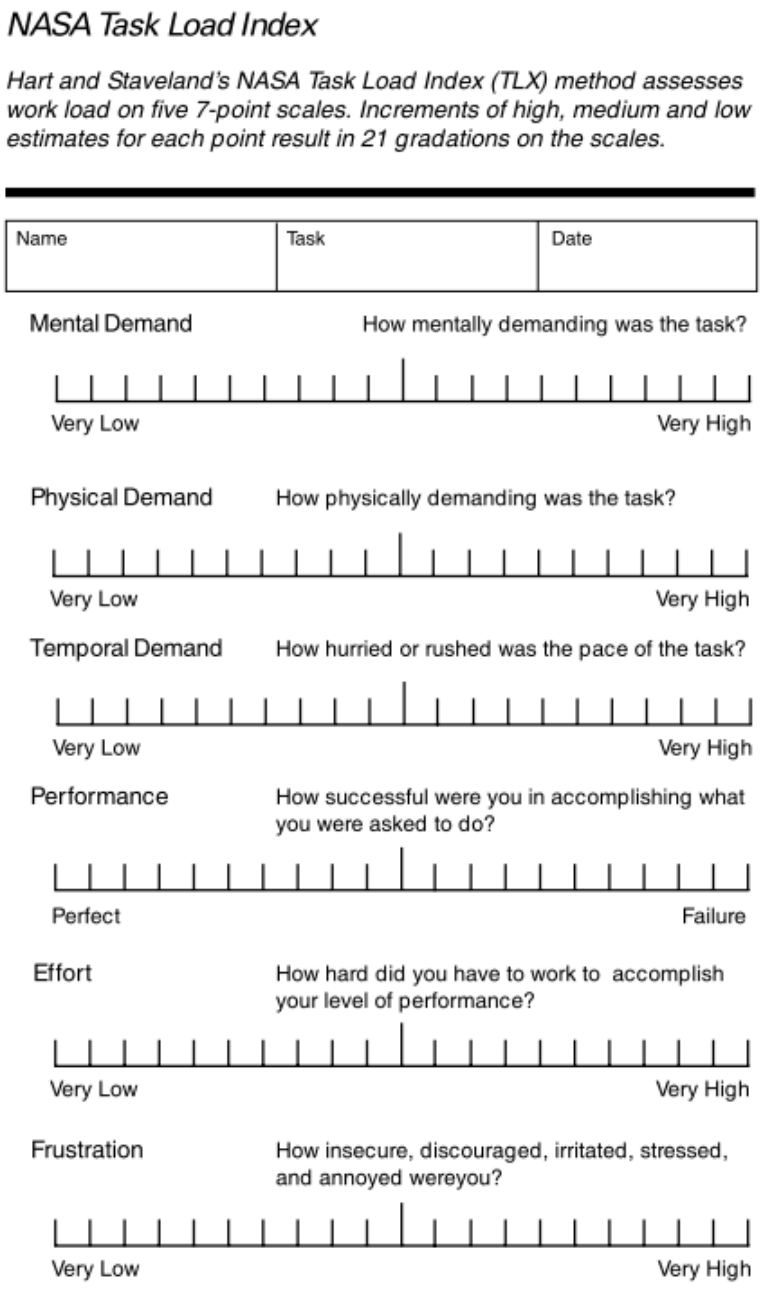


Figure 2.1: Sample of NASA TLX survey

2.4.2 Physiological Measures

The fundamental idea behind physiological measures is that changes in workload can be directly observable given changes in the central or the autonomic nervous system (H. Ursin & Ursin, 1979). Physiological measures have also been shown to be sensitive to subtle increases in demand before overt breakdowns in driving performance (Mehler, Reimer, Coughlin, & Dusek, 2009). Heart rate and skin conductance scale relatively linearly with increasing cognitive demand in a delayed digit recall task (N-back) (Mehler, Reimer, & Coughlin, 2012). Heart rate measurement was also used to observe differences in driver workload among different age groups (Reimer, Mehler, Pohlmeyer, Coughlin, & Dusek, 2006).

Electroencephalography (EEG) measures the recording of electrical activity within the human brain. Changes in the EEG were identified that were reliably associated with levels of cognitive workload (Berka et al., 2004). EEG workload also increased with increasing working memory load and during problem solving, integration of information, analytical reasoning, and may be more reflective of executive functions (Berka et al., 2007). Lin et al. used the EEG to measure cognitive effort, engagement, and workload in a simulated traffic light experiment (2007).

Pupil diameter is also a measure of cognitive load. When people are faced with a challenging task, their pupils dilate. This phenomenon is called Task Evolved Pupillary Response (TERP) (Beatty, 1982). Larger pupil diameters was observed when study participants had to recall more numbers in a digit span recall task (Granholm, Asarnow, Sarkin, & Dykes, 1996). Kun et al. measured the cognitive workload of a speech interaction using pupillometry and observed that filler words such as *"um"* had lower cognitive workload compared to complete phrase utterances (Kun, Palinko, Medenica, & Heeman, 2013).

The advantage of using physiological measures is that data can be collected continuously without ever interrupting the participant in the middle of the task. Using a an electro cardiogram (ECG) to measure heart rate will not produce a shortage of data points. The

downside to physiological measures such as ECG is that they are relatively expensive and the data it generates can be noisy. In addition, using ECG can be intrusive as they require electrodes and conductance gels to be placed on the body. EEG for instance requires the participant to wear a cap full of electrodes.

2.4.3 Performance Measures

A third class of cognitive workload measurement is to analyze primary and secondary task (dual task) performances. In the distracted driving research domain, the primary task is driving the vehicle safely, while the secondary task would be to perform something distracting like using an In-Vehicle Information System (IVIS). The effect of the distracting secondary task would be used to evaluate the performance in the primary task, as well as providing some insight into the limitation of human operators (Merat et al., 2007). Dual task studies are based on the assumption that if two tasks share the same working memory resource, performance in one or both task deteriorates when two task are done together, compared when each task is done alone (Merat et al., 2007).

Performance measurements for the primary task are associated with metrics ascribed to the highest degree of priority (Levy & Pashler, 2008). In a driving study, primary task performance measurements may include driving speed, lane maintenance or braking reaction times. Primary tasks measurements however do not provide indication what cognitive workload might be. Under the dual task setting, impairment in one task (e.g. lane maintenance) is an indication of workload imposed by another task (e.g. using an IVIS) (Vilimek, Schäfer, & Keinath, 2013).

2.4.4 Detection Response Tasks

The Detection Response Task (DRT) is a cognitive workload performance measure based on the dual task methodology. In a Detection Response Task, the participant is asked to drive a

vehicle (primary task) and to respond to a frequent and randomly occurring stimuli as quickly as possible (secondary task). The ISO standards (ISO/DIS 17488, 2015) have defined that the stimuli appears randomly every 3-5 seconds and the stimuli can be either visual (LEDs), tactile (vibration motor), or acoustic (blip). The participant responds to the stimuli by pressing a micro-switch button. Based on the reaction time to the stimuli, conclusions about the visual and/or mental workload can be drawn as DRT reaction time and miss rates will be elevated with increasing cognitive workloads (Harbluk, Burns, Hernandez, Tam, & Glazduri, 2013). NHTSA has recommended the use of a Detection Response Task over the Lane Change Task, due to its sensitivity to changes in the cognitive load for auditory/vocal memory-scanning task (Ranney, Baldwin, Vasko, & Mazzae, 2009). The DRT protocol has also been used in a tertiary task setting (Harbluk et al., 2013; Strayer, Turrill, Coleman, Ortiz, & Cooper, 2014; Jahn, Oehme, Krems, & Gelau, 2005). In a tertiary task setting, the primary task is driving, the secondary task is operating the IVIS, and the tertiary task is responding to the DRT stimuli.

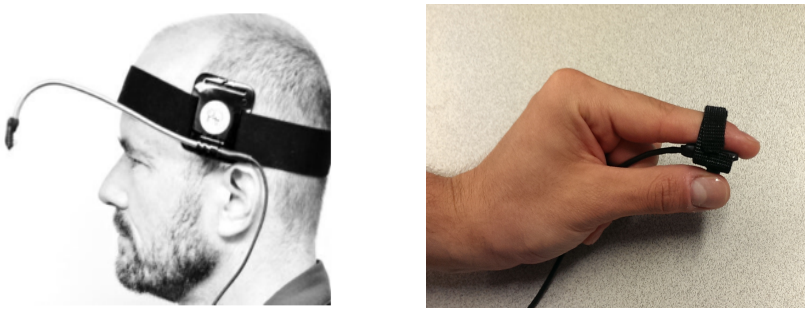
Visual Detection Response Task (VDRT)

The Visual Detection Task (VDRT) is a type of Detection Response Task that was developed based on the idea that functional visual field decreases with increasing workload (van Winsum, Martens, & Herland, 1999). In a Visual Detection Task, the subject responds to a visual stimulus like a LED light by pressing a micro-switch attached to the index finger. The VDRT can be made more complex by changing the location where the LED light appears or by having the participant simultaneously respond to a light and make a decision of some kind.

There are two variants of the Visual Detection Response Task. The first one is the Remote Detection Response Task (RDRT), where the LED stimuli signal located in the peripheral field of view (van Winsum et al., 1999). The LED is often reflected off the windshield

or presented graphically on the simulator screen. The issue with the Remote Detection Response Task (RDRT) is that visual eccentricity varies with eye and head movement. The RDRT does not tell with certainty whether the effect was caused by looking away or by internal interference (Rupp, 2011). Another issue with the RDRT is that when its applied to the field, the LED reflected on the windshield strongly depends on the lighting condition and background contrast.

The second variant of the VDRT is called the Head Mounted Detection Response Task (HDRT) (see Figure 2.2a) . The HDRT solves some of the problems caused by visual eccentricity and non optimal lighting conditions, by fixing the LED to the head with a head band (Victor, Engström, & Harbluk, 2009; Harbluk et al., 2013). The Head Mounted Detection Response Task however increases intrusiveness as the head band may also interfere with video-based eye tracking systems (Rupp, 2011).



(a) Head Mounted DRT (b) Microswitch Button

Figure 2.2: Head Mounted DRT (Cooper, Ingebretsen, & Strayer, 2014) with microswitch button

Auditory Detection Response Task (ADRT)

The Auditory DRT uses a sound stimuli, instead of light, thus it can also resolve the problem of visual eccentricity, and it is also less intrusive than the Head Mounted Detection

Response Task. Despite being able to resolve visual eccentricity and equipment intrusiveness, the ADRT is less useful in driving studies as the sound stimuli can be masked by the background noise from driving. Furthermore, the ADRT is will frequently interfere with a voice interface as the participant needs to listen or speak to the system. While ADRT has not been extensively evaluated and studied when compared to the visual counterparts, Vilimek found the ADRT to be sensitive to increases in cognitive workloads (Vilimek et al., 2013).

Tactile Detection Response Task (TDRT)

In a Tactile Detection Response Task, a tactile or vibration is used in place of a light or sound stimuli. This is sometimes called Tactile Detection Task (TDT), but for this dissertation, the term TDRT will be used. A small vibrating tactor is often taped to the side of neck or wrist and it also resolves the problems of visual eccentricity, non-optimal lighting condition, intrusive equipment, and risk having a sound stimuli being masked by background noise.

In addition to resolving environmental issues and equipment intrusiveness associated with other DRT variants, there is evidence that the TDRT is more sensitive to cognitive task load such as counting and answering a yes/no question (Engström, Aberg, Johansson, & Hammarback, 2005). According to a study conducted by Ranney, Baldwin, Smith, Mazzae, and Pierce, the TDRT had the highest level of test-retest reliability when compared to the Remote and Head Mounted DRT (2014). The TDRT was the only DRT variant that was sensitive between a baseline task, 0-back, and 1-back when compared to the Remote and Head Mounted DRT (Harbluk et al., 2013). One practical issue with the TDRT concerns with the placement of the vibrating tactor. The TDRT tactor has been proposed to be placed on the wrist or side of the neck. Taping the tactor in the wrist can create interference from wiring. Taping the tactor on the side of the neck can create interference due to talking-induced sound vibrations (Rupp, 2011).

Table 2.1: Examples of different cognitive workload measurement methodologies

Measure	Result of Incresed Workload	Benefits	Drawbacks
Survey			
Rating Scale of Mental Effort (RSME)	Higher Rating	Easy to administer, inexpensive, quick	Not continuous data, not as sensitive as multidimensional scales
NASA-TLX	Higher Rating	Easy to administer, inexpensive	Not continuous data
Physiological			
Heart Rate (ECG)	Increases	Continuous data	Not reliable, intrusive, expensive
Electroencephalogram (EEG)	Alpha waves replaced by Beta waves	Continuous data	Very intrusive, expensive
Perfromance (DRT)			
Remote Detection Response Task (RDRT)	Increased reaction time	Continuous data, inexpensive, not intrusive	Need optimal lighting condition, visual eccentricity
Head Mounted Detection Response Task (HDRT)	Increased reaction time	Continuous data, inexpensive, resolves visual eccentricity	Intrusive
Auditory Detection Response Task (ADRT)	Increased reaction time	Continuous data, inexpensive, resolves visual eccentricity	Sound stimuli can be masked by ambient noise
Tactile Detection Response Task (TDRT)	Increased reaction time	Continuous data, inexpensive, resolves visual eccentricity, not impacted by external environment	Uncertain regarding placement of tactor

2.5 Gaps in Literature

The impact of VCS recognition errors on cognitive workload has yet to be assessed using the DRT. VCS are imperfect as ambient noise, acoustic similarity of commands, and length of spoken word can all undermine recognition accuracy (Gellatly & Dingus, 1998). Poor speech recognition accuracy can lead decrements in driving performance. Kun et al. observed larger steering wheel angle variance with low speech recognition accuracy (2007). Higher number of collisions was observed in a lead vehicle driving simulator study when a VCS operated at 56% accuracy (McCallum, Campbell, Richman, Brown, & Wiese, 2004). Slow system response times along with poor recognition accuracy can push users towards manual input

(Ginosar & Hearst, 2014). Given the adverse effects of speech recognition errors, it is vital to quantify the cognitive distraction of VCS errors.

The DRT protocol has shown to be sensitive varying demands of cognitive workload, but few studies have applied the TDRT to explore age related cognitive deficits. Older adult drivers for example have more trouble multitasking as they showed a significantly decreased ability to divide attention in a dual task driving simulator study (Brouwer, Waterink, Wolffelaar, & Rothengatter, 1991). Stemlach et al. observed that older adults executed limb movements with less speed and precision (Stelmach & Nahom, 1992). Cerella surveyed numerous reaction-time literature, and concluded that the more complex tasks result in greater performance deficits for the elderly (Cerella, Poon, & Williams, 1980).

Numerous studies were able to demonstrate the DRT's sensitivity to increasing cognitive loads for secondary driving tasks as a way to assess in-vehicle interfaces. However, most studies used counting and a number recall task such as the n-back task as surrogates for VCS interaction (Ranney et al., 2014; Bengler, Kohlmann, & Lange, 2012; Merat et al., 2007; Engström, Aberg, et al., 2005; R. A. Young, Hsieh, & Seaman, 2013; Schindhelm & Schmidt, 2015, 7; Bengler et al., 2012). VCS interactions are complex, as they are typically multi-modal interfaces that provide both visual feedback and audio feedback. In addition, VCS's can make recognition errors causing task resets. Performing mental arithmetic with a counting task or storing two to three single digit numbers in short term memory for the n-back task may not be an accurate proxy for an actual VCS interaction.

There are a few studies that applied the DRT to evaluate the cognitive workload of commercially available VCS (Harbluk et al., 2013; Strayer et al., 2014; Strayer, Cooper, Turrill, Coleman, & Hopman, 2015). The bulk of these studies however compared the performance across systems (e.g. Chevy Equinox vs VW Passat) or between tasks (e.g. radio vs navigation). For example, Strayer et al. concluded that the VCS for the Chevy Equinox had lower cognitive workload when compared to VW Passat. However, data analysis aggregated at the

system level (e.g. cognitive workload of Chevy Equinox) does not explain *why* a VCS has higher cognitive workload when compared to another. Cognitive workload analysis should not just be performed at the system level or task level, but also at the subtask level. VCS interactions are complex, and understanding the cognitive workload at the subtask level provides better guidance where design improvements can be made.

2.6 Study Objectives and Specific Aims

VCS's are multifaceted interactions that consists of diverse cognitive processes such as listening, speaking, and reading. A VCS interaction can be further complicated with errors such as mis-recognitions or system time outs. The main objective of this dissertation is to understand the cognitive workload associated with in-vehicle VCS use while accounting for complexities such as VCS errors, and also accounting for the cognitive processes that make up a VCS interaction. The following Specific Aims define common themes found in an on-road observational study, measure the cognitive workload of a VCS interaction, and present novel ways to analyze and assess VCS's.

Aim 1: *Identify drivers' use patterns with voice control systems (VCS's) that exist in current automobiles and to document any usability issues with VCS's that may affect driving safety.* An on-road contextual interview was conducted with 64 participants across two study sites. Participants were asked to use their own VCS in their own vehicle, while video and audio was recorded. This qualitative study was designed to document the types of non-optimal user/system interactions that occur for experienced VCS users.

Aim 2: *Assess the cognitive workload of non-optimal VCS interactions.* A driving simulator study with 48 participants was conducted to assess the cognitive workload of VCS interactions with recognition errors and system time outs.

Aim 3: *Examine the cognitive workload of VCS subtask.* VCS's are multifaceted interaction where the user can be performing cognitive processes such as listening, speaking, and

reading all within the same task. The objective is to determine whether a VCS interaction can be characterized by its cognitive processes (e.g. listening, speaking, waiting).

Aim 4: *Testing the limits VCS subtasks.* A second driving simulator study was conducted with 32 participants to determine the amount of speaking, listening, and reading the driver can engage in before task performance deteriorates and cognitive workload drastically elevates.

Chapter 3

VCS CONTEXTUAL INTERVIEW STUDY

3.1 Introduction

The goal of the contextual interview study was to identify drivers' use patterns with voice control systems (VCS's) that exist in current automobiles and to document any usability issues with VCS's that may affect driving safety. 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. It was not designed to evaluate any specific VCS, nor was it designed to directly compare performance of different VCS's. The VCS's that were evaluated included OEM (VCS embedded in-vehicle) and 3rd party nomadic devices (primarily cell phones) connected to the vehicle or operating independently within the vehicle. The findings of this research was designed to inform development of VCS evaluation methods and development of NHTSA guidelines for driver distraction.

3.2 Study Design

Drivers in Seattle, WA and Rockville, MD 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's. A large part of the video recorded interview consisted of having the participant demonstrate how he or she typically uses their 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 (Beyer & Holtzblatt, 1997).

3.3 Participants

A total of 64 drivers (ages 19 to 65) were interviewed for this study. The participants included 34 women and 30 men. All held a valid U.S. driver's license 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, Maryland. Participants in Washington were recruited via Craigslist, and by posting flyers on bulletin boards around King County, Washington. 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.

3.4 Driving Route

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.

Maryland Driving Route

The on-road driving portion of the interview was approximately 30 minutes and two driving routes were used (Figure 3.1). The routes included a mixture of driving on a highway,

arterial streets, 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.

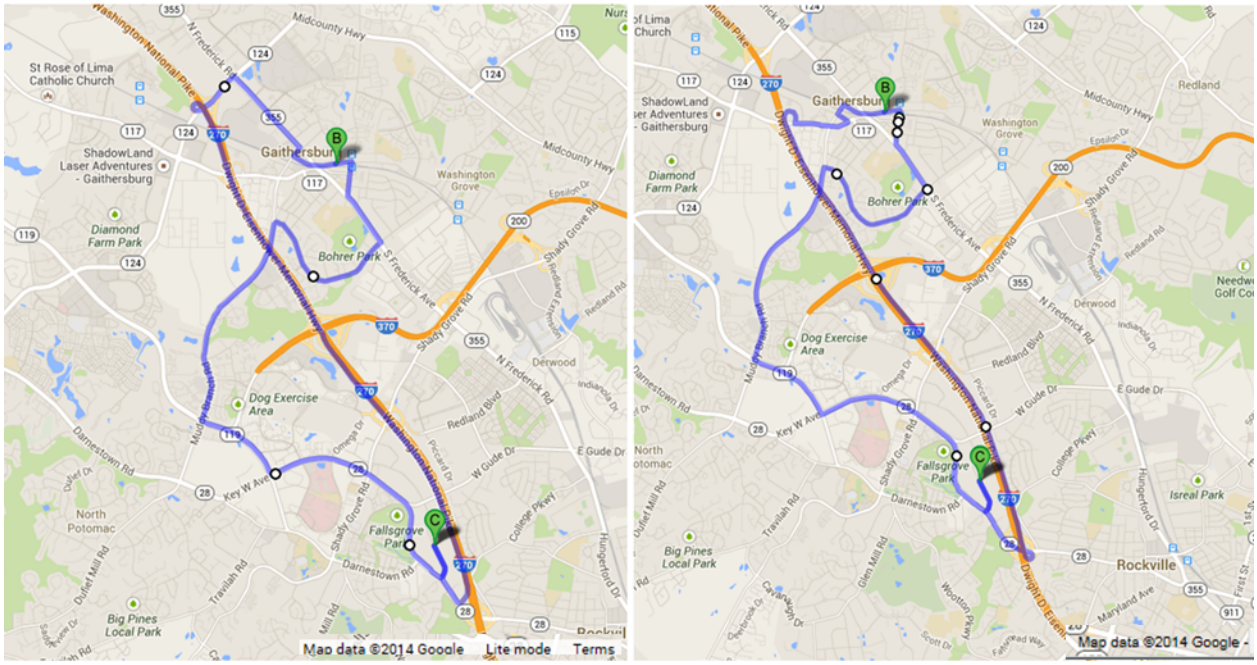


Figure 3.1: Maryland driving routes. Route 1 (left) and Route 2 (right)

Washington Driving Route

The route in Washington (Figure 3.2) was 13.2 miles long and took approximately 30 minutes to complete in light traffic. Participants were randomly assigned 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.

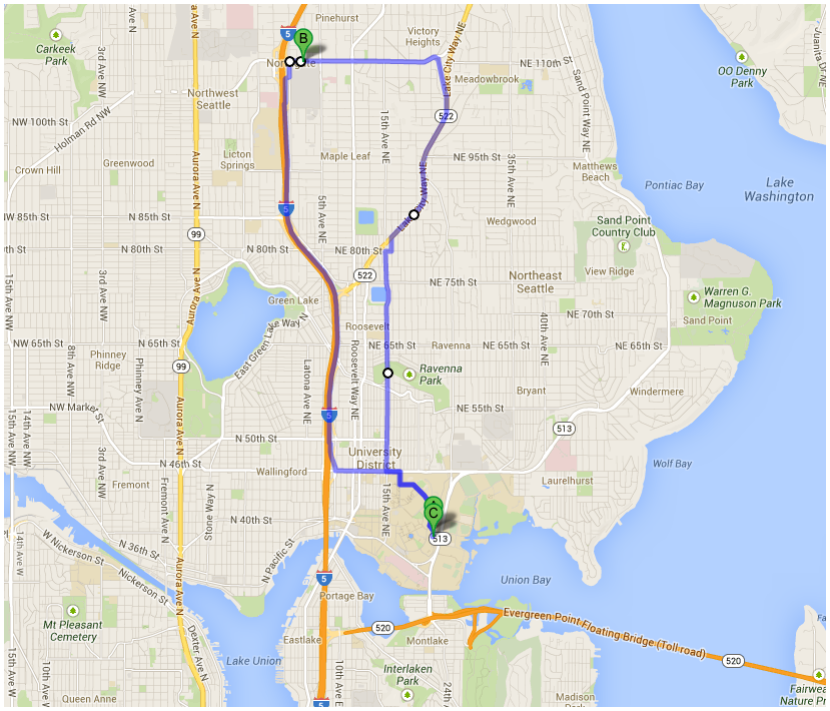


Figure 3.2: Washington driving route

3.5 Instrumentation

At the beginning of each contextual interview session, two battery-operated video cameras were mounted inside 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. The researcher and participant each wore a lavaliere microphone clipped onto their shirt, as it allowed for clear audio recording of all interactions with the system and between the researcher and participant.

Figure 3.3: Composite camera views of the steering wheel, dashboard, forward roadway, and participant's face

3.6 Procedure

Each participant completed an informed consent 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 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 setup 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 researcher reviewed the informed consent form with the participant. The participant was shown a map of the route to give them an idea of where they would be driving. While the researcher installed the video cameras in the participant's vehicle, the participant filled out a voice control use questionnaire. 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. While the vehicle was still parked, the researcher asked the participant to demonstrate a few of the voice control tasks. The researcher 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 whenever they were comfortable, and in driving situations when they would most likely perform the tasks in real life.

To summarize: Participants only performed VCS tasks that they said they were familiar with 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. Additionally, the researcher asked the participant about VCS use tactics, errors encountered, and about their experience with the system in general.

3.7 Data Reduction

Videos recorded from the GoPro cameras were combined into a single composite video suitable for video coding as shown in Figure 3.3. A video coding software *Morae* was used to code errors participants made during the task. The coding markers included start and end of a task, task type, task error, successful task completion, and number of attempts to complete the task.

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 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.

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 my repeating the original command. There were four types of tasks commonly performed by drivers when interacting which are as follows:

Communication Participants used voice to communicate via phone or by text messages to an individual outside the vehicle. Participants might make a phone call using their list of contacts, check voice messages, or to send or check text messages.

Navigation Participants used voice to interface with a navigation system to find a route to a specific address or a point of interest.

Information Participants used voice to obtain the status of road and weather conditions. Command example: "Check weather" or "Check traffic".

Entertainment Participants used voice to control the entertainment media, to play music or radio, change or end song, or adjust volume. Command Example: "Play FM radio", or "Play next song".

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". The list below describes error each type:

Clarification System has "misrecognition" and asks the driver to clarify what was said. This requires the participant to respond with some sort of open-ended input. For example, the system might state, "please repeat".

Premature Speaking Participant speaks before prompted by the system. E.g. hits the button and speaks immediately instead of waiting for the tone to sound or speaks before hitting the button at all.

Read Off Option System has "misrecognition" and provides a list of alternative/potential commands from which the driver can select. Requires the driver to respond with a "canned" specific option. For example, "Sorry, there is a list of POI [point of interest] categories to choose from..."

System Time Out Participant engages the voice system but does not provide feedback in the provided timeframe (not quick enough). This may result in the system saying something such as, "please say a command".

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.

E.g., participant tries to perform a navigation task while still in the phone menu and the system will not execute.

Wrong Task System has a "misrecognition" and executes the wrong task. For example, driver wants to navigate to a certain radio station and the system leads them to a different station.

In summary, data coders marked the beginning and end of a task to determine the duration of the voice interaction. Coders also marked the type of voice task (e.g. Communication, Navigation, Information, and Entertainment). If an error occurred during the voice interaction, the coder marked the error type (e.g. Wrong Task, System Time Out, Clarification).

3.8 Results

3.8.1 Quantitative Description

Study participants used a variety of VCS's which include *OEM Installed*, *Smartphone*, and *Smartphone Hybrid*. Table 3.1 lists the number of each type of VCS's at each study site. *OEM Installed* are VCS's that was integrated with the vehicle. The systems represented in this study include Ford SYNC, UConnect, Lexus Premium Total Technology Package, Entune, BlueLink, Infiniti System, Nissan, BMW I - System, Honda Hands- Free BlueTooth, AcuraLink, Tesla Model S, Subaru, VW System, and Onstar. For these systems, participant would initiate voice task by pressing the push-to-talk button on the steering wheel.

Smartphone refers to devices such as Apple iPhone or Android phones which have voice control capabilities via software systems such as Siri or Google Now. Participants using this system setup would place their device within reaching distance (e.g. in hand, cupholder, phone holder, or lap).

Smartphone Hybrid are smartphones that are connected to the vehicle via AUX cable or Bluetooth. The *Smartphone Hybrid* setup allows audio to be projected through the speakers

of the car instead of the phone, and participant can often initiate voice task interaction via the push-to-talk button on the steering wheel, instead of pressing a button on the smartphone.

Table 3.1: VCS's used by participants

Site	OEM Installed	Smartphone	Smartphone Hybrid
WA	10	6	16
MD	27	0	5

The Seattle, WA study site had more participants with *Smartphone* and *Smartphone Hyrbrid* VCS setup, while Rockville, MD study site had more participants with *OEM* embedded system.

Numerous errors occurred during VCS interaction. Table 3.2 and Table 3.3 show the four tasks commonly performed by the participants segmented by completion status. That is, whether they were completed with some errors (took more than one attempt to complete the task) or completed error-free. If the voice control device executed the driver's command on the very first attempt, the task completion status was classified as "without error." Participants in Seattle, WA had lower error rates than those in Rockville, MD. At both sites, navigation and communication tasks resulted in a higher error rate than information and entertainment tasks. Information tasks such as checking weather status had the smallest percent errors.

Navigation and Communication tasks had higher error rates as these tasks are typically more complex interaction. Navigation tasks for instance often require participants to state a house number, street name, along with city and state. This adds to the task complexity and lowers the likelihood to complete the task on first attempt. Figure 3.4 shows the types of errors that can occur. The majority of errors in the Washington site were *System Time Out* errors. Most participants recruited in Washington used a smartphones or hybrids.

Table 3.2: Task completion status in Seattle, WA

Task Type	W/error	Error-free	Total	%Error
Communication	25	54	79	37%
Navigation	29	66	95	31%
Information	7	36	43	16%
Entertainment	10	28	38	26%

Table 3.3: Task completion status in Rockville, MD

Task Type	W/error	Error-free	Total	%Error
Communication	37	34	71	52%
Navigation	35	32	67	52%
Information	12	19	31	39%
Entertainment	31	46	77	40%

Based on the contextual interviews *System Time Outs* can occur when there is poor internet connection, Bluetooth disconnects, failure to detect voice command input, or database synchronization problems. The Rockville, Maryland study site experienced a lot of *Clarification* and *Wrong Task* errors. Since 27 out of 32 participants used OEM installed VCS, the relative high frequency of *Clarification* and *Wrong Task* errors suggest that the voice recognition accuracy of OEM are not as good as smartphone devices.

The presence of errors increases VCS interaction time. Figure 3.5 compares mean VCS interaction time when errors are present and errors are absent. Participants tended to spend most of their time performing navigation related tasks. Navigation tasks took an average of 33.3 seconds when no errors were encountered, but took an average of 51.0 seconds for all

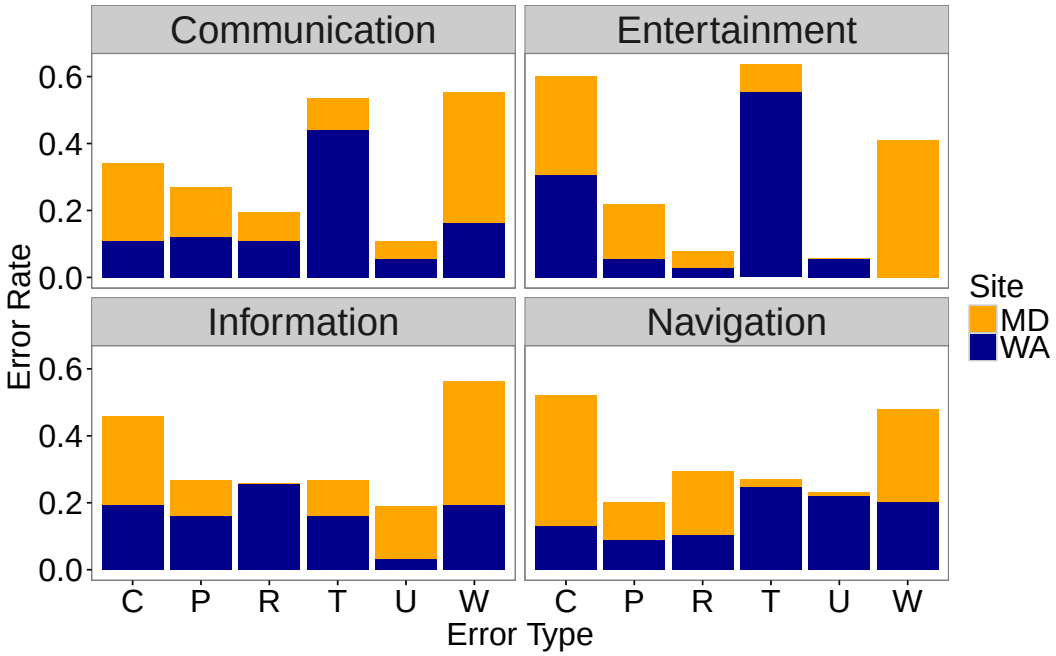


Figure 3.4: The six errors identified for each task (C: Clarification, P: Premature, R: Read off Option, T: Time out, U: Uncertainty, W: Wrong task)

observed navigation task interactions. By contrast, A/C related tasks had a mean of only 11.2 seconds without errors and 14.7 seconds for all observed A/C interactions.

3.8.2 Qualitative Summary

It should be noted that many different VCS's were used and many different VCS-enabled tasks were demonstrated. This study documented both the variety and commonality of user experiences with VCS's while driving. Experimenters' notes from both the Washington and Maryland site were analyzed to identify common themes that related to the behavior of participants and usability of VCS. Most themes relate to use of both embedded VCS's as well as VCS's on smartphones.

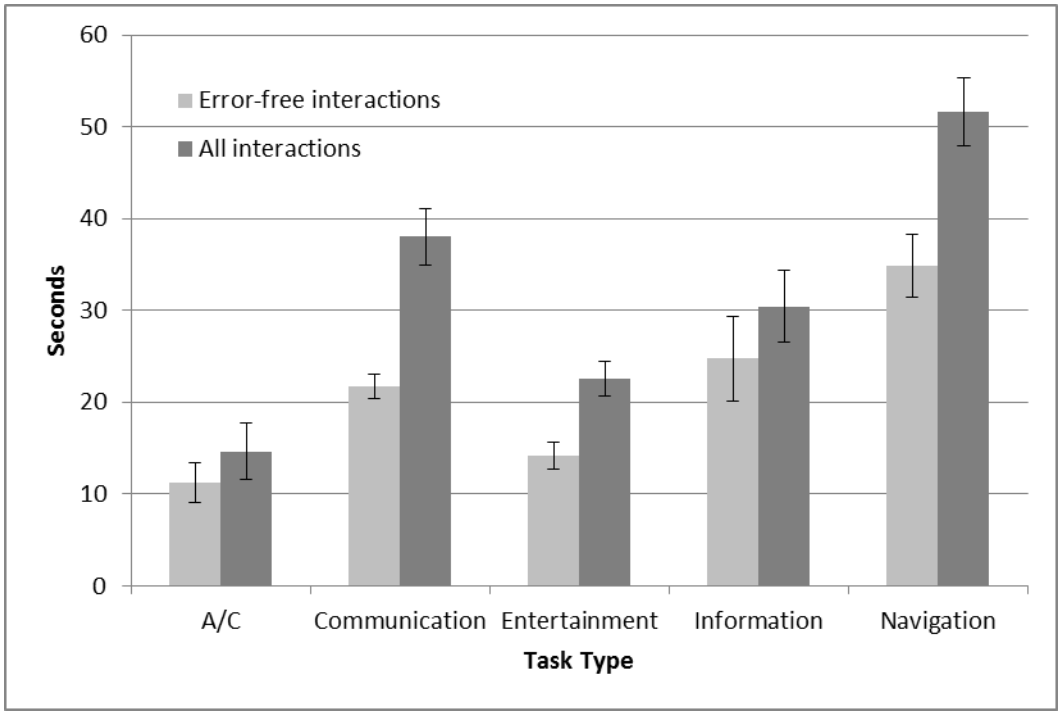


Figure 3.5: Mean Interaction Time by Task Type

Theme 1 - Anthropomorphism of VCS

- It was common for users to reference their VCS's as another human being. This seems to influence their mental model of system function.
- Users have a need for immediate and frequent system feedback in their interaction with VCS's, as if in a conversation.
- One user said that he needed to speak loudly and clearly to the system, "almost as if you are talking to an idiot."
- Other participants said things like, "Sometimes she is stubborn," "She doesn't listen very well," and "What's wrong with you? Bad Navi!"
- Throughout the drive, one participant kept referring to her VCS as "Isabella," and

when the system didn't work properly, she said that this behavior was, "One of Isabella's moods."

Theme 2 - Users' expectations for system performance were modest

- "Voice systems are still in their infancy."
- "If there is a technology and it is hard to use it, I use it anyway."
- "The errors that the system makes are to be expected. This is a computer after all."

Theme 3 - Users tend to blame themselves for non-optimal user/system interactions

- "I have an accent."
- "I didn't say that correctly."
- "That could be human error where, you know, I may not say it clearly enough."
- "It was my fault. I need to get out of USB mode boefore going into directions."
- "It's a robot, you should accept mistakes and try again."
- "I'm sure by the time I get my next car, the system will be much more advanced."

Theme 4 - System performance seems off when compared to other days

- "This isn't the way it normally works. This typically works pretty smoothly but the system is crashing today."
- "Maybe your [data collection] equipment is screwing with my Navi's head. This [poor performance] is unusual."
- May indicate that users' perception of system performance exceeds actual performance

The various themes demonstrates that VCS performance is less than ideal. Despite non-optimal VCS performance, study participants were not completely discouraged from using VCS. Some participants humanized their VCS and gave it a name. Others put up and accepted VCS performance limitations. Majority of study participants use VCS while driving because they believe VCS is safer to use than a visual manual interface.

3.9 Discussion

Using VCS's to call a friend or enter an address into a GPS navigation system are popular and useful features to perform while driving, yet the results from the contextual interviews show that VCS errors are common. Over 50% of Communication and Navigation interactions had one or more errors in Rockville, MD study site. Seattle, WA had lower error rates, but 30% of Communication and Navigation interactions had one or more error. A possible explanation for the difference in error rates between Rockville and Seattle is that participants recruited in the Seattle study site used smartphone technologies which are capable of processing natural speech. The VCS embedded with the OEM often require a very specific voice command. The qualitative responses also revealed some of the frustration and limitations associated with using a VCS while driving.

While VCS's tend to be safer than visual manual interfaces (Barón & Green, 2006; Carter & Graham, 2000), they are not distraction free. The presence of errors extends the task duration, which means drivers spend more time allocating cognitive resources using the VCS's instead of focusing on driving safely. Lee et al. observed higher brake reaction time when driver was composing a message using voice (2001). Lane departures were higher compared to a baseline condition when drivers were entering an address using speech (Tsimhoni, Smith, & Green, 2004). Lane variance and steering wheel angles increased when with poor voice recognition accuracy (Kun, Paek, & Medenica, 2007).

A limitation of the study is that the errors were not linked or attributed to a specific VCS.

The Maryland study site had over 50% error rates for Communication and Navigation tasks (Table 3.3). The results doesn't explain if the majority of errors are caused by a VCS made by a specific OEM (e.g. Ford SYNC) or all the OEM's had equally poor VCS's. Another study limitation is the results also does not differentiate between VCS that are capable of processing natural speech (e.g. "Go to 120 Main St" and "Navigate to 120 Main St" both work), or require specific commands (e.g. "Navigate to..." is the only command that works). Understanding the underlying technology behind the VCS better helps explain the causes of frequent error.

Despite the study limitations, the contextual interview demonstrated that VCS's which are commonly available in the market have performance issues as numerous interactions result in some kind of error. The VCS's that are embedded with the vehicle (OEM installed) often have poor recognition accuracy since they contain dated technologies. Smartphones on the other-hand have updated voice recognition technology, but they are often not designed and optimized for in-vehicle use. The frequency of errors from VCS interaction create cognitive distractions as the driver is allocating cognitive resources towards the interface instead of driving safely.

3.10 Chapter Summary

A contextual interview study was conducted in Rockville, MD and Seattle, WA in order to identify patterns of VCS use while driving. In the contextual interview, study participants drove a predetermined route, using their own vehicle, and operated their own VCS. A researcher who rode along with the participant would provide navigation directions and asked the participant to perform a VCS task. Video and audio was recorded, and the researcher made notes on the participants interaction with the VCS.

Study results showed that errors frequently occurred with VCS interactions as participants were unable to complete the task on their first attempt. Over 50% and 30% of

navigation task resulted with one or more error in Rockville, MD and Seattle, WA respectively. VCS's frequently experienced recognition errors and at times provided no feedback to a prompt. Participants commented that the VCS can be frustrating to use and acknowledged the limitations of the technology.

Since VCS interactions from the contextual interview frequently included errors, additional research is needed in order to understand the impact of VCS errors on cognitive distraction. While VCS's allow drivers to keep their eyes on the forward roadway, drivers may still thinking about how to resolve the VCS error instead of focusing on driving safely. The following chapter explores the impact of VCS errors on cognitive workload in a driving simulator study.

Chapter 4

VCS DRIVING SIMULATOR STUDY ONE - COGNITIVE WORKLOAD OF VCS ERRORS

4.1 Introduction

The overall goal of this chapter is to determine if VCS errors elevates cognitive distractions while driving. While VCS allow the drivers to keep their hands on the steering wheel and eyes on the road while interacting with an in-vehicle interface, they can still induce cognitive distractions where driving performance can be compromised (Lee et al., 2001; Tsimhoni et al., 2004). This problem is further compounded by the frequency of performance errors such as mis-recognitions and system timeouts observed in the contextual interview study in Chapter 3. Poor speech recognition accuracy for instance can reduce driving performance (Kun et al., 2007; McCallum et al., 2004).

A driving simulator study was conducted to measure the cognitive workload of VCS use. Drivers interacted with a VCS that was prone to frequent recognition and system timeout errors. Cognitive workload was measured using the Detection Response Task (DRT) protocol (ISO/DIS 17488, 2015), which has been extensively used to measure cognitive workload of a VCS (Engström, 2010; Ranney et al., 2014; Bengler et al., 2012; Merat et al., 2007). Most studies however used the n-back task or counting task as a surrogate for a VCS interaction. This study simulated more applicable voice tasks that are often performed while driving.

There are two main research question in this chapter. The first is to determine if the DRT protocol is sensitive to changes in cognitive workload for a VCS interaction. The second research question is to determine if VCS errors elevates cognitive workload. To answer the research questions, this driving simulator study tests the following hypothesis:

1. Is there a difference in TDRT performance among different types of voice tasks such as radio channel selection, address navigation, and calendar appointment scheduling. (H1)
2. Does inducing poor voice recognition accuracy and system timeouts increase TDRT response times and miss rates. (H2)

4.2 Methods

The experiment was based on a driving simulator study with a secondary task. The study participant is expected to multi-task among the following:

1. Drive a vehicle using the driving simulator
2. Perform a voice task
3. Respond to a randomly occurring tactile stimuli

The primary task is driving the vehicle, while secondary task is interacting with the voice control system (VCS). Responding to a randomly occurring stimuli (TDRT) is the tertiary task that allows us to infer the cognitive workload of the secondary task. This chapter will go over the participant recruitment criteria, describe the hardware used for the experiment, and detail the experimental procedures along with the study design.

4.3 Participant Recruitment

There were 48 participants recruited for the study (24 males, 24 females). The mean age of the participants was 38 years old and Table 4.1 shows the distribution of participants across different age groups and gender.

The study was approved through the UW Internal Review Board (#45851). Participants were recruited for the study via emails, flyers, and online classified ads. In order to be qualified for the study, participants also had to fulfill the following inclusion criteria:

Table 4.1: Number of Participants across Age Group and Gender

	Male	Female
Age 18-24	6	6
Age 25-39	6	6
Age 40-54	6	6
Age 55-75	6	6

- Be in good general health (no heart condition, seizure, epilepsy, Méniére's Disease, or narcolepsy)
- Be in the age range of 18-75 years of age, inclusive
- Be an active driver with a valid US driver's license
- Drive a minimum of 3,000 miles per year
- Be comfortable using computer, touchscreen, and using voice control systems
- Be comfortable communicating via text messages (SMS), voice input, keypad input, or a combination of both
- No participation in any driving simulator studies in the past 6 months
- Be a native English speaker

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 identified themselves as having a high likelihood of experiencing simulator sickness were excluded from being eligible in the study. The recruiting strategy followed the 2012 NHTSA Visual Manual guidelines (NHTSA, 2012).

4.4 Study Setup

During the experiment, participants were required to drive a car using the simulator, issue a voice command, and respond to a tactile stimuli. The *Study Setup* section covers the hardware and software used for the experiment.

4.4.1 Driving Simulator

The driving portion of the study was handled by the NADS MiniSim, which is a fixed based driving simulator. The NADS MiniSim includes three projection screens (3.0' (wide) by 1.7' (tall) each) that are placed about 4.5' away from the driver's eye point.

Driving Scenario

The simulated environment had the following parameters:

- Four lanes, undivided
- A solid double yellow line down the center, solid white lines on the outside edges, and dashed white lines separating the two lines in the same direction
- Flat, straight road (no horizontal curves)

Drivers were asked to follow a lead vehicle and maintain a two second headway at 50 mph (80 km/h). The driving scenario did not include any lead vehicle braking events, but the lead vehicle varied its speed based on a sinusoidal function.

Face Video

Video and audio of the participant's face was recorded using a GoPro Hero3 camera at 720p resolution. The video footage of the participant's face was used to help track eye glances whether the eyes were on or off the road.

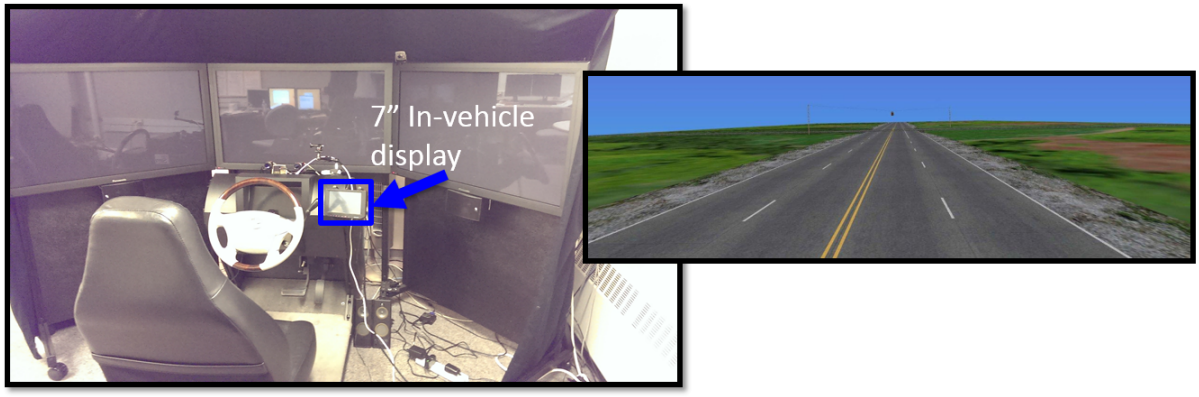
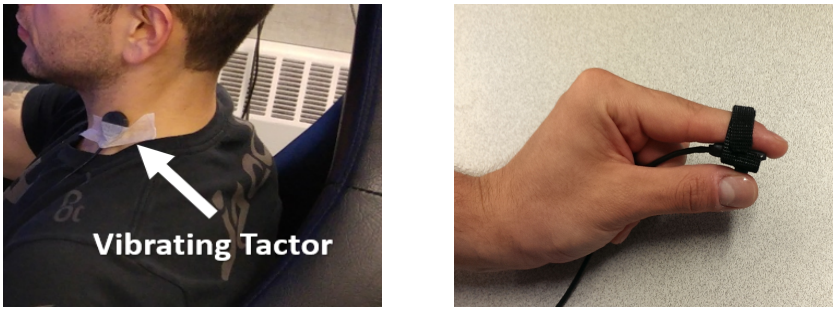


Figure 4.1: NADS MiniSim

4.4.2 Tactile Detection Response Task (TDRT)

The cognitive workload of the task was assessed using the Tactile Detection Response Task (TDRT) protocol. The hardware or setup consisted of a vibrating tactor and a microswitch button (Figure 4.2). The TDRT was set up per ISO standard (ISO/DIS 17488, 2015), hence the tactor was taped to the side of the neck above the collarbone and the sensor vibrated randomly once every 3-5 seconds.

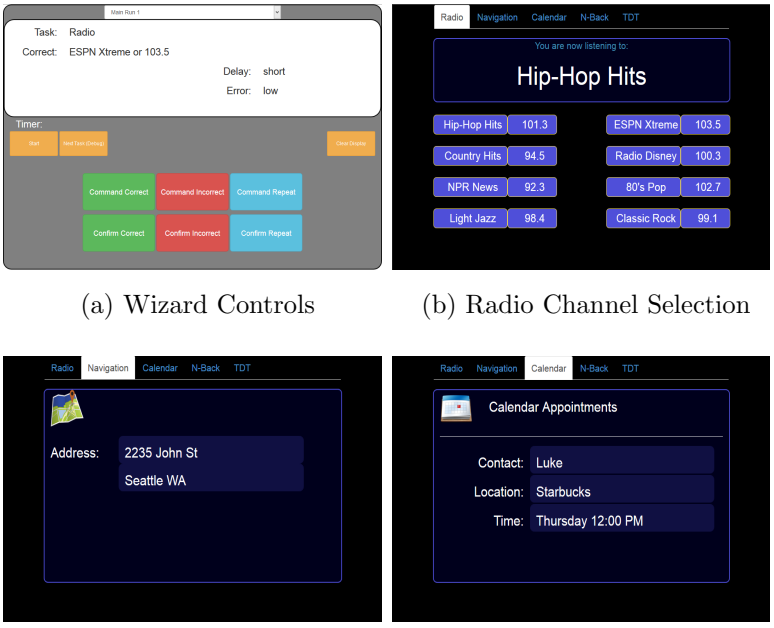


(a) Vibrating Tactor (b) Microswitch Button

Figure 4.2: Participant needs to respond to vibrating sensor taped to the neck by pressing a microswitch button

4.4.3 In-Vehicle Interface

Participants had to interact with an In-Vehicle Information System using voice commands for the secondary tasks. Participants were asked to select a radio channel, navigate to a specific address, and create a calendar appointment all using voice input.



(a) Wizard Controls (b) Radio Channel Selection

(c) Navigation (d) Calendar Scheduling

Figure 4.3: A human or wizard acts as the voice recognition system unbeknown to the participant, and the wizard controls the outcomes for voice tasks such as radio (b), navigation (c), and calendar appointments (d)

Instead of using a commercial software to handle voice recognition, the study implemented the Wizard of Oz protocol where the participant believes that he or she is interacting automated voice recognition system, when in reality the behavior of the system is controlled by a research assistant (Fraser & Gilbert, 1991).

Even though commercially available speech recognition software such as Google Now and

iPhone Siri have fairly quick and accurate speech recognition capabilities, the Wizard of Oz allows the precise control of recognition error, is relatively quicker to implement for rapid testing and prototyping, and the experiments can be highly repeatable. Figure 4.3 shows the interface that the study participant interacts with along with a screen-shot of the Wizard Controls which allows the researcher to act behalf the voice recognition system and control the flow of the interaction.

4.5 Procedure

Prior to the start of experiment, participants had to read and sign an informed consent form. Participants underwent training where they had the opportunity to get familiar with the handling of the vehicle inside in driving simulator, get comfortable with operating the VCS, and get familiar with the TDRT. The training session typically lasted between 10 to 15 minutes. During the drive, participants were asked to follow the lead vehicle while maintaining 50 mph (80 km/h) while simultaneously performing a voice task and responding to the tactile stimuli.

The main experiment consisted of 24 radio, 12 navigation, and 12 calendar tasks for a total of 48 voice tasks. After completing all the voice tasks, participants were compensated $30 USD per hour for a study that could last anywhere between 1 - 1.5 hours. The study was split into two sessions with a 10 minute break between each session.

4.6 Independent Variable

4.6.1 Voice Task: 3 Levels

VCS's induce cognitive distraction, and the amount of cognitive distraction was determined by the complexity of the voice tasks. In this study, participants had to select a radio channel, navigate to a specified address, and schedule an appointment all using voice commands. Each voice task varies in complexity and are performed as follows:

Radio To operate the radio using voice, participants needed to say the name of the radio station (e.g. *"Tune to Hip-Hop Hits"*). The radio task consists of **1 chunk** of information which is the name of the radio station.

Navigation Participant inputs an address that consisted of a four-digit house number and a short generic street name (e.g., "Navigate to 5435 Main St") via voice. Navigation task consist of **2-3 chunks** of information (street number, street name).

Calendar To create a calendar appointment using voice commands, participant needed to input a contact name, location, time, and day (e.g. "Schedule an appointment with Luke at Starbucks on Friday 12 PM") using voice. The calendar task consists of **4 chunks** of information (contact name, location, time, and day).

The radio, navigation, and calendar voice tasks were designed to be short term memory recall tasks. An automated voice prompt gave instruction to the participant telling him or her to "Go to 5435 Main St", and the participant immediately proceeded to issue the proper voice command and say "Navigate to 5435 Main St." Cognitive workload was varied by changing the amount of information required for recall. The radio task consisted of one chunk of information (name of radio station), while the calendar tasks were a lot more complex and consisted of four chunks of information (name, location, day of the week, and time).

4.6.2 Time or System Delay: 2 Levels

Results from the contextual interview in Chapter 3 showed that VCS's commonly experienced system time outs, where the VCS is slow or fails to respond to a voice command input. Time delay is an experimental control factor that simulates a system time out. Time delay is defined as the time required for the VCS to respond after the user has issued a voice command. All participants encountered both **Short** and **Long** time delays, where

a **Short time delay** had response time under 2 seconds as the Wizard of Oz based VCS required a research assistant to listen to participant's voice command and make a selection via mouse click. **Long time delays** on the other hand had a delayed response of 8 seconds. For example, the participant would issue a voice command and say "Tune to Classic Rock station". The research assistant acting behalf the VCS will make the selection and the voice command executes 8 seconds later. The 8 second delay was based on slow VCS interaction observed in the contextual interview in Chapter 3.

4.6.3 Recognition Error: 2 Levels

A recognition error was either **Present** or **Absent**. In the **Present** condition, the error rate was forced to be 66 % error, or 2/3 of the tasks contained system recognition errors. The **Absent** condition did not include any system recognition errors. This independent variable was a between-subject factor with 24 participants in the **Present** condition, and 24 different participants in the **Absent** condition.

When a recognition error was **Present**, the system made a word substitution error. For example, the participant would give a voice command and say "*Tune to Classic Rock station*". The system would replace *Classic Rock* and substitute it with something incorrect and respond with "*You are now listening to Jazz Hits"*. Radio station names, addresses, contact names, meeting time, and meeting dates can all experience substitution errors.

4.7 Dependent Variable

Two dependent variables was used for the analysis: *Response Time*, a continuous variable, *TDRT Miss* events, a binary outcome. Response time was defined as stimuli responses within the 100-2500 ms after onset. This variable was based on the ISO 17488 (ISO/DIS 17488, 2015), where a tactile stimuli appears every 3-5 seconds.

As per ISO 17488, a TDRT miss is any instance where the participant does not even press

the microswitch button despite the presence of an onset tactile stimuli. Cognitive demand can be high enough where the participant fails to attend the TDRT. In addition, response times outside the 100-2500 ms interval are also marked as a miss. Response time and TDRT misses increases with higher cognitive workloads (Merat et al., 2007).

TDRT Hit

- Any response time within 100-2500 ms interval

TDRT Miss

- Failing to respond respond to onset stimuli
- Any response time outside 100-2500 ms interval

4.8 Data Analysis

This is a 3×4 mixed-effects model with repeated measures, where *Time Delay* and *Voice Tasks* are within subject factors, and *Recognition Error* is a between subject factor. A mixed effects logistic regression was used to analyze the binary outcome, *TDRT Miss*.

The R statistical package *lme4* (Bates, Mächler, Bolker, & Walker, 2015) was used to perform both the linear mixed effects analysis and the mixed effects logistic regression analysis. *Voice Task*, *Time Delay*, *Recognition Error*, *Gender*, and *Age Groups* were entered as fixed effects while *participant ID* was entered as random effect. Eqn 4.1 and Eqn 4.2 are the models explored in the study where Eqn 4.1 looked at demographic information while Eqn 4.2 modeled various voice task conditions.

$$y = \beta_0 + \beta_1(Gender) + \beta_2(AgeGroup) + \gamma_0 + \gamma_1(ParticipantID) + \epsilon \tag{4.1}$$

$$\begin{aligned} y =& \beta_0 + \beta_1(VoiceTask) + \beta_2(TimeDelay) + \beta_3(RecognitionError) \\ &+ \gamma_0 + \gamma_1(ParticipantID) + \epsilon \end{aligned} \tag{4.2}$$

β**:** Fixed effects

γ**:** Random effects
ϵ**:** Residual Error
y**:** Response variable, either TDRT Response Time or TDRT Miss

4.9 Results

4.9.1 Descriptive Analysis

The mean age of the participants was 40.5 (SD=16.7) for females and 39.4 (SD=16.0) for males, while the age ranged from 19 to 73 years old. There were 12,810 TDRT responses collected across all 48 participants, with an average of 261.4 tactile stimuli events per subject. Long time delays and system recognition errors extended task duration, thus increasing the number of tactile stimuli events. Subjects in the Recognition Error *Absent* condition averaged 229.1 tactile stimuli events while subjects under Recognition Error *Present* averaged 292.5 tactile stimuli events.

Table 4.2: Response Time (RT) in seconds and Miss Counts summary across Gender and Age Groups

Gender	Age Group	Total Count	Mean(RT)	SD(RT)	Miss Count	Miss Rate (%)
Female	18-24	1476	0.52	0.32	96	7
	25-39	1307	0.48	0.30	62	5
	40-54	1452	0.50	0.33	131	9
	55-75	1460	0.60	0.35	194	13
Male	18-24	1741	0.69	0.50	95	5
	25-39	1567	0.53	0.37	128	8
	40-54	1466	0.56	0.36	121	8
	55-75	1247	0.58	0.38	267	21

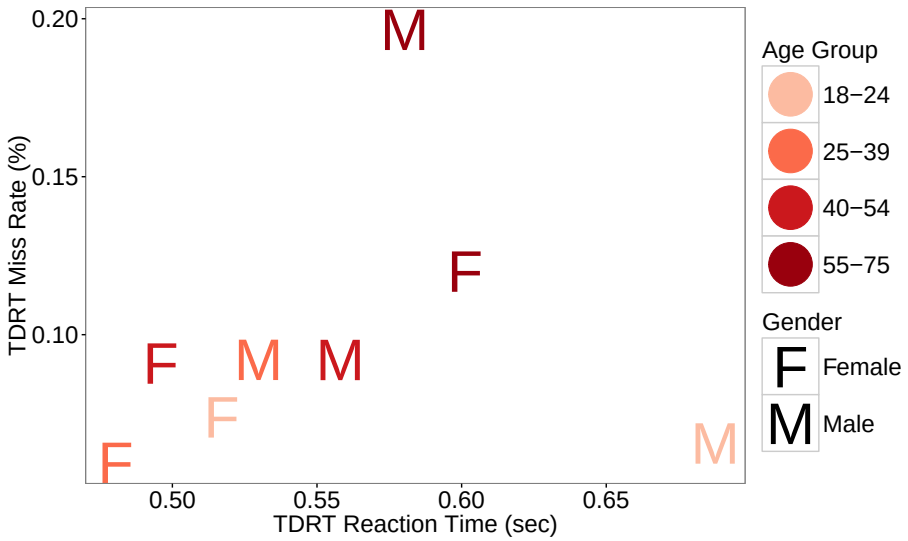


Figure 4.4: Mean TDRT Reaction Time vs Mean TDRT Miss Rate

Table 4.2 shows the mean TDRT response time, miss count, and miss rates for participants grouped by gender and age. Overall, the miss rates were not very high. The average miss rate for all participants was approximately 7%. Older drivers (ages 55-75) tend to have longer response times, higher miss count and miss rates. Females drivers had faster mean response times on average when compared to male drivers. Figure 4.4 plots mean TDRT reaction time against mean miss rates. While higher miss rates were observed among older drivers, no strong correlation can be seen between response time and miss rates.

The interaction plot in Figure 4.5 demonstrated that TDRT response time increased when the participant has to simultaneously drive and engage in a voice task. Furthermore, the navigation and calendar tasks exhibited longer response times when compared to radio as they are more complex because participants needed to recall more chunks of information.

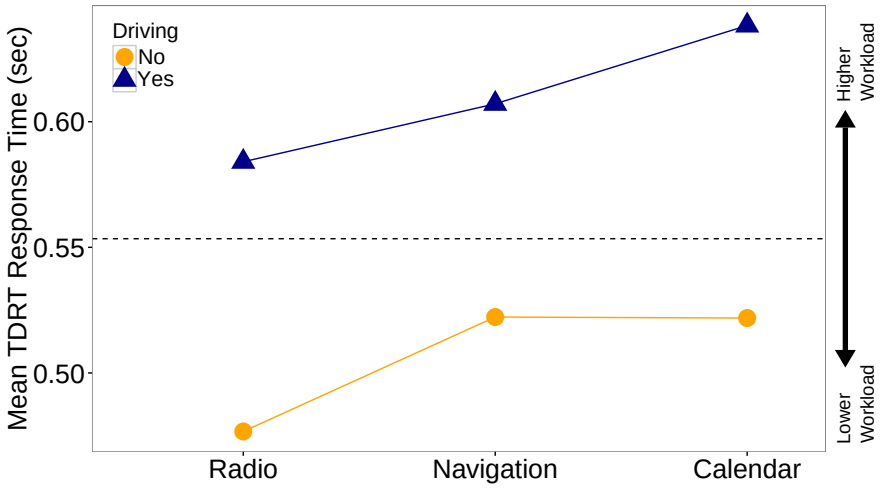


Figure 4.5: Cognitive workload is higher when participant has to drive the vehicle. Dashed line is the overall mean TDRT Response Time.

4.9.2 TDRT Response Time Analysis

A mixed effects linear model was used to examine TDRT reaction times. The TDRT reaction time was log transformed to meet the assumption of normality. Age Group and Gender were not significant factors ($p < 0.05$) in predicting TDRT response time.

Even though Age and Gender were not significant, there were differences observed in the Time Delay ($z = -1.85, p < 0.05$) and interaction of Voice Task $\times$ Recognition Errors ($z = 2.24, p < 0.05$). Figure 4.5 indicated that TDRT response times increased with the number of information chunks the participant needed to recall. The inclusion of recognition error and time delay to the voice task also impacted response time (see Figure 4.6). TDRT response times were quicker with *Long* Time Delays when compared to *Short* Time Delays. Furthermore, cognitive workload was signigicantly lower (lower TDRT reaction time) for calendar tasks with Recognition Error *Present* than with Recognition Error *Absent*.

Table 4.3: Mixed Effects Model for TDRT Response Time

Variables	Estimate	Std. Error	z-value	p-value
(Intercept)	-0.65	0.07	-9.75	p<0.001
Long Delay vs Short Delay	-0.05	0.03	-1.85	p<0.05
Navigation vs Radio	0.04	0.03	1.54	p<0.07
Calendar vs Radio	0.04	0.03	1.62	p<0.06
Error Absent vs Error Present	0.01	0.1	0.11	N.S.
Long Delay * Err Absent vs Short Delay * Err Present	0.01	0.04	0.33	N.S.
*Short Delay*Radio vs*				
Long Delay * Navigation	0.04	0.03	1.35	N.S.
Long Delay * Calendar	0.02	0.03	0.74	N.S.
*Radio * Error Present vs*				
Navigation * Error Absent	-0.05	0.04	-1.1	N.S.
Calendar * Error Absent	0.09	0.04	2.24	p<0.05
*Short Delay * Error Present * Radio vs*				
Long Delay * Error Absent * Navigation	-0.01	0.05	-0.15	N.S.
Long Delay * Error Absent * Calendar	-0.06	0.05	-1.1	N.S.
$AIC = 8287 \quad LogLikelihood = -4127$				
$AIC_0 = 13186 \quad LogLikelihood_0 = -6590$				

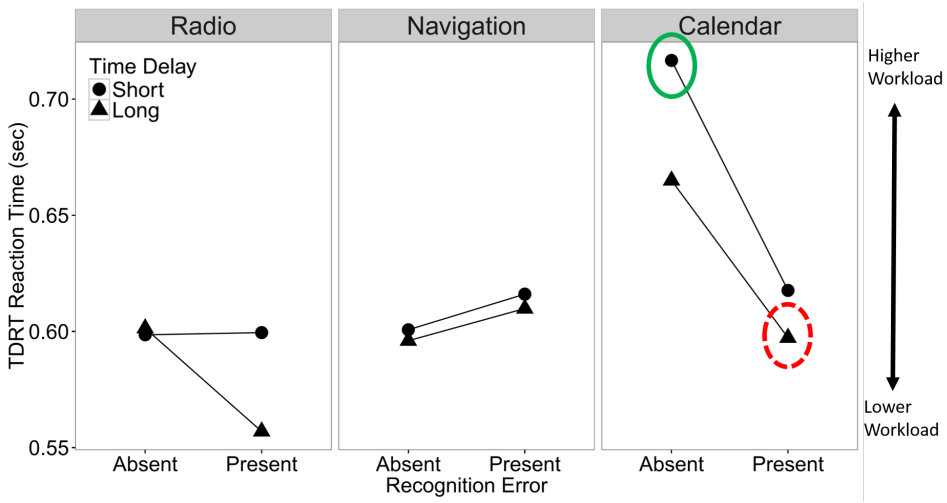


Figure 4.6: A calendar task with VCS imperfections (dashed circle) had lower cognitive workload than a VCS with no imperfections (solid circle)

4.9.3 TDRT Miss Analysis

A mixed effects logistic regression was used to examine TDRT miss events where the subjects was identified as random effect. Age Group and Gender were not significant in predicting TDRT misses (at $p < 0.05$).

Voice Task was a significant main effect (Table 4.4). More specifically, drivers engaged in the navigation and calendar task were 1.67 and 1.84 times more likely to experience TDRT miss event when compared to the radio task. Since the navigation and calendar tasks are cognitively more demanding than the radio task, more TDRT misses are expected which is in line with findings observed in Vilimek et al. (2013) and Ranney et al.(2014).

An unexpected result is that time delay and recognition error had no impact on TDRT misses. This outcome suggests that poor VCS performance did not increase driver's cognitive workload.

Table 4.4: Mixed Effects Logistic Regression Model for TDRT Miss

Variables	Estimate	Std. Error	z-value	p-value
(Intercept)	-2.73	0.33	-8.40	<0.001
Navigation vs Radio	0.51	0.14	3.57	<0.001
Calendar vs Radio	0.61	0.14	4.28	<0.001
Long Delay vs Short Delay	-0.16	0.12	-1.41	N.S.
Recognition Error Present vs Absent	-0.58	0.43	-1.33	N.S.
$AIC = 4865$	$LogLikelihood = -2424.5$			
$AIC_0 = 6591$	$LogLikelihood_0 = -3293.5$			

4.10 Discussion

Voice control systems can include time delays and recognition errors, which can negatively impact workload for the driver as they interact with these systems while driving. To date, there is little research on the workload associated with voice control systems given these system imperfections. Previous DRT studies primarily focused on determining which type of DRT stimuli is most sensitive to changes in cognitive workload using VCS proxy tasks like the counting or n-back (Vilimek et al., 2013; Ranney et al., 2014; Merat and Jamson, 2008; Harbluk et al., 2013). As demonstrated in the contextual interview in Chapter 3, VCS's are very susceptible to errors and the likelihood of being distracted by system imperfections can have an impact on safety, satisfaction, and overall usability. This study examined cognitive workload while drivers were engaged in various in-vehicle tasks that were prone to errors.

The first research question is to determine if the DRT protocol is suitable to measure the cognitive workload of a VCS interaction. Study results show that the navigation and calendar tasks had higher TDRT response times and generated more TDRT misses when compared to

the radio task. This outcome was expected since the navigation and calendar tasks required subjects to recall three and four chunks of information respectively when compared to one chunk for radio. This outcome also shows that the design of speech interaction can impact cognitive distraction. It is important for interface designers to be aware that voice command complexity can increase cognitive distraction.

The second research question is to determine if VCS performance errors elevate cognitive workload. Study results showed that the presence of time delay and recognition error appeared to reduce cognitive workload as demonstrated by a quicker TDRT response time. A system time delay during a VCS interaction may be annoying but that does not necessarily relate to higher cognitive load. In fact, the expectation of a delay, may actually free up other attentional resources for the driver to conduct other subtasks such as listening, remembering task prompts, or speaking to the VCS. In this study, cognitive workload was a function of the amount of information needed for memory recall. The presence of recognition errors however did not increase the number chunks of information needed during recall.

In this study, an imperfect VCS had lower cognitive workload when compared to a VCS free of time delays and recognition errors; this finding is counter intuitive. That said, annoyance was not a factor examined in this study. Participants may actually not have been annoyed at all by the experimental manipulations but that may not be the same experience if they were using a similar system in actual driving conditions.

The study followed ISO protocol which should be robust and sensitive to accommodate a vast array of different interfaces, interactions, and designs; however, study results indicates that the ISO protocol is not robust enough to universally evaluate voice control systems. The ISO protocol is based on experiments conducted with n-back task (ISO/DIS 17488, 2015). The n-back task which is a fixed paced memory recall task may not be very representative of an actual VCS where the pace of task can greatly vary. During a voice interaction for example, there could be task resets due to recognition errors which increases task duration.

The study results also suggests that the cognitive demand for a VCS interaction is not uniform. A calendar task can be cognitively more demanding than a radio channel selection; however, the increase in cognitive demand can be masked by other elements such as time delays.

With uneven pacing and non-uniform cognitive workload, analyzing of a single complete task such as navigating to a specific address does not provide sufficient resolution to discern dynamic changes in cognitive workload. For example in a navigation task, the participant could be speaking, listening, or waiting for system response, which all contribute to different pacing and cognitive workloads. Future work should consider separating the voice interaction into subtasks (e.g. listening, speaking, and waiting), and analyze the cognitive workload of the subtasks to isolate the more cognitively intense interactions.

In summary, the TDRT protocol shows promise for detecting differences in cognitive workload for auditory interfaces. However, this study also highlights some of its shortcomings. System imperfections within an auditory interface does not necessarily lead to higher cognitive workload. In fact, the cognitive workload may not be uniformly distributed over the course of the VCS interaction. Further research would be needed to identify the moment to moment changes in workload over the course of the voice interactions. With increasing use of voice within the vehicle, these subtle differences will become more important.

4.11 Chapter Summary

A driving simulator study was conducted to assess the cognitive workload of VCS use. Study participants had to follow a lead vehicle, operate a VCS, and respond to a randomly occurring tactile stimuli (TDRT). Furthermore, VCS performance was manipulated by creating forced recognition errors along with system time outs.

Study results showed that the TDRT protocol was sensitive to cognitive workload associated with VCS use. Cognitive workload was higher when participants had to drive when

compared to no driving. The navigation and calendar tasks were cognitively more demanding than the radio task as they involve more complex voice commands.

However, when taking VCS errors into consideration, there were situations* where a poorly performing VCS (interaction with recognition error and system time outs) had lower cognitive workload than a VCS with no performance issues. This outcome questions the robustness of the DRT ISO protocol. The discrepancy between poor VCS performance leading to lower cognitive workloads highlights the multifaceted characteristics of a VCS interaction. During a VCS interaction, the user could be performing various cognitive processes such as listening to the system, speaking or issuing a command, glancing at the display, or even waiting for the VCS to respond (Figure 4.7). Each action can potentially have different contribution to overall cognitive workload.

Driving Simulator Study 1 showed that VCS errors do not necessarily result in higher cognitive workload. Poor VCS performance might have worse user satisfaction, however that doesn't mean there is a strong correlation between poor user satisfaction and high cognitive workload. A VCS interaction on the other hand is made up of many different cognitive processes. Chapter 5 seeks to analyze the cognitive workload of different cognitive processes such as listening, speaking, and waiting. The goal of Chapter 5 is to investigate whether the different cognitive processes can be used to explain the variability in cognitive workload for a VCS interaction.

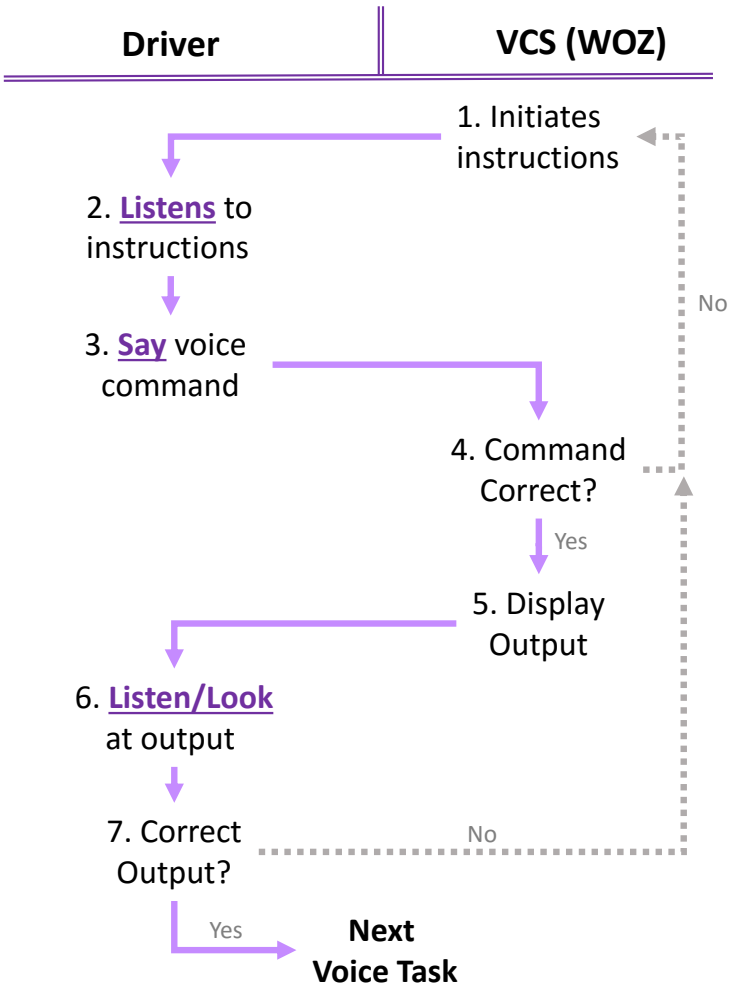


Figure 4.7: Task analysis of Driving Simulator Study 1. Drivers perform cognitive processes such as listening, speaking, and reading.

Chapter 5

COGNITIVE WORKLOAD OF VCS SUBTASKS

5.1 Introduction

The driving simulator study in Chapter 4 showed situations where a poorly performing VCS (interaction with recognition error and system time outs) had lower cognitive workload than a VCS with no performance issues. The contradiction of a poor performing VCS with lower cognitive workload highlights some issues associated with measuring the cognitive workload of a VCS interaction. Cognitive workload was measured using the Tactile Detection Response Task (TDRT) protocol which followed the ISO standard (ISO/DIS 17488, 2015). The goal of the ISO standard is to guide OEM and after market device makers on how to evaluate in-vehicle interfaces. The contradiction of a poor performing VCS with lower cognitive workload can discredit the usefulness and robustness of the ISO standard.

The results observed in Chapter 4 also highlights the diverse cognitive processes that constitutes a VCS interaction. A user for example could be performing various cognitive actions such as listening, speaking, or waiting in a single voice task. Each individual action or *subtask* can have different contribution to cognitive workload as some subtask may elevate, while other subtasks may lessen cognitive workload. It is important to characterize a VCS interaction by its subtasks to gain a better understanding how cognitive workload changes throughout the VCS interaction over time.

The goal of Chapter 5 is to explore if different cognitive processes such as as speaking and listening can be used to explain variability in cognitive workload for a VCS interaction. To answer this research question, this chapter tests the following hypothesis:

1. Is there a difference in TDRT response time and miss rates for VCS subtasks such as listening to instructions, speaking to the VCS, waiting for VCS to respond, and giving a confirmation? (H1)

2. Is there a difference in TDRT response times and miss rates for *listening, speaking, waiting*, and *confirmation* among different voice tasks such as radio, navigation, and calendar? (H2)

5.2 Methods

VCS Driving Simulator Study 1 in Chapter 4 generated over 12,000 data points or observations. The data was recoded to include information about subtask being performed (e.g. listening, speaking, waiting, and confirmation), and new data analysis was conducted. This section goes over the data reduction method for determining whether the participant was either listening, speaking, waiting, or giving a confirmation. The methods section will also review the methodology for data analysis.

5.2.1 Data Reduction

A single VCS interaction such as entering an address into the navigation system, consists numerous cognitive processes. The driver could be listening to the passenger for instructions. Next the driver speaks to the system and glances at the system to make sure the address has been entered correctly. Table 5.1 shows a typical interaction between the participant and the VCS in Driving Simulator Study 1 in Chapter 4. In addition, Table 5.1 shows each part of the VCS interaction was mapped to a subtask.

It is important to note that the subtasks of *listening*, *speaking*, *waiting*, and *confirmation* are labels to a step or interval of time in the VCS interaction. The subtask names are not a precise descriptor of the cognitive process that is taking place. For example, during the *confirmation* subtask, the participant could be "speaking" by saying the word *"Yes"* as

Table 5.1: Interaction between Participant and VCS

Step	Actor	Dialog/Action	Subtask
1	Instructions:	*"Please go to 5376 Main St, Seattle WA"*	Listening
2	Participant:	*"Navigate to 5376 Main St, Seattle WA"*	Speaking
3	VCS:	(System Time Out 8 Seconds)	Waiting
4	VCS:	*"Do you want to navigate to 5376 Main St, Seattle WA?"*	Confirmation
5	Participant:	*"Yes"*	Confirmation

shown in Step 5 (Table 5.1). The "speaking" that is taking place in Step 2 (Table 5.1) on the other hand requires the participant to utter a lot more words for memory recall, instead of a simple yes/no response as seen in Step 5.

5.2.2 Independent Variables

Subtasks: 4 Levels

During a given moment of the VCS interaction, participant could either be *listening* to for instructions, *speaking* or issuing a voice command, *waiting* for the system to respond, or giving a *confirmation*. Since cognitive workload has been observed to be non-uniform (Kun et al., 2013), it is hypothesized that each subtask will have different cognitive demands. The subtasks that are considered for the study are as follows:

Listening An automated voice is giving instructions to the participant to perform a VCS task (e.g. "Please go to 5376 Main St, Seattle WA"). During the *listening* subtask, the participant needs to encode the information in working memory.

Speaking During the *speaking* subtask, the participant issues the voice command to complete the voice task. The participant needs to be able to recall and verbalize the

information encoded during the *listening* subtask.

Waiting After the participant issues the voice command, the VCS can timeout for 8 seconds. During the system timeout, the participant is waiting for the VCS to respond.

Confirmation Confirmation subtask is a way for the participant to check if the VCS received the correct inputs for the voice command (e.g. "Do you want to schedule an appointment with Luke at McDonalds on Tuesday at 5 PM?").

Voice Task: 3 Levels

Cognitive workload is manipulated by the amount of information the participant needs to recall from short term or working memory. It is hypothesized that cognitive workload increases when more chunks of information needs to be recalled.

Radio To operate the radio using voice, participants needed to say the name of the radio station (e.g. *"Tune to Hip-Hop Hits"*). The radio task consists of **1 chunk** of information which is the name of the radio station.

Navigation Participant inputs an address that consists of a four-digit house number and a short generic street name (e.g., "Navigate to 5435 Main St") via voice. Navigation tasks consist of **2-3 chunks** of information (street number, street name).

Calendar To create a calendar appointment using voice commands, participant needs to input a contact name, location, time, and day (e.g. "Schedule an appointment with Luke at Starbucks on Friday 12 PM") using voice. The calendar task consists of **4 chunks** of information (contact name, location, time, and day).

5.2.3 Dependent Variables

Two dependent variables was used for the analysis: *Response Time*, a continuous variable, *TDRT Miss* events, a binary outcome. Response time defined as stimuli responded within the 200-2500 ms onset. This variable was based on the ISO 17488 (ISO/DIS 17488, 2015), where a tactile stimuli appears every 3-5 seconds.

As per ISO 17488, a TDRT miss is any instance where the participant does not even press the microswitch button despite the presence of a tactile stimuli. Cognitive demand can be high enough where the participant fails to attend the TDRT. In addition, response times outside the 100-2500 ms interval are also marked as a miss. Response time and TDRT misses increases with higher cognitive workloads (Merat et al., 2007).

TDRT Hit

- Any response time within 100-2500 ms interval

TDRT Miss

- Failing to respond respond to onset stimuli
- Any response time outside 100-2500 ms interval

5.2.4 Data Analysis

This is a 3×4 mixed-effects model with repeated measures, where *Subtasks* and *Voice Tasks* are within subject factors. A mixed effects logistic regression was used to analyze the binary outcome, *TDRT Miss*.

The R statistical package *lme4* (Bates et al., 2015) was used to perform both the linear mixed effects analysis and the mixed effects logistic regression analysis. *Voice Task*, *Subtask*, *Gender*, and *Age Groups* were entered as fixed effects while *participant ID* was entered as random effect. The analysis in this chapter is very similar to the data analysis found in Chapter 4. The only difference is the inclusion of *VCS subtask*.

$$y = \beta_0 + \beta_1(Gender) + \beta_2(AgeGroup) + \gamma_0 + \gamma_1(ParticipantID) + \epsilon \tag{5.1}$$

$$y = \beta_0 + \beta_1(VoiceTask) + \beta_2(Subtask) + \gamma_0 + \gamma_1(ParticipantID) + \epsilon \tag{5.2}$$

β**:** Fixed effects

γ**:** Random effects

ϵ**:** Residual Error

y**:** Response variable, either TDRT Response Time or TDRT Miss

5.3 Results

5.3.1 Descriptive Analysis

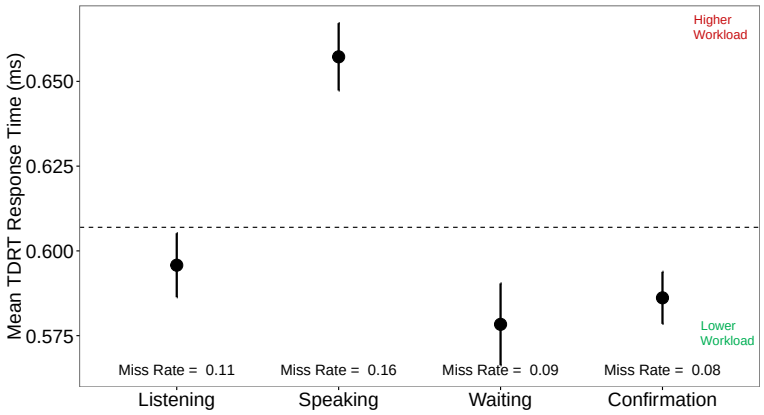


Figure 5.1: Speaking subtask is cognitively most demanding

Figure 5.1 shows that *speaking* subtask had the highest mean TDRT response times and miss rates. This outcome suggests that *speaking* is cognitively more demanding when

compared to other VCS subtasks such as *listening, waiting,* and *confirmation*. No significant differences was observed between *listening, waiting,* and *confirmation* subtasks.

Figure 5.2 plots mean TDRT response times of VCS subtasks for different age groups. The *speaking* subtask consistently has the highest TDRT response times for all age groups. No significant differences was observed (at $p < 0.05$) between *listening, waiting,* and *confirmation* subtasks for each age groups.

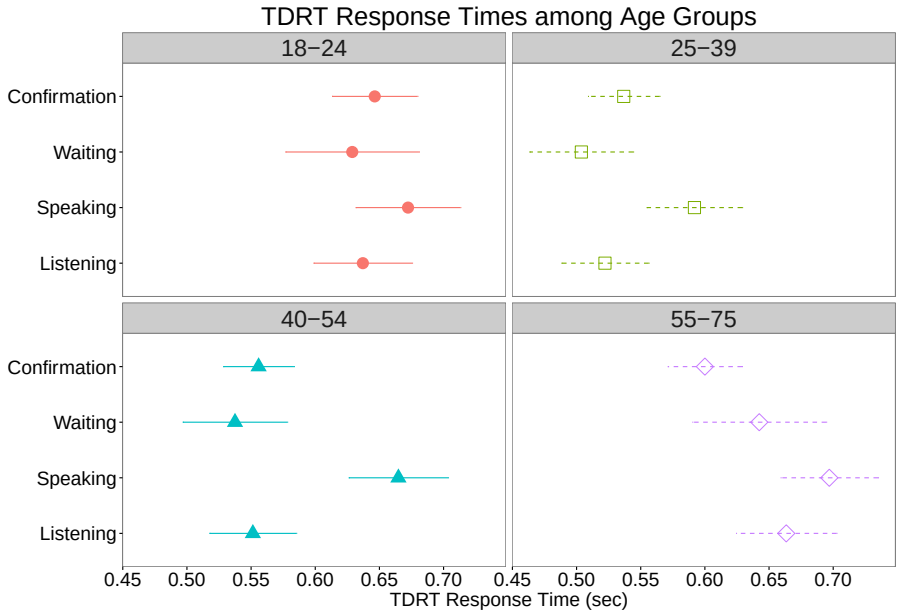


Figure 5.2: Speaking subtask has highest mean TDRT Response Time for each Age Group

5.3.2 TDRT Response Time Analysis

A mixed effects linear model was used to analyze TDRT reaction time outcomes. TDRT reaction time was log transformed to meet the assumption of normality. Age group and gender were not significant factors (at $p < 0.05$) in predicting response time.

Voice task (e.g. radio, navigation, and calendar) and subtasks (e.g. listening, speaking,

waiting, and confirmation) on the other hand were significant factors. The mixed linear model in Table 5.2 showed significant differences ($z = 6, p < 0.001$) in TDRT response time between the calendar task ($M = 0.64, SD = 0.009$) and radio task ($M = 0.58, SD = 0.007$). The mean TDRT response times for *speaking* subtask was signifcantly higher than *listening* ($z = -3.97, p < 0.001$), *waiting* ($z = -3.78, p < 0.001$), and *confirmation* ($z = -4.27, p < 0.001$).

Furthermore, interaction terms between Voice Task × Subtask was also significant. Response times for $Calendar \times Speaking$ ($M = 0.70, SD = 0.020$) was significantly higher than $Radio \times Listening$ ($M = 0.60, SD = 0.016$), $Radio \times Waiting$ ($M = 0.56, SD = 0.016$), $Radio \times Confirmation$ ($M = 0.58, SD = 0.011$), and $Navigation \times Listening$ ($M = 0.55, SD = 0.016$). Figure 5.3 further illustrates that *speaking* for the radio task was significantly lower than the *speaking* for the calendar ($z = 6.00, p < 0.001$) and navigation ($z = 5.67, p < 0.001$) task.

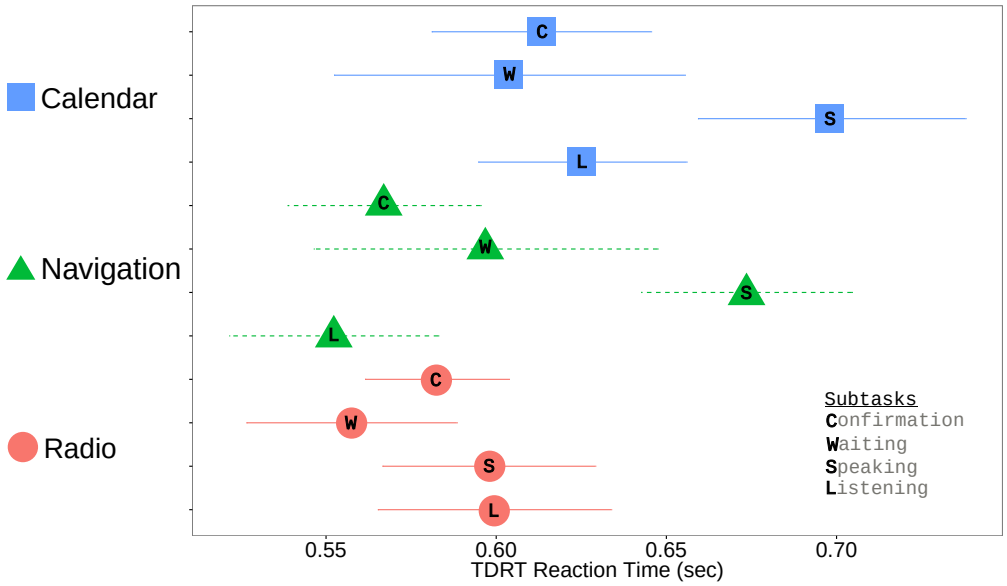


Figure 5.3: Speaking subtask for Navigation and Calendar is cognitively more demanding

5.3.3 TDRT Miss Analysis

Since a TDRT *Miss* is a binary outcome, a mixed effects logistic regression was used to predict tactile TDRT miss events, and participant was set as random effect. Figure 5.4 shows the outcome of mixed logistic regression model.

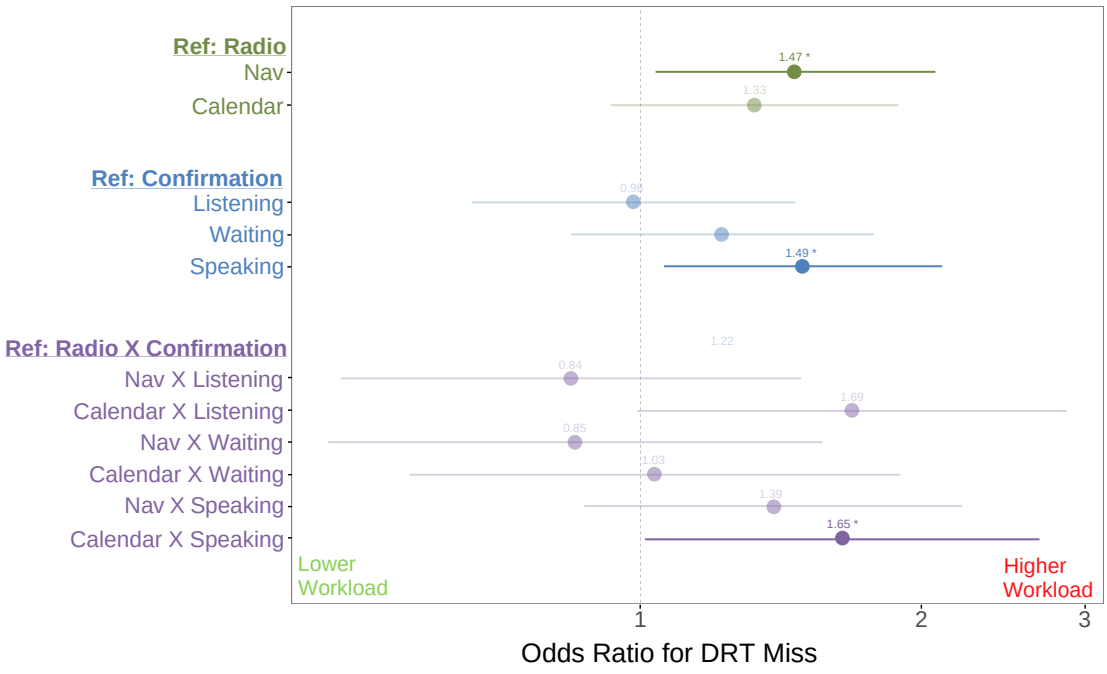


Figure 5.4: Odds of TDRT miss for *Speaking* is 1.49 times or 49% higher than *Confirmation*

Drivers performing the navigation task are 1.47 times more likely to *miss* TDRT signals when compared to the radio task. Furthermore, drivers who are engaged in *speaking* are 1.49 times more likely to *miss* TDRT signals when compared to *confirmation* subtask. Finally, the interaction between Voice Task × Subtask was significant. The odds for a TDRT *miss* for drivers who are *speaking* during the calendar task is 65% higher than drivers who are engaged with the *confirmation* prompt for the radio task.

Table 5.2: Mixed Effects Model for log of TDRT Response Time

Variables	Estimate	Std. Error	z value	p-value
(Intercept)	-0.51	0.05	-10.33	p<0.001
Ref: Calendar				
Radio	-0.15	0.02	-6.00	p<0.001
Navigation	-0.02	0.02	-0.94	N.S.
Ref: Speaking				
Listening	-0.10	0.02	-3.97	p<0.001
Waiting	-0.12	0.03	-3.78	p<0.001
Confirmation	-0.11	0.02	-4.27	p<0.001
Ref: Calendar × Speaking				
Radio × Listening	0.09	0.03	2.47	p<0.05
Navigation × Listening	-0.08	0.03	-2.49	p<0.05
Radio × Waiting	0.08	0.04	2.09	p<0.05
Navigation × Waiting	0.02	0.04	0.55	N.S.
Radio × Confirmation	0.10	0.03	3.19	p<0.01
Navigation × Confirmation	-0.04	0.03	-1.07	N.S.
$LogLikelihood = -4113$	$AIC = 8255$			
$LogLikelihood_0 = -4186$	$AIC_0 = 8378$			

5.4 Discussion

This chapter characterizes a VCS interaction as a multi-step process that encompasses different cognitive processes. Cognitive workload was analyzed for different cognitive processes or VCS subtasks encountered in Driving Simulator Study 1 in Chapter 4. Results showed that the *speaking* subtask had a higher mean detection response time and miss rate when compared to *listening*, *waiting*, or *confirmation*. Speaking or language processing can be considered a cognitively intricate process. A speaker needs to begin with an idea they want to convey. They then need to arrange the sequence of sounds that constitutes the phonological content of the word to expresses their idea (Ferreira & Pashler, 2002). Language processing can be subjected to central processing bottlenecks, which can cause a delay in performing a secondary task (Pashler, 1994; Ferreira & Pashler, 2002). The observed outcome of higher cognitive workload for *speaking* subtask was consistent with the findings of Nunes and Recarte (2002), and Strayer and Johnson (2001), where vehicle driving performance was compromised when subjects were involved in a speaking task (e.g., telling a story or engaging in a hands free cellphone conversation).

While results showed that *speaking* increases cognitive workload, the increase however is dependent on the complexity of the voice command. No significant differences in TDRT response times was observed between *listening, speaking, waiting*, and *confirmation* for the radio task. Since the participant only needed to recall one chunk of information, using voice command to select a radio channel (i.e. *speaking*) was not cognitively more demanding than *waiting* for an idle VCS to respond.

The navigation and calendar task on the other hand required the participant to recall 2-4 chunks of information, therefore *speaking* had significant higher cognitive workloads relative to *listening* and *confirmation*. Furthermore, *speaking* for the navigation and calendar task was cognitively more demanding than *speaking* for radio channel selection.

There are numerous ways to characterize the cognitive workload of a VCS interaction.

Strayer et al. examined cognitive workload from a system level as they compared the cognitive workload of iPhone Siri with another system like the Ford Sync (2015). Harbluk et al. (2013) examined the cognitive workload at a voice task level as they compared the cognitive workload of a iPhone Siri calendar appointment scheduling task with iPhone Siri simple query task (e.g "What is the weather?"). A system level or voice task level analysis does not provide insight as to why one system or task may have higher cognitive workload. Characterizing a VCS interaction into different cognitive processes such as listening, speaking, and waiting on the other hand allows interface designers identify moments in the VCS interaction where cognitive workload is greatly elevated, and provides better insights as to where improvements can be made.

5.5 Chapter Summary

VCS interactions can contain many communication elements. In the same voice task, the driver could either be listening to the system, speaking a voice command, or even waiting for the VCS to respond. Prior DRT studies have not analyzed the cognitive workload for different VCS cognitive processes. The goal of this chapter is to determine if different cognitive processes such as listening, speaking, waiting, and confirmation) can be used to explain the variability in cognitive workload of a VCS interaction.

Using the data collected in Driving Simulator Study 1 in Chapter 4, the data was recoded to include information on which subtask the participant was engaged in. Results showed that the speaking subtask generally had the highest cognitive workload. However, the cognitive demand for speaking was also dependent on whether the driver is selecting a radio channel, navigating to an address, or entering a calendar appointment. If the driver was using voice command to select a radio channel, the cognitive workload of speaking was not different than waiting for a VCS response during a system time out. On the other hand if the driver was performing a more complicated voice command like entering an address or scheduling a

calendar appointment, the speaking subtask would become cognitively more demanding.

The analysis in this chapter demonstrated that VCS interaction can be characterized in terms of different cognitive processes or subtasks. Identifying various VCS subtasks is a useful way to explain the rapid changes in cognitive workload for a VCS interaction. Chapter 6 investigates the limits of feasible subtask for a VCS interaction. How much listening, speaking, and reading can a driver carry out before cognitive workload greatly elevates and task performance diminishes. This research question is answered with a second driving simulator experiment.

Chapter 6

VCS DRIVING SIMULATOR STUDY TWO: TESTING THE LIMITS OF VCS SUBTASKS

6.1 Introduction

Using different cognitive processes to characterize a VCS interaction is a useful way to explain the variability in cognitive workload. The analysis in Chapter 5 demonstrated that speaking had higher cognitive demands when compared to listening, waiting, or confirmation. Cognitive workload of speaking however is also dependent on the complexity of the voice command itself. A simple voice command such as saying the name of a radio station is cognitively less demanding than using voice to input an address or schedule calendar appointment. The research question now shifts to how much listening, speaking, and visual search a driver can engage in before cognitive workload starts to elevate and driving and task performance starts to deteriorate. Understanding the limits of listening, speaking, or visual searching while driving can provide more useful and in depth guidelines to improve VCS designs.

A second driving simulator study was conducted to test limits of a feasible VCS interaction. Study participants had to perform a point of interest navigation task (e.g. "Find nearest Korean restaurant"). Task difficulty or complexity was greatly increased by having the participant navigate to a restaurant with more complex search criterias (e.g. "Find a Korean restaurant with a 4 star rating and less than 5 miles away"). Furthermore, this driving simulator included a significant visual component as the participant had to look on a display and visually search for a restaurant that best matches the search criteria. To answer the research question whether complex VCS interactions impacts the cognitive workload of

different cognitive processes (e.g. listening, speaking, visual search), Chapter 6 tests the following hypothesis:

1. Does TDRT response time and miss rate increase with more complex and lengthy voice commands? (H1)

2. Is there a difference in TDRT performance for VCS subtasks such as listening, speaking, visual search and confirmation? (H2)

3. Does TDRT response time and miss rate increase when visual scanning effort is increased during the visual search subtask? (H3)

6.2 Methods

The experiment is based on a secondary-task driving simulator study. The study participant was expected to multi-task between the following:

1. Drive a vehicle using the driving simulator
2. Perform a voice task (Point of Interest Navigation Task)
3. Respond to a randomly occurring tactile stimuli

This driving simulator study used the same hardware setup and participant recruiting strategy as VCS Driving Simulator Study 1 in Chapter 4. Please refer to Sections 4.3 and Sections 4.4 for more details on participant screening and equipment setup. The main difference between study and VCS Driving Study 1 is the design of the voice task interaction. Instead of a radio channel selection, address entry, or calendar appointment entry, the participant had to do a point of interest navigation task (e.g. "Find nearest Korean restaurant").

6.2.1 VCS Navigation System

Participants had to perform a secondary task which a point of interest navigation task using using voice commands. Instead of entering an address with a house number and street name, the participant searched for a restaurant using voice command like *"find nearest Italian restaurant"* or *"find a restaurant with 4 stars"*. A list of restaurants that matches the query would then appear on a 7 inch screen display located on the right side of the driver (similar location to a central stack). Similar to the VCS driving simulator study in Chapter 4, the study also used the Wizard of Oz protocol instead of a VCS that is commercially available. Figure 6.1 is a screen shot of the interface the driver is interacting with.

Figure 6.1: Point of Interest navigation task

6.3 Procedure

Prior to the start of experiment, participants had to read and sign an informed consent form. Participants underwent training where they had the opportunity to get familiar with driving the vehicle in the simulator, get comfortable with operating the VCS, and get familiar with the TDRT. During the main drive, participants were asked to follow the lead vehicle while maintaining 50 mph (80 km/h) while simultaneously performing point of interest navigation task and also responding to the tactile stimuli.

Table 6.1 shows a typical interaction between the participant and the VCS. The point-of-interest navigation task is a multi-step process that involves various cognitive processes such as listening to instructions, speaking or using voice to input a command, and visually scanning a list.

Table 6.1: Interaction between Participant and VCS

Step	**Actor**	**Dialog/Action**	**Subtask**
1	Instructions:	*"Find a three star Thai restaurant"*	Listening
2	Participant:	*"Find a three star Thai restaurant"*	Speaking
3	VCS:	*"Here are the listings of restaurants nearby"*	
4	Particpant:	(Participant visually scans a list of restaurant to find one that best matches the instructions)	Visual Search
5	Participant:	*"Select **Best Thai** restaurant"*	
6	VCS:	*"Starting Navigation"*	Confirmation

The point-of-interest navigation tasks essentially are short term memory recall tasks combined with a visual search task. The participant first needs to recall the voice command given in the instructions. Next, the participant visually searches a list of restaurant (see

Figure 6.1) and selects the restaurant that best matches the instructions.

The main experiment consisted of a two 20 minute sessions, where there were 30 navigation tasks in each session for a total of 60 voice tasks. After completing all the voice tasks, participants were compensated $30 USD for their time, and earned additional $0.10 USD for each correct restaurant selection made (for maximum possible compensation of $36 USD). The goal of performance incentive bonus is to help motivate or nudge participants to perform a relatively distracting action of taking their eyes off road to visually search a list of restaurant and find one that best matches the instructions. Otherwise participants can simply ignore performing a meticulous visual search and just randomly select a restaurant from the list.

6.4 Independent Variable

6.4.1 Subtasks: 4 Levels

A VCS interaction is a multistep procees where the participant could either be *listening* to for instructions, *speaking* or issuing a voice command, *visually searching* a list of restaurants, or making a selection and listening to a *confirmation*. The subtasks that are considered for the study are as follows:

Listening An automated voice gives instructions to the participant to perform a point-of-interest navigation task (e.g. "Find a three star Thai restaurant"). During the *listening* subtask, the participant needs to encode the information in working memory.

Speaking After being given instructions, the participant issues the proper voice command to complete the voice task task. During the *speaking* subtask, the participant needs to be able to retrieve the voice command from working memory and verbalize it.

Visual Search A list of restaurants appear on the display after a correct voice command has been given. The participant then proceeds to visually search a list of restaurants

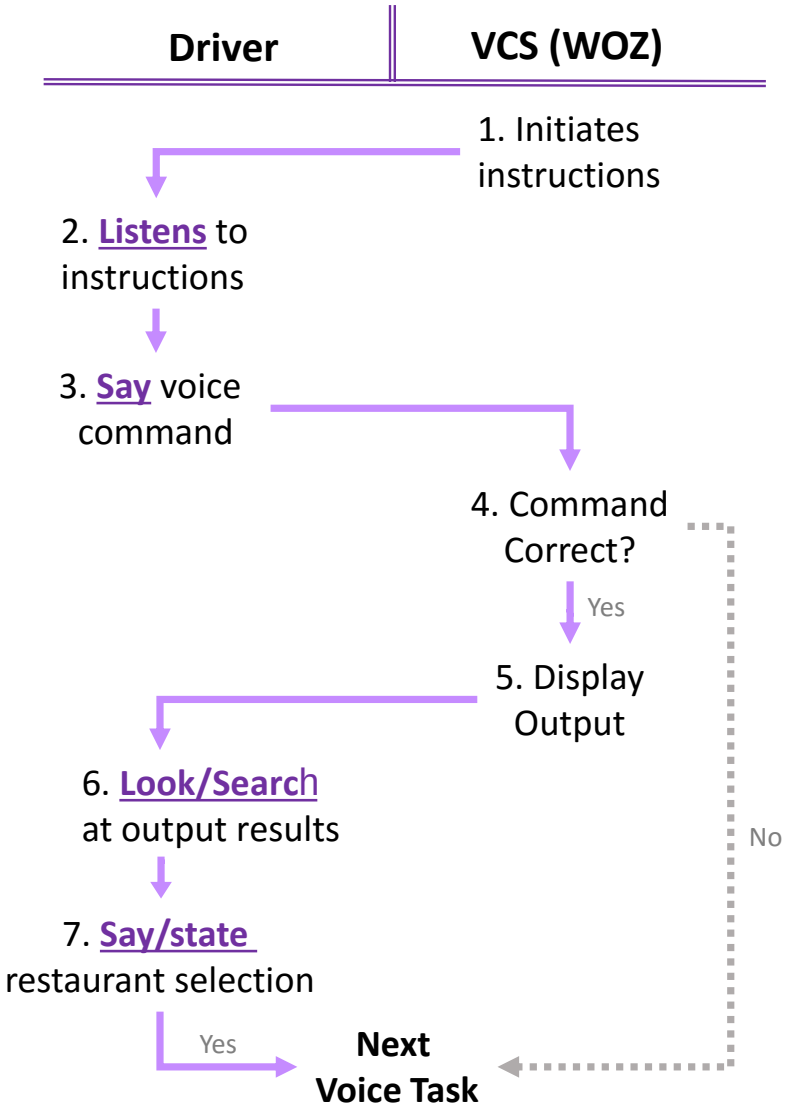


Figure 6.2: Task Analysis of Driving Simulator Study 2. In Driving Simulator Study 1, participant has the option to look or listen to VCS in Step 6 (see Figure 4.7). In Driving Simulator Study 2, looking at the display is mandatory.

as shown in Figure 6.1 and selects a restaurant that best matches the instructions from the *listening* subtask. During this subtask, participant is forced to take eyes off the forward road scene and make glances on a secondary display.

Confirmation Once a restaurant selection has been made, the VCS visually and auditorily provides the following confirmation prompt: *"Starting Navigation"*. The participant then waits for the next voice task to start.

Table 6.1 maps the dialog or actions to a subtask. It is important to note that the subtasks of *listening*, *speaking*, *visual search*, and *confirmation* are labels to a step or interval of time in the VCS interaction. The subtask names are not a precise descriptor of the cognitive process that is taking place. For example, during the *confirmation* subtask, the participant could be performing the cognitive process of "listening" to the VCS saying *"Starting Navigation"*. The "listening" that is taking place in Step 1 in Table 6.1 is information encoding where the information needs be recalled later. The "listening" in Step 6 (*"Staring Navigation"*) is just filler words that is not required for any recall, and the participant can choose to ignore it.

6.4.2 Cognitive Complexity: 5 Levels

Cognitive Complexity describes the difficulty of voice command. An example of a simple voice command is to *"find a Thai restaurant"*. Voice commands are made more complex by increasing the amount of information or number of search parameters. The list below shows voice commands with increasing levels of cognitive complexity.

C1: *"Find a Thai restaurant"*

C2: *"Find a restaurant with 55 reviews and is 15 miles away"*

C3: *"Find a 3 star Thai restaurant with 2 dollar signs"*

C4: *"Find a restaurant 12 miles away, 3 dollar signs, 4 stars and is Thai"*

C5: *"Find a 3 star Thai restaurant with 2 dollar signs, is 15 miles away, and has 40 reviews"*

Cuisine type, star rating, price rating (dollar sign), distance, and number of reviews are the only viable search parameters. The voice commands are short term memory recall tasks and at higher complexity levels, the participant needs to recall more chunks of information. If the participant fails to recall the instruction prompt, the VCS will skip to the next navigation task.

6.4.3 Scanning Effort: 2 Levels

Scanning effort only applies to the visual search subtask and it is the degree of difficulty to visually locate and select the restaurant that best matches the instruction prompt. There are two levels, *Scanning Effort Low* and *Scanning Effort High*.

Table 6.2: Scanning Effort: Low vs High

	Scanning Effort	
	Low	High
Line location for best restaurant	Always line 1	Any line #
# of restaurants that match task instruction	1 or 2	Always 3

Under *Scanning Effort Low* condition, the restaurant that best matches the task instructions is always displayed on Line 1. Using Figure 6.1 as a reference, suppose the instruction is to "*find a Indian restaurant*". There is only one *Indian*, restaurant and it is listed in Line 1. With *Scanning Effort High*, there will always be 3 restaurants that matches the instructions. Participant needs determine out of the unused categories, which restaurant is the

best. In order to find the best restaurant, the driver or participant is always forced to make comparisons between similar alternatives. He or she needs to figure out which restaurant has the highest star rating, lowest price, closest distance, and most reviews among 3 similar alternatives.

Using Figure 6.1 as a reference, an example of the *Scanning Effort High* condition is being instructed to *"find a restaurant with 4 stars and 1 dollar sign"*. There are 3 restaurants that fulfill the condition of having 4 stars and 1 dollar sign (Line 2, Line 3, and Line 4). Out of the restaurants that have 4 stars and 1 dollar sign, the participant must look at the remaining unused categories (e.g. number of reviews, and distance), and determine which is the best restaurant. Line 2 (Mod Pizza) is the best choice since it has more reviews, and it is closest distance to the driver.

6.5 Dependent Variable

TDRT performance metrics were used for the analysis: *Response Time*, a continuous variable, *DRT Miss* events, a binary outcome. Response time defined as stimuli responded within the 200-2500 ms onset. This variable was based on the ISO 17488 (ISO/DIS 17488, 2015), where a tactile stimuli appears every 3-5 seconds.

The list below shows the conditions for DRT hits and misses. DRT hits are any responses within the 100-2500 ms interval. DRT miss are any instances where participant fails press the microswitch button despite the presence of a tactile stimuli. To prevent cheating and random guessing, multiple button presses or any response time outside the 100-2500 ms interval are marked as a DRT miss. Response time and DRT misses increases with higher cognitive workloads (Merat et al., 2007; Ranney et al., 2014).

DRT Hit

- Any response time within 100-2500 ms interval

DRT Miss

- Failing to respond respond to onset stimuli
- Any response time outside 100-2500 ms interval
- Multiple (>1) microswitch button presses during an interval

Since this study included a visual search task where the participant had to correctly pick out the best restaurant from a visual display, visual search performance was analyzed. Visual search accuracy binary outcome and it measures if the participant is able to select the restaurant that best matches the task prompt. A visual search task is coded as a failure under the following conditions:

- Selecting a restaurant that does not match the task prompt (e.g. instructions are to *"Find a 4 star Greek restaurant"*, and participant selects a 3 star Greek restaurant).
- Selecting a restaurant that matches task prompt, but it is not the best one given other available alternatives. For example, instructions are to navigate to a *4 star Greek restaurant*, and three search results match the instructions (Scanning Effort High condition). Participant fails the visual search task when the restaurant that is selected is not the closest, lowest price, and has most reviews.

6.6 Data Analysis

A 5×4 mixed-linear model with repeated measures was used to analyze TDRT response times (Eqn 6.1). *Subtasks* and *Cognitive Complexity* are within subject factors that were entered as fixed effects. Participant ID was entered as a random effect. A mixed effects logistic regression was used to analyze the binary outcome, *TDRT Miss* (Eqn 6.2). Once again, *Subtasks* and *Cognitive Complexity* were entered as fixed effects and Participant ID was entered as a random effect.

$$RT = \beta_0 + \beta_1(CognitiveComplexity) + \beta_2(Subtask) + \gamma_0 + \gamma_1(ParticipantID) + \epsilon \quad (6.1)$$

$$Miss = \beta_0 + \beta_1(CognitiveComplexity) + \beta_2(Subtask) + \gamma_0 + \gamma_1(ParticipantID) + \epsilon \quad (6.2)$$

In addition to TDRT metrics, visual search accuracy was also analyzed. A mixed logit model was used to analyze the binary outcome of whether or not participant selected the "best" restaurant. Age, Cognitive Complexity, and Scanning Effort was entered as fixed effects. Participant ID was entered as a random effect (Eqn 6.3).

$$\begin{aligned} VisualSearch = &\beta_0 + \beta_1(Age) + \beta_2(ScanningEffort) + \\ &\beta_3(CognitiveComplexity) + \gamma_0 + \gamma_1(ParticipantID) + \epsilon \end{aligned} \quad (6.3)$$

β**:** Fixed effects
γ**:** Random effects
ϵ**:** Residual Error

The R statistical package *lme4* (Bates et al., 2015) was used to perform both the linear mixed effects analysis and the mixed effects logistic regression analysis.

6.7 Results

6.7.1 Descriptive Analysis

Thirty two participants were recruited for the study (18 males, 14 females). The age ranged from 19-75 years old where the mean age was 39.0 ($SD = 17.5$) for males and 38.6 ($SD = 16.9$) for females. Table 6.3 shows the distribution of participants across different age groups and gender.

Table 6.3: Number of participants across Age Group and Gender

Age Group	Male	Female
18-24	6	5
25-39	3	3
40-54	5	3
55-75	4	3

Drivers were asked to drive on a straight road and maintain 50 mph (80 km/h), while using a VCS. Table 6.4 shows the mean speed and mean lateral position (MLP), and standard deviation in lateral position (SDLP) for participants of different age groups and genders. Despite being distracted with a VCS task, participants were able to maintain a driving speed very close 50 mph, with small degree of lane deviation. This outcome is expected since the driving scenario only consisted of straight road with no traffic, and the VCS for the most part allow drivers to keep eye glances on the forward road scene.

TDRT data was collected throughout the study, and there were a total of 11,312 TDRT responses collected for 32 participants. Participants averaged 955 TDRT events over a drive that lasted about 40 minutes. Table 6.5 shows the mean TDRT response times and miss rates for among different age groups and gender. In general, males had faster TDRT response times when compared to females. Furthermore, females in the 55-75 age group had the highest mean TDRT response times and miss rates.

It is also important to note that there are a lot of individual differences in TDRT response time performance. Figure 6.3 shows the mean TDRT response times for every participant when they are only driving the vehicle. Most participants were able instantaneously respond to the vibrating tactor and maintain a mean response time between 250-350 ms. However there were a number of participants whose mean response times exceed 500 ms despite not

Table 6.4: Speed (mph) and Lateral Position (ft) summary across Gender and Age Groups

Gender	Age Group	Speed	SD(Speed)	MLP	SDLP
Female	18-24	49.02	0.026	-0.336	0.977
	25-39	50.04	0.036	-0.332	0.869
	40-54	50.13	0.050	-0.635	0.974
	55-75	49.67	0.049	-0.585	1.083
Male	18-24	49.06	0.028	-0.411	0.825
	25-39	50.01	0.027	-0.327	1.011
	40-54	49.07	0.035	0.045	1.243
	55-75	49.89	0.051	-0.317	1.667

Table 6.5: Response Time (RT) in milliseconds and Miss Counts summary across Gender and Age Groups

Gender	Age Group	Total Count	Mean(RT)	SD(RT)	Miss Count	Miss Rate (%)
Female	18-24	2281	651.25	358.64	524	0.23
	25-39	1374	521.98	297	295	0.21
	40-54	1421	542.51	344.26	493	0.35
	55-75	1362	661.77	376.31	523	0.38
Male	18-24	2797	549.41	359.24	510	0.18
	25-39	1426	466.98	269.77	132	0.09
	40-54	2282	472.29	321.48	580	0.25
	55-75	1988	440.64	305.27	562	0.28

having to perform a voice task.

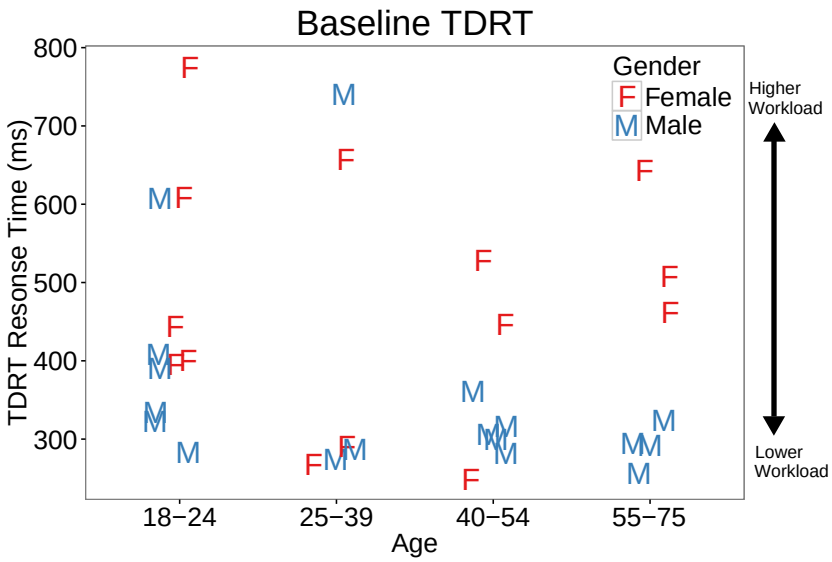


Figure 6.3: Mean Tactile DRT Response Time for Driving Only

6.7.2 Cognitive Complexity Analysis

The VCS navigation tasks are essentially memory recall tasks. Table 6.6 shows that memory recall decreased with higher levels of cognitive complexities or when participant had to remember more chunks of information. Every participant was able to recall a simple C1 level task prompt ("Find a 4 star restaurant"). However, out of all the C5 leveled voice tasks (e.g. "Find a 5 star Thai restaurant with 2 dollar sign, 100 reviews, and is 4 miles away"), only 29% of the total navigation voice tasks were successfully recalled.

It is clear from the memory recall results in Table 6.6, that drivers have difficulty encoding a voice command with many search parameters in the working memory and immediately recalling it verbally. This observed outcome is consistent with memory recall studies where recall performance decreased with word length (Baddeley, Thomson, & Buchanan, 1975;

Cowan et al., 1992). At higher levels of cognitive complexities, participants would also often mix up numbers between categories. Instead of saying the correct command of "Find a restaurant with **4 stars** and **2 dollar signs**", participants mix up numbers and say "Find a restaurant with **2 stars** and **4 dollar signs**".

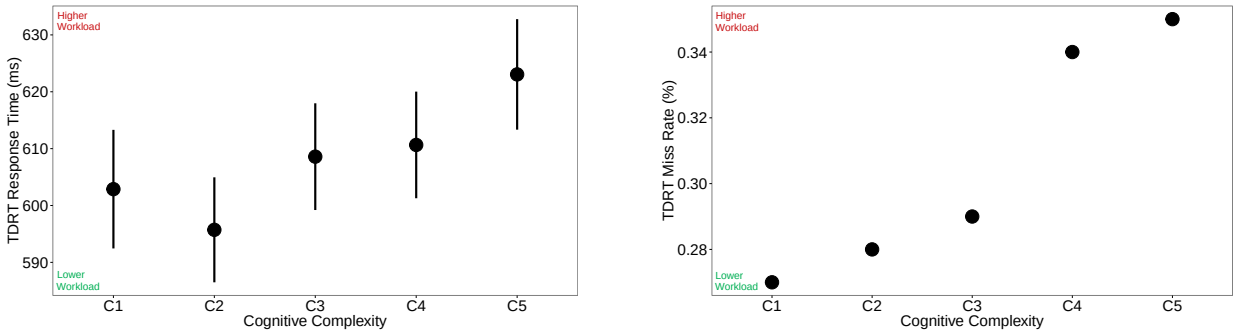
Table 6.6: Memory recall diminishes with increasing Cognitive Complexity

Cognitive Complexity	C1	C2	C3	C4	C5
Recall Success	100%	96%	84%	52%	29%

The broader question is whether the observed decline in memory recall correlates to increased workload as described in the ISO standards using TDRT metrics. The mixed linear model output in Table 6.7 showed that cognitive complexity was not a significant factor in predicting tactile TDRT response times. The mean TDRT response time for a C1 level voice task was not statistically different than a C5 (at $p < 0.05$). Figure 6.4a illustrates that there was high variability in TDRT response times across all cognitive complexity levels.

While TDRT response time was not sensitive to changes in cognitive complexity, significant differences was observed for TDRT misses. The mixed logit model in Table 6.8 showed that C4 and C5 leveled voice tasks are 1.43 ($z = 4.64, p < 0.001$) and 1.52 ($z = 5.29, p < 0.001$) times more likely to experience a TDRT miss when compared to C1 voice task. Furthermore, post hoc analysis in Figure 6.5 revealed that C4 and C5 voice tasks are even more likely to experience TDRT miss when compared to C2 and C3.

Since cognitive complexity was not significant in predicting TDRT response times, but significant factor for predicting TDRT miss events, this outcome suggests that TDRT misses is a more sensitive measure to detect differences in cognitive workload.



(a) Mean TDRT Response Time

(b) Miss Rate

Figure 6.4: TDRT Response Time and Miss Rates with Cognitive Complexity

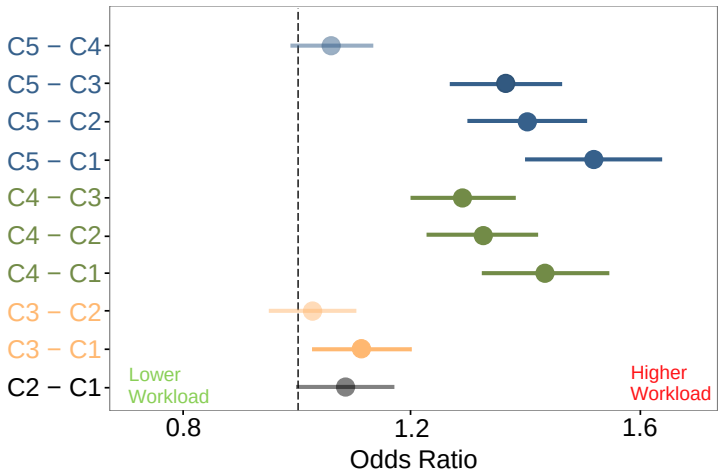


Figure 6.5: Post Hoc Tukey Contrast for TDRT Miss. C5 and C4 have 1.2 to 1.5 times more TDRT misses compared to C1, C2, and C3.

6.7.3 VCS Subtask Analysis

Mean TDRT response time was the highest for speaking subtask and lowest for the confirmation subtask as shown in Figure 6.6. TDRT miss rates also followed a similar trend where the speaking, and visual search subtask had the highest miss rates, which was then followed up listening and confirmation.

The mixed linear model in Table 6.9 affirmed that subtask was a significant main effect in predicting TDRT response times. Mean TDRT response time for listening was lower than speaking and visual search. The listening subtask however had higher mean TDRT response times when compared to the confirmation subtask.

Post hoc analysis in Figure 6.7 visualizes all possible Tukey contrasts for listening, speaking, visual search, and confirmation subtasks. Figure 6.7 demonstrated that response time for the confirmation subtask was significantly lower than listening, speaking, and visual search. TDRT response time for speaking however was not significantly different from visual search.

Table 6.7: Mixed Linear Model for log of TDRT Response Time

Fixed Effects	Estimate	StdError	t-value	p-value
(Intercept)	6.262	0.049	127.43	p<0.001
Ref: C1				
C2	0.008	0.017	0.43	N.S.
C3	0.020	0.017	1.15	N.S.
C4	0.028	0.018	1.58	N.S.
C5	0.042	0.018	2.34	0.079
$AIC = 9435.3$	$LogLik = -4707.9.1$			
$AIC_0 = 9547.8$	$LogLik_0 = -4770.9$			

Table 6.8: Mixed logit model for TDRT Miss. Cognitive complexity is significant main effect for predicting TDRT miss rates.

Fixed Effect	Estimate	StdError	z-value	p-value
(Intercept)	-1.256	0.203	-6.20	p<0.001
Ref: Cognitive Complexity C1				
C2	0.080	0.079	1.00	N.S.
C3	0.106	0.078	1.35	N.S.
C4	0.361	0.078	4.64	p<0.001
C5	0.418	0.079	5.29	p<0.001
$AIC = 11306$	$LogLik = -5643.7$			
$AIC_0 = 11590$	$LogLik_0 = -5793.2$			

This outcome suggests that the speaking subtask and visual search subtask require similar amount of cognitive demands.

VCS subtasks was also a significant main effect for predicting TDRT miss events. The mixed logit model in Table 6.10 demonstrated that the speaking and visual search are 1.22 and 1.56 times more likely to experience a TDRT miss when compared to listening. The confirmation subtask on the other hand is 2.17 times less likely to experience TDRT miss when compared to listening. The TDRT miss rate analysis mirror the outcomes for TDRT response time analysis.

6.7.4 Visual Search Performance

Visual search was one of the subtasks with the highest TDRT response times and miss rates, and it was isolated for additional analysis. The visual search subtask itself is multi-layered interaction, as it requires various cognitive processes to complete. The participant needs to be

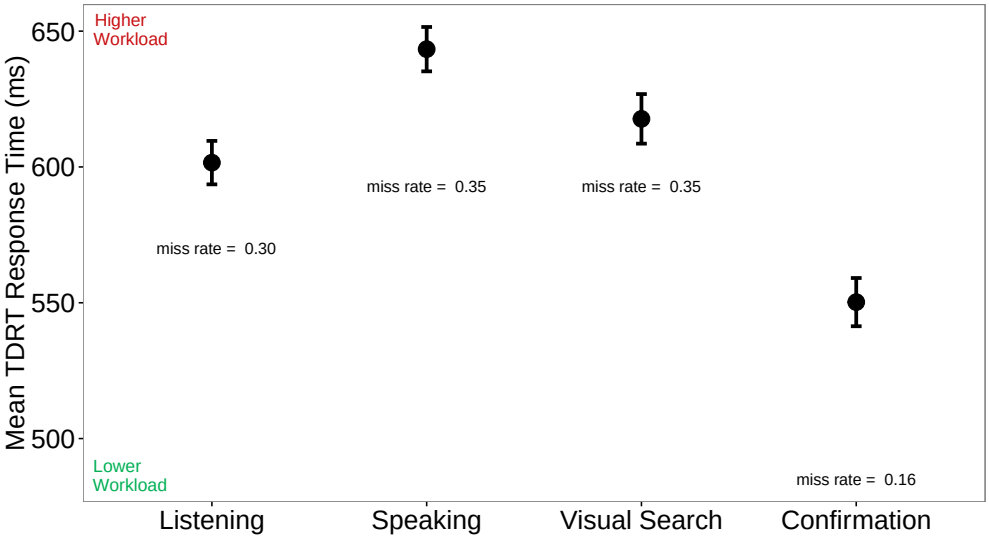


Figure 6.6: Mean TDRT Response Time and Miss Rates for VCS Subtasks

able to recall the task prompt (instructions from the listening subtask), map the task prompt information to the display as shown in Figure 6.1, and filter out irrelevant information. In this experiment, the visual search subtask was also manipulated with scanning effort level.

Table 6.11 shows that visual search accuracy decreased with scanning effort. When participants were presented a list of restaurant with 3 similar alternatives (Scanning Effort High), they frequently failed to identify the best restaurant. Most participants simply selected a restaurant that matches the task prompt and moved on to the next task instead of carefully comparing and contrasting other search category to determine which restaurant is the best among 3 similar alternatives.

Figure 6.8 shows that mean eyes-of-road glance duration were longer for high scanning effort, however the differences between low and high scanning effort were not significant (at $p < 0.05$). Since visual search accuracy decreased with scanning effort but mean eyes-of-road glance durations remain relatively the same, this outcome suggests that participants are not

Table 6.9: Mixed Linear Model results for log TDRT Response Time. VCS subtask is a significant main effect for predicting TDRT response times.

Fix Effects	Estimate	StdError	t-value	p-value
(Intercept)	6.285	0.0474	132.62	p <0.001
Ref: Listening				
Speaking	0.075	0.014	5.32	p <0.001
Visual Search	0.063	0.015	4.22	p <0.001
Confirmation	-0.086	0.016	-5.34	p <0.001
$AIC = 9434.1$	$LogLik = -4711.1$			
$AIC_0 = 9547.8$	$LogLik_0 = -4770.9$			

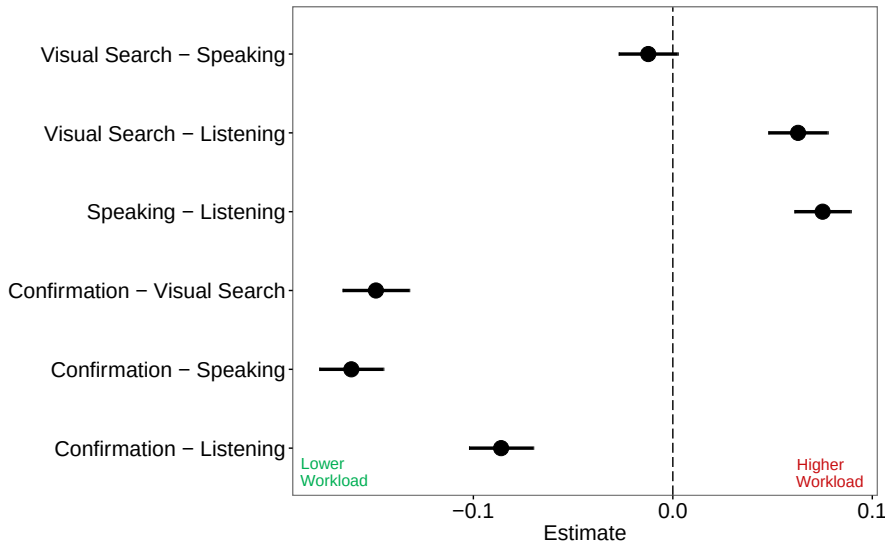


Figure 6.7: Post Hoc Tukey Contrast for TDRT Response Time

Table 6.10: Mixed Logit Model for TDRT Miss. VCS subtask is a significant main effect for predicting TDRT misses.

Fixed Effect	Estimate	StdError	z-value	p-value
(Intercept)	-1.256	0.203	-6.20	p<0.001
Ref: Listening				
Speaking	0.197	0.059	3.34	p<0.001
Visual Search	0.447	0.064	6.99	p<0.001
Confirmation	-0.774	0.087	-8.89	p<0.001
$AIC = 11306$	$LogLik = -5643.7$			
$AIC_0 = 11590$	$LogLik_0 = -5793.2$			

making the necessary effort in identifying the best restaurant.

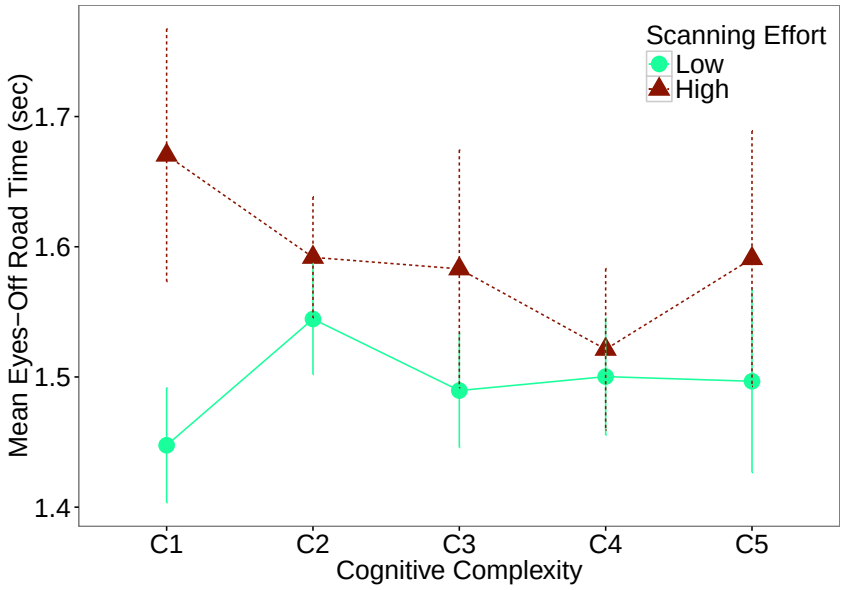


Figure 6.8: Mean Eyes-Of-Road Time

Table 6.11: Visual Search performance

	Cognitive Complexity				
Scanning Effort	C1	C2	C3	C4	C5
Low	95%	86%	92%	92%	88%
High	64%	48%	63%	67%	73%

Table 6.12 is the output of a mixed logit model for visual search performance. Visual search success decreased with older age ($z = -3.02, p < 0.001$). Younger drivers were more successful at reading an interface like Figure 6.1 and selecting a restaurant that best matches the task prompt.

Based on results of visual search accuracy analysis, it is evident that scanning effort increased cognitive workload of a visual search task. TDRT performance metrics (response time and miss rate) on the other-hand were not sensitive to increased cognitive workload for the a visual search task. Figure 6.10 illustrated that TDRT response time and miss rates do not increase with scanning effort. Furthermore, the Chi-Squared likelihood of ratio tests in Table 6.13 confirmed that models with cognitive complexity or scanning levels as predictors were not significantly different from the null model (only use intercept as predictor).

Table 6.12: Mixed Logit Model for Visual Search performance

Fixed Effects	Estimate	StdError	z-value	p-value
(Intercept)	1.450399	0.333546	4.348	p<0.001
Age	-0.021251	0.007036	-3.02	p<0.01
Ref: Scanning Effort High				
Low	2.473531	0.367438	6.732	p<0.001
Ref: Cognitive Complexity C1				
C5	0.360562	0.351845	1.025	N.S.
C3	-0.096725	0.241367	-0.401	N.S.
C4	0.053588	0.29021	0.185	N.S.
C2	-0.76016	0.22168	-3.429	p<0.001
Ref: Scanning Effort High × C1				
Low×C5	-1.610128	0.62705	-2.568	p<0.05
Low×C4	-0.96769	0.532788	-1.816	N.S.
Low×C3	-0.359826	0.496974	-0.724	N.S.
Low×C2	0.253391	0.482297	0.525	N.S.
$AIC = 1239.3$	$LogLik = -607.65$			
$AIC_0 = 1467.7$	$LogLik_0 = -731.85$			

Table 6.13: Chi-Square Tests comparing various models

Model	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	p-value
$Null_{RT}$	3	2344	2360	-1169	2338			
$Mod1$	4	2346	2368	-1169	2338	0.0051	1	0.9429
$Mod2$	7	2350	2389	-1168	2336	1.6428	3	0.6497
$Null_{Miss}$	2	3080	3092	-1538	3076			
$Mod3$	3	3082	3100	-1538	3076	0.0281	1	0.8669
$Mod4$	6	3083	3119	-1536	3071	4.5543	3	0.2075

$Null_{RT}$	$RT = 1 + \gamma(ParticipantID)$
$Mod1$	$RT = \beta(Scanning) + \gamma(ParticipantID)$
$Mod2$	$RT = \beta(CognitiveComplexity) + \gamma(ParticipantID)$
$Null_{Miss}$	$Pr(TDRTMiss) = 1 + \gamma(ParticipantID)$
$Mod3$	$Pr(TDRTMiss) = \beta(Scanning) + \gamma(ParticipantID)$
$Mod4$	$Pr(TDRTMiss) = \beta(CognitiveComplexity) + \gamma(ParticipantID)$

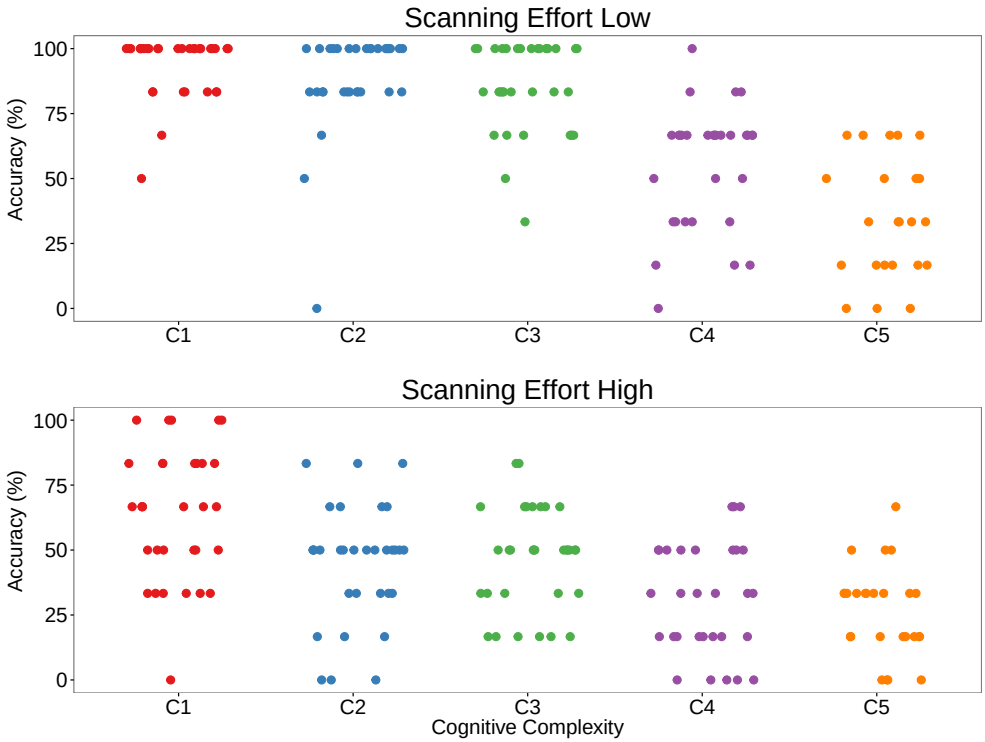
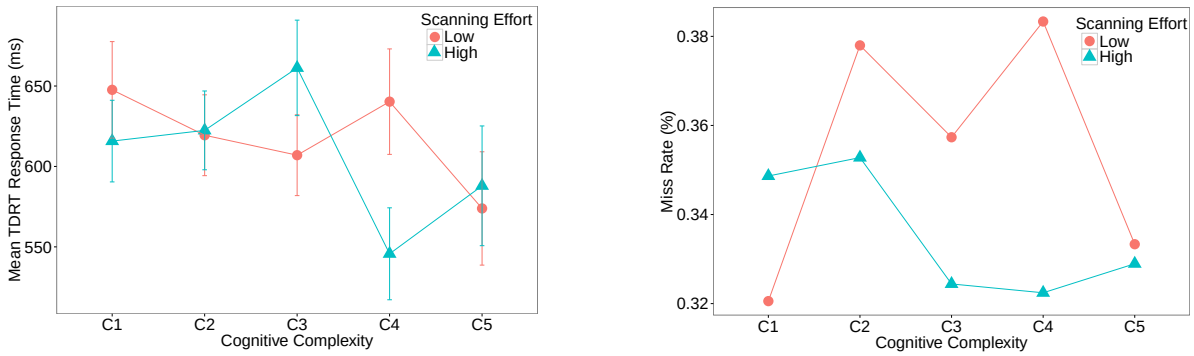


Figure 6.9: Visual Search Accuracy diminishes with High Scanning Effort

6.8 Discussion

VCS's are multi-layered interactions as a single task can involve various cognitive processes such as listening, speaking, or reading, and each cognitive process can have different contribution to cognitive workload. The driving simulator study in Chapter 6 explored how much listening, speaking, visual search a driver can engage in before cognitive workload elevates. Are people able to perform a verbal search query such as *"Find a Greek restaurant with 4 stars, 2 dollar signs, over 100 reviews and less than 10 miles away"*? Are people able to visually scan a list while driving and pick out the best restaurant out of a list with similar alternatives? Finally is the TDRT protocol capable of picking up the differences between simple and difficult VCS interactions?



(a) Mean TDRT Response Time

(b) Miss Rate

Figure 6.10: TDRT performance for Visual Search Subtask

Cognitive complexity is the amount of information needed for memory recall. The first hypothesis is that cognitive workload increases with cognitive complexity (H1). The results showed that memory recall performance diminished with higher cognitive complexity levels. Drivers were unable to consistently recall and say a long and complicated voice command such as *"Find a Japanese restaurant with 4 stars, 2 dollar signs, 75 reviews, and is 5 miles away"*. The TDRT response time metric however was not sensitive to cognitive complexity, as mean response time not statistically different for different cognitive complexity levels. TDRT miss events on the other hand was sensitive to cognitive complexity. The probability of a TDRT miss increased with C4, and C5 cognitive complexity levels.

The results from cognitive complexity analysis suggests that TDRT miss is a more sensitive metric for cognitive workload when compared to response time. When a participant is multi-tasking between driving, using a VCS, and responding to the TDRT stimuli, it becomes difficult to prioritize between the three tasks, thus leading to inconsistent and highly variable TDRT response times. Should the participant focus on the TDRT and sacrifice memory recall performance or should the participant focus on saying the correct voice command and worry less about fast TDRT response times? A TDRT miss on the other hand signifies

that the participant is overwhelmed with driving and speaking to the VCS, therefore he or she is incapable of attending the TDRT. TDRT miss is a better metric when participant is multi-tasking with at least 3 activities or when task prioritization becomes too difficult.

The second hypothesis tests if there are differences in cognitive workload for different VCS subtasks such as listening, speaking, visual search, and confirmation (H2). Study results showed that cognitive workload (both TDRT response time and TDRT Misses) was highest for speaking and visual search, while confirmation had the lowest cognitive workload. The cognitive workload of listening was in between speaking and confirmation. This study outcome is consistent with the findings in Chapter 5. Speaking or language processing is an intricate cognitive process. The elevated cognitive workload associated with speaking for this study was consistent with findings for research studies that involve driving and engaging in cellphone conversations (Nunes & Recarte, 2002; Strayer & Johnston, 2001). Visual search is also an intricate cognitive process. The driver needs to recall the task prompt and visually find a match between the task prompt and the list of restaurant that is displayed on the center console. In the mean time the driver also needs to look on the forward road scene from time to time in order to maintain lane position.

Given that visual search subtask is a complicated and intricate cognitive process, the visual search subtask was isolated for more in depth analysis including measures for cognitive workload, eyes-off-road glance durations, and visual search accuracy. It is hypothesized that cognitive workload increases with higher scanning effort (H3). The results however indicated that TDRT response times and miss rates did not significantly increase when going from a low scanning effort to a high scanning effort. There was also no significant differences in mean eyes-off-road glance durations between scanning effort low and scanning effort high. While there was no significant outcomes observed for TDRT metrics and eyes-off-road glance durations, visual search accuracy decreased when going from low scanning effort to high scanning effort. Participants a lot of difficulties selecting the best restaurant among three

similar looking restaurant choices.

The TDRT was not very sensitive to the increase in cognitive demand for the visual search subtask. The TDRT's lack of sensitivity for the visual search subtask can be explained with Wicken's Multiple Resource Theory. Since responding to a tactile stimuli (TDRT) and visually scanning a list use two different sensory inputs, it becomes easier to perform both tasks concurrently (Wickens, 2008). The results are also accordant with a study conducted by Young et al. where the TDRT was not sensitive to cognitive workload for a visual search task (2013).

Experiment results suggest that the TDRT protocol based on the ISO standard can be used to detect difference in cognitive workload between broadly defined cognitive process such as listening, speaking, and visual search. The TDRT protocol however is less sensitive to changes in cognitive workload within the same cognitive process (e.g. speaking with 1 chunk vs speaking with 5 chunks of information). It is important to use multiple performance metrics to get a better understanding of the cognitive workload associated with a VCS interaction. The Visual Detection Response Task, which uses light stimuli instead of a tactile stimuli, could potentially be a more sensitive to the changes in cognitive workload for the visual search subtask.

A study limitation is that participants were not explicitly instructed to prioritize TDRT over driving the vehicle and using the VCS to select the best restaurant, thus leading to higher variability with TDRT response time performance. If participants were told to prioritize the TDRT above everything else, the TDRT response times potentially become less variable and more sensitive to subtle changes in cognitive workload. However, if participants are told to prioritize the TDRT, they are then no longer incentivized to select the best restaurant and can potentially select restaurants at random. It remains to be determined if participants are even capable of deciding which task to perform first given driving, voice task, and TDRT all happening at the same time.

While there are inconsistencies with TDRT response times, TDRT miss is potentially a more robust performance metric for cognitive workload. A TDRT miss means that the task is so demanding that the participant is incapable of paying attention to the TDRT, therefore it is an absolute indicator of high cognitive workload. The downside to TDRT miss is that it is contingent on the fact that the voice task is difficult. If the voice tasks are too easy, there will be no or too few TDRT miss events to make any meaningful statistical analysis.

Despite the limitations with the TDRT protocol, the experiment nevertheless revealed limits to voice commands. First, study participants were not able to consistently say out loud really complex voice commands (e.g. "Find a Greek restaurant with 4 stars, 2 dollar signs, over 100 reviews and is 10 miles away"). Given the infinite possibility of a voice query, designers can consider limiting the voice command complexity to 3 chunks of information (C3: "Find a Greek restaurant with 4 stars, 2 dollar signs"), as more chunks of information required significantly higher cognitive demands.

The visual search subtask analysis also provided insights on how to design an interface for an in-vehicle display. Given that search accuracy decreased with higher scanning effort while TDRT metrics and eyes-off-road glance durations remained the same, this outcome suggests that participant did not make the necessary effort to find the best restaurant among 3 similar alternatives. In fact, participants compensated for the increased cognitive workload by adopting a guessing strategy despite being given a performance incentive bonus. This challenges the notion of how many search results should be allowed in a in-vehicle display as drivers had difficulties making visual comparisons between similar 3 search results. Displaying only 1 or 2 search results with fewer information (e.g. only list distance and rating) can greatly reduce cognitive workload of visual search.

The driving simulator study experiment in this chapter highlights the diversity of cognitive processes and the complexity of a VCS interaction. Understanding the cognitive workload of different VCS subtasks along with the complexity of a VCS verbal command can help guide

interface designers identify where improvements and refinements should be made. However, additional research is needed to better parse out the subtle differences in cognitive workload.

6.9 Chapter Summary

VCS interactions contain numerous cognitive processes that include listening, speaking, and visual search. A driving simulator experiment was conducted to test if a long and complicated voice command impacts (e.g. "Find a Japanese restaurants with 4 stars, 3 dollar signs, 200 reviews and is 7 miles away") listening, speaking, and visual search, and if such command is feasible to perform while driving a vehicle. Voice task performance such as memory recall rate and visual search accuracy was recorded, and cognitive workload was measured using the Tactile Detection Response Task (TDRT) protocol.

Memory recall rate decreased when a voice command was more complex. Based on study findings, it is recommended that voice commands should be limited to 3 chunks of information as cognitive workload (as measured by TDRT miss events) elevated and memory recall drastically diminished for voice commands with 4 and 5 chunks of information.

Visual search accuracy decreased when high visual scanning effort was required. Study participants had trouble identifying the best restaurant when visually comparing and contrasting three similar restaurants while driving the vehicle. Interface designers should be cognizant of the number of search results displayed.

The TDRT protocol was sensitive to changes in cognitive workload for different cognitive processes (e.g. listening, speaking, visual search, confirmation), but less sensitive to subtle changes in within the same cognitive process (e.g speaking with 1 chunk vs speaking with 5 chunks of information). Given the diverse cognitive processes that make up a VCS interaction, other cognitive workload measurements and performance metrics are needed to supplement TDRT results to get a better understanding of in-vehicle cognitive distraction.

This experiment used various performance metrics such as task completion rate, cognitive

workload, and eyes-off-road glance measures to gain insights to the decision making process and compensatory behavior of the driver during a visual search subtask. Cognitive workload and eyes-off-road glance metrics did not increase with higher scanning effort requirements, however, visual search accuracy still greatly decreased. This trend suggests that participants compensated for the increased complexity of the visual search subtask completing the tasks in a quicker and more efficient manner (e.g. making educated guesses), instead of carefully peering over each selection choices for greater accuracy to maximize monetary gains for the performance incentives.

Chapter 7
GENERAL CONCLUSIONS

This chapter summarizes overall findings of the dissertation, discusses the contribution of the results to the research field, addresses study limitations, and presents future research topics that relate to the research aims.

7.1 Overall Summary

The overall objective of this dissertation is to measure the cognitive workload associated with VCS use. Furthermore cognitive workload was assessed for VCS interactions with performance errors. Finally this dissertation measured the cognitive workload of different VCS cognitive processes such listening, speaking, waiting, and visual search. It was hypothesized that cognitive workload is elevated when recognition errors and system time outs are present throughout the interaction. It was also hypothesized that there would be differences in cognitive workload for different VCS cognitive processes such as listening and reading. One contextual interview study and two driver simulator studies were conducted for this purpose, and the key findings are summarized as follow.

1. VCS errors were frequent occurrences in an on-road contextual interview study. Drivers used their own VCS', performed VCS tasks that they were familiar with, and drove their own vehicle. Despite being familiar with their VCS, up to 31% and 52% of navigation tasks resulted in an error for Maryland and Washington study sites respectively.

2. The TDRT protocol was observed to be sensitive to the cognitive demands of using a VCS in a driving simulator study. Cognitive workload was lower for the radio channel selection task when compared to a navigation and calendar appointment scheduling task.

This outcome was expected since selecting a radio channel is an easier voice command to perform compared to navigation and calendar appointment scheduling.

3. For some situations, cognitive workload was lower for a VCS interaction that had errors compared to a VCS interaction that had no errors. During a system time out error for example, the driver is waiting for an idle VCS to respond, and during the idle period the driver can allocate available attentional resources towards other activities such as the TDRT. This outcome suggests that VCS interactions are multifaceted and complex. Failing to account for the complexity can lead to confusing and counterintuitive outcomes.

4. A VCS interaction is complex as the driver can be performing a sequence of actions such as listening, speaking, waiting, and confirmation within the same task. TDRT outcomes showed that the speaking and reading subtasks had the highest cognitive workload, followed by listening, and waiting. The TDRT protocol is a suitable method to measure the cognitive workload of different VCS cognitive processes.

5. Voice commands should be limited to 3 chunks of information. Cognitive workload elevated and memory recall drastically diminished for longer and more complex voice commands that included 4 or 5 chunks of information. TDRT misses was a more sensitive metric to detect differences in voice command complexity when compared to TDRT response time.

6. Visual search accuracy decreased for a *point of interest* navigation task when high scanning effort was required. Drivers had difficulty identifying the best restaurant when visually comparing and contrasting between three similar restaurants while driving. Interface designers should be cognizant of the number of search results displayed, as higher scanning effort reduced search accuracy. Drivers compensated for the increased scanning effort demands of the visual search subtask by applying educated guessing strategies.

7.2 Theoretical Implications

VCS's that are commercially available provides poor user experiences as they frequently encounter performance errors. This research attempted to model the increase in cognitive workload associated with poor VCS performance. Tactile detection response task results showed situations where a poorly performing VCS had lower cognitive workload than a VCS with no performance problems. A low performing VCS does not necessarily correspond to higher cognitive workload. For example, during a VCS system time out when no engagement is needed with the VCS, the driver has actually freed up attentional resources for other activities. This dissertation investigates how varying attentional resources impacts whether VCS interactions are considered cognitively distracting or merely annoyances.

Prior research has examined the cognitive workload of an entire system such as the iPhone Siri (Strayer et al., 2014; Strayer et al., 2015) or the cognitive workload of a VCS task such as scheduling an appointment using voice (Harbluk et al., 2013). A system level or task level cognitive workload characterization hides and obscures the complexity of a VCS interaction, as it is a multi-step process with rapid changes in cognitive workload. This dissertation proposes to characterize VCS's in terms of its different cognitive processes (e.g. listening, speaking, visual search). This framework helps identify which cognitive process of the VCS interaction elevates cognitive workload and which cognitive process lowers cognitive workload.

Measuring cognitive workload during VCS use can provide insight on how to improve safety and design. However, examining only cognitive workload does not reveal the compensatory behavior that drivers undertake during increased cognitive loads. This dissertation further refines the understanding of attentional demands required during VCS use by analyzing multiple performance metrics (e.g. cognitive workload, task accuracy, eye glance, driving metrics). Using multiple performance metrics provides insights to the decision making process and how the driver compensates for the increase in cognitive workload when

performing a distracting VCS task.

7.3 Contributions

Voice control systems are becoming more accessible to drivers and contain more complex features that would not be normally allowed with a visual manual interface. While VCS's have demonstrated to be less distracting than visual manual interface, they still impose cognitive distractions on the driver. Cognitive workload was measured for a VCS interaction using the Tactile Detection Response Task (TDRT) protocol. Furthermore, cognitive workload was measured for a poorly performing VCS that contained recognition errors and system timeouts. Study outcomes showed a situation where a poorly performing VCS with recognition errors and system time outs had lower cognitive workload than a VCS interaction without any performance issues. This suggests that VCS's are complex multi-step interaction where different steps of the interaction can have different cognitive workload demands. The driving simulator study of measuring cognitive workload of non-optimal VCS interaction has been submitted to *Transportation Research Part F.*

This dissertation also describes a way to characterize and analyze a VCS interaction by segmenting it into different cognitive processes such as listening, speaking, reading, and waiting. Previous studies compared the cognitive workload between systems (e.g. iPhone Siri vs Ford SYNC), and between tasks (e.g. radio vs navigation). Comparing the cognitive workload between two different system or two different tasks does not necessarily explain why one is better than another. Characterizing a VCS interaction in terms of its cognitive process helps identify moments in the interaction that elevate or lower cognitive workload. This type of analysis helps better guide designers to areas where improvements can be made. The analysis for measuring the cognitive workload of speaking, listening, waiting, and confirmation of a VCS interaction will be submitted to *Human Factors.*

Prior research have not identified thresholds to the amount of listening, speaking, and

visual searching a driver can undertake. This dissertation assessed the cognitive workload and task performances for varying complexity levels of listening, speaking, and visual search. When a driver is speaking, the verbal command complexity should be limited to 3 chunks of information as cognitive workload elevated and memory recall declined for complex verbal commands that contained 4 or 5 chunks of information. Drivers also had difficulties visually identifying the best restaurant when similar restaurants choices were presented. This outcome questions the number of restaurant search results that should be displayed. The findings related to VCS Driving Simulator Study 2 in Chapter 6 will be submitted to *Accident Analysis & Prevention*.

7.4 Limitations

Driving Simulator Study 1 in Chapter 4 showed that there were significant difference in TDRT response times between the radio task and calendar task. This outcome was expected since the radio task is a voice command that consists of 1 chunk of information (*"Tune to* ***Classic Rock*** *station*"), while the calendar task contains 4 chunks of information (*"Meet* ***Luke*** *at* ***Starbucks*** *on* ***Tuesday*** *at* ***4 PM***"). Driving Simulator Study 2 in Chapter 6 however observed no significant difference in TDRT response times between C1 (*"Find a Japanese restaurant"*) voice task and a C5 (*"Find a Japanese restaurant with 4 stars, 2 dollar signs, 120 reviews, and is 6 miles away"*). The two driving studies were similar in design regarding the scaling of task difficulty, which was manipulated by increasing chunks of information required for recall. However, the results between the two studies were not consistent with each other. A possible explanation is that both studies used different TDRT instruments making cross study comparisons a challenge.

Another possible explanation is that point of interest navigation task is a very different cognitive process when compared to address or calendar appointment entry, and the complexity of the voice task should not be defined by the number chunks of information required

for memory recall. In Driving Simulator Study 1 for example, the participant for example is asked to navigate to a specific address (e.g. "Navigate to 4315 Main St, Seattle WA"). The information given to the participant is always structured in the order of street number, street name, city, and state. This ordered structure of a street address follows a format that most people are familiar with. Calendar appointment entry always follow the order of contact name, location, time, and day (e.g. "Meet Luke at Starbucks on Monday at 3 PM"). The point of interest navigation in Driving Simulator Study 2 however does not always deliver information in an uniform orderly structure. Participant could be told to "*find a Japanese restaurant with 3 stars and has more than 50 reviews*" or they could be told to "*find a restaurant with more than 50 reviews, has 3 stars, and is Japanese*". The inconsistent phrase structure found in the point of interest navigation task can lead to greater variability in cognitive demands.

There are also limitations with the experimental procedures. Study participants had to multi-task between driving a vehicle, using the VCS, and responding to the TDRT, however, no instructions were given on which task they should prioritize. If participants were instructed to prioritize the TDRT, this may lead to more consistent TDRT response time values. However, it remains to be determined if drivers are able to consistently prioritize TDRT over using the VCS or driving the vehicle, especially if a voice task and a TDRT stimuli happen to start at the same time.

The low sampling frequency and the randomness of the TDRT protocol can also lead to less consistent TDRT response times. The ISO standard defines that the onset stimuli appears randomly every 3-5 seconds, which is approximately a 0.25 Hz sampling rate. Pupillometry, which uses pupil diameter measurements to infer cognitive workload, logs data at 60 Hz or 60 data points every second. Furthermore, suppose the participant needs to listen and remember the following task prompt: *"Find a restaurant with 4 stars, 2 dollar signs, 150 reviews, and is 7 miles away"*. If the random stimuli appears during the statement *"Find*

a restaurant", this can be considered filler text and the TDRT response time are expected to be very quick. If a random TDRT stimuli appears within the statement *"4 stars"*, this is vital information and may slow the TDRT response times as the entire chunk is considered important. Not every word in a voice command has high contribution to cognitive workload, and the random sampling nature of the TDRT produces less consistent response time values.

There are also limitations to the practicality and implications of the study results. In the two driving simulator studies, participants received instructions on what voice task to perform. This is based on the scenario where the passenger tells the driver to perform a VCS task. Most often, drivers already know in advance what task they would like to perform, therefore the driving simulator studies presented in this dissertation may not be representative of common VCS use. For example, most drivers already know he or she wants to find a Korean restaurant with four stars instead of being told to do so.

Finally, the two driving simulator experiments in this dissertation defined listening, speaking, waiting, visual search, and confirmation as the subtasks or cognitive processes that constitutes a VCS interaction. There could be numerous other types of cognitive processes taking place throughout the VCS interaction that haven't been considered yet. This studied only analyzed listening, speaking, waiting, visual search, and confirmation subtasks because they were the only subtasks that were observable in an interval of time. An alternative approach is to analyze the cognitive workload of filler words and compare it to the cognitive workload of phrases that contain vital information.

7.5 Future Research

Driving Simulator Study 1 observed differences in cognitive workload between radio, address navigation entry, and calendar appointment scheduling, while Driving Simulator Study 2 observed no significant differences between a C1 and C5 point of interest task. Future research should compare the cognitive workload of a point of interest navigation task with

radio, address navigation, and calendar appointment scheduling task all using the same TDRT equipment and experimental procedures. This study would test if a less structured way of presenting information for recall as found in a point of interest navigation, has higher cognitive demands than a very structured way of presenting information like an address that consists of street number, street name, city, and state.

Future studies should compare the sensitivity of DRT with physiological cognitive workload measure such as heart rate or pupil diameter, which are capable of logging data at 60 Hz (60 data points every second) or more. VCS interactions are complex and more data resolution is needed to better understand the complexities. The DRT methodology based on the ISO standard is essentially random sample of cognitive workload, as one data point is collected randomly every 3-5 seconds (or 0.20 to 0.33 Hz). It is not clear that the random sampling nature of the DRT methodology captures the changes in cognitive workload. A continuous measure may be able to capture more subtle differences as drivers transition among subtasks.

Finally future studies should also consider the cognitive workload of each individual word or phrases. The studies presented in this dissertation measured the cognitive workload of a complete voice command instead of individual words. Speech production however is a very complex cognitive process, where a speaker needs to begin with an idea they want to convey and then arrange the sequence of sounds that constitutes the phonological content of the word to expresses their idea (Ferreira & Pashler, 2002). Understanding how each word or phrase contribute to cognitive workload provides insight on how to design a better VCS dialog system.

BIBLIOGRAPHY

Alm, H. & Nilsson, L. (1995). The effects of a mobile telephone task on driver behaviour in a car following situation. *Accident Analysis & Prevention*, *27*(5), 707–715.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975, December). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 575–589. doi:10.1016/S0022-5371(75)80045-4

Barón, A. & Green, P. (2006). *Safety and usability of speech interfaces for in-vehicle tasks while driving: a brief literature review* (tech. rep. No. UMTRI-2006-5). University of Michigan, Transportation Research Institute.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276.

Bengler, K., Kohlmann, M., & Lange, C. (2012, January). 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*, 4919–4923. doi:10.3233/WOR-2012-0786-4919

Berka, C., Levendowski, D. J., Cvetinovic, M. M., Petrovic, M. M., Davis, G., Lumicao, M. N., ... Olmstead, R. (2004). Real-time analysis of eeg indexes of alertness, cognition, and memory acquired with a wireless eeg headset. *International Journal of Human-Computer Interaction*, *17*(2), 151–170.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., ... Craven, P. L. (2007). Eeg correlates of task engagement and mental workload in vi-

gilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, *78*(Supplement 1), B231–B244.

Beyer, H. & Holtzblatt, K. (1997). *Contextual design: defining customer-centered systems*. Morgan Kaufmann Publishers.

Brookhuis, K. A., de Vries, G., & de Waard, D. (1991). The effects of mobile telephoning on driving performance. *Accident Analysis & Prevention*, *23*(4), 309–316.

Brouwer, W. H., Waterink, W., Wolffelaar, P. C. V., & Rothengatter, T. (1991, October). Divided Attention in Experienced Young and Older Drivers: Lane Tracking and Visual Analysis in a Dynamic Driving Simulator. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *33*(5), 573–582. doi:10.1177/001872089103300508

Burns, P., Parkes, A., Burton, S., Smith, R., & Burch, D. (2002). *How dangerous is driving with a mobile phone?: benchmarking the impairment to alcohol* (tech. rep. No. TRL547). TRL.

Carter, C. & Graham, R. (2000, July). 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), 3–286–3–289. doi:10.1177/154193120004402016

Cerella, J., Poon, L. W., & Williams, D. M. (1980). Aging in the 1980s: psychological issues. In L. W. Poon (Ed.), (Chap. Age and the complexity hypothesis, pp. 332–340). Washington DC: American Psychological Association.

Cooper, J., Ingebretsen, H., & Strayer, D. (2014, October). *Mental Workload of Common Voice-Based Vehicle Interactions across Six Different Vehicle Systems*. AAA Foundation for Traffic Safety. Washington, DC.

Cowan, N., Day, L., Saults, J. S., Keller, T. A., Johnson, T., & Flores, L. (1992, February). The role of verbal output time in the effects of word length on immediate memory. *Journal of Memory and Language*, *31*(1), 1–17. doi:10.1016/0749-596X(92)90002-F

Engström, J. (2010). The tactile detection task as a method for assessing drivers cognitive load. *Performance Metrics for Assessing Driver Distraction: The Quest for Improved Road Safety*, 90–103.

Engström, J., Aberg, N., Johansson, E., & Hammarback, J. (2005). Comparison between visual and tactile signal detection tasks applied to the safety assessment of in-vehicle information systems. In *Driving assessment 2005: 3rd international driving symposium on human factors in driver assessment, training, and vehicle design* (pp. 232–239).

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. & Markkula, G. (2007). Effects of visual and cognitive distraction on lane change test performance. In *Proceedings of the 4th international driving symposium on human factors in driver assessment, training, and vehicle design* (pp. 199–205).

Ferreira, V. S. & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1187–1199. doi:10.1037/0278-7393.28.6.1187

Foley, J. P. (2009). Now you see it, now you don't: visual occlusion as a surrogate distraction measurement technique. In M. A. Regan, J. D. Lee, & K. Young (Eds.), *Driver distraction: theory, effects, and mitigation* (pp. 107–121). CRC Press.

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, May). Evaluation of different speech and touch interfaces to in-vehicle music retrieval systems. *Accident Analysis & Prevention*, *42*(3), 913–920. doi:10.1016/j.aap.2009.12.022

Gellatly, A. W. & Dingus, T. A. (1998). Speech recognition and automotive applications: using speech to perform in-vehicle tasks. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 42, *17*, pp. 1247–1251). SAGE Publications.

Ginosar, S. & Hearst, M. (2014). A study of the use of current speech recognition in an information-intensive task. *Chi '14: proceedings of the sigchi conference on human factors in computing systems*, 61–68. ACM.

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology*, *33*(4), 457–461.

Haigney, D., Taylor, R., & Westerman, S. (2000). Concurrent mobile (cellular) phone use and driving performance: task demand characteristics and compensatory processes. *Transportation Research Part F: Traffic Psychology and Behaviour*, *3*(3), 113–121.

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 *Proceedings of the seventh international driving symposium on human factors in driver assessment, training, and vehicle design* (pp. 78–84).

Harbluk, J. L., Noy, Y. I., & Eizenman, M. (2002). *The impact of cognitive distraction on driver visual behaviour and vehicle control* (tech. rep. No. TP13889E). Transport Canada.

Hart, S. G. & Staveland, L. E. (1988). Development of nasa-tlx (task load index): results of empirical and theoretical research. *Advances in psychology*, *52*, 139–183.

ISO/DIS 17488: Road vehicles – Transport information and control systems – Detection-Response Task (DRT) for assessing attentional effects of cognitive load in driving. (2015). International Organization for Standardization. Geneva, Switzerland.

Jahn, G., Oehme, A., Krems, J. F., & Gelau, C. (2005, May). Peripheral detection as a workload measure in driving: Effects of traffic complexity and route guidance system use

in a driving study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*(3), 255–275. doi:10.1016/j.trf.2005.04.009

Jamson, A. H., Westerman, S. J., Hockey, G. R. J., & Carsten, O. M. (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.

Jenness, J. W., Lattanzio, R. J., O'Toole, M., Taylor, N., & Pax, C. (2002). Effects of manual versus voice-activated dialing during simulated driving. *Perceptual and Motor Skills*, *94*(2), 363–379.

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

Kun, A. L., Palinko, O., Medenica, Z., & Heeman, P. A. (2013). On the feasibility of using pupil diameter to estimate cognitive load changes for in-vehicle spoken dialogues. In *Interspeech* (pp. 3766–3770).

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

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.

Levy, J. & Pashler, H. (2008). Task prioritisation in multitasking during driving: opportunity to abort a concurrent task does not insulate braking responses from dual-task slowing. *Applied Cognitive Psychology*, *22*(4), 507–525.

Liang, Y. & Lee, J. D. (2010). Combining cognitive and visual distraction: less than the sum of its parts. *Accident Analysis & Prevention*, *42*(3), 881–890.

Lin, C.-T., Chung, I.-F., Ko, L.-W., Chen, Y.-C., Liang, S.-F., & Duann, J.-R. (2007). Eeg-based assessment of driver cognitive responses in a dynamic virtual-reality driving environment. *Biomedical Engineering, IEEE Transactions on*, *54*(7), 1349–1352.

Mattes, S. & Hallen, A. (2009). Surrogate distraction measurment techniques: the lange change test. In M. A. Regan, J. D. Lee, & K. Young (Eds.), *Driver distraction: theory, effects, and mitigation* (pp. 107–121). CRC Press.

McCallum, M., Campbell, J., Richman, J., Brown, J., & Wiese, E. (2004). Speech recognition and in-vehicle telematics devices: potential reductions in driver distraction. *International Journal of Speech Technology*, *7*(1), 25–33. doi:10.1023/B:IJST.0000004804.85334.35

McEvoy, S. P. & Stevenson, M. R. (2009). Measuring exposure to driver distraction. In M. A. Regan, J. D. Lee, & K. Young (Eds.), *Driver distraction: theory, effects, and mitigation* (pp. 73–83). CRC Press.

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

Mehler, B., Reimer, B., Coughlin, J. F., & Dusek, J. A. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record: Journal of the Transportation Research Board*, *2138*(1), 6–12.

Merat, N. & Jamson, A. H. (2008, February). The effect of stimulus modality on signal detection: implications for assessing the safety of in-vehicle technology. *Human factors*, *50*(1), 145–158.

Merat, N., Johansson, E., Engström, J., Chin, E., Nathan, F., & Victor. (2007, January). *Specification of a secondary task to be used in safety assessment of ivis* (tech. rep. No. IST-1-1507674-IP). AIDE.

Miller, S. (2001). *Workload measures* (tech. rep. No. N01-006). National Advanced Driving Simulator.

NHTSA. (2012). *Visual-manual nhtsa driver distraction guidelines for in-vehicle electronic devices* (tech. rep. No. NHTSA-2010-0053). National Highway Traffic Safety Administration (NHTSA), Department of Transportation (DOT).

Nunes, L. & Recarte, M. A. (2002, June). Cognitive demands of hands-free-phone conversation while driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *5*(2), 133–144. doi:10.1016/S1369-8478(02)00012-8

Speech-Based Interaction with In-Vehicle Computers: The Effect of Speech-Based E-Mail on Drivers' Attention to the Roadway. (2001, December). *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(4), 631–640. doi:10.1518/001872001775870340

Pashler, H. (1994). Dual-task interference in simple tasks: data and theory. *Psychological bulletin*, *116*(2), 220.

Patten, C. J., Kircher, A., Östlund, J., & Nilsson, L. (2004). Using mobile telephones: cognitive workload and attention resource allocation. *Accident analysis & prevention*, *36*(3), 341–350.

Patten, C. J., Kircher, A., Östlund, J., Nilsson, L., & Svenson, O. (2006, September). Driver experience and cognitive workload in different traffic environments. *Accident Analysis & Prevention*, *38*(5), 887–894. doi:10.1016/j.aap.2006.02.014

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

Putze, F. & Schultz, T. (2012). Cognitive dialog systems for dynamic environments: progress and challenges. In J. H. Hansen, P. Boyraz, K. Takeda, & H. Abut (Eds.), *Digital*

signal processing for in-vehicle systems and safety (pp. 133–143). Springer New York. doi:10.1007/978-1-4419-9607-7_8

Ranney, T., Baldwin, G. S., Vasko, S. M., & Mazzae, E. N. (2009, December). *Measuring distraction potential of operating in-vehicle devices* (Technical Report No. HS-811 231). National Highway Traffic Safety Administration.

Ranney, T., Baldwin, S., Smith, L., Mazzae, E., & Pierce, R. (2014, November). *Detection Response Task (DRT) Evaluation for Driver Distraction Measurment Application* (tech. rep. No. DOT-HS-812-077). National Highway Traffic Safety Administration.

Ranney, T., Harbluk, J., & Noy, Y. I. (2005). Effects of voice technology on test track driving performance: implications for driver distraction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *47*(2), 439–454.

Reimer, B., Mehler, B., Pohlmeyer, A., Coughlin, J. F., & Dusek, J. (2006). The use of heart rate in a driving simulator as an indicator of age-related differences in driver workload. *Advances in Transportation Studies*, 9–29.

Rupp, G. (2011). *Performance Metrics for Assessing Driver Distraction: The Quest for Improved Road Safety*. SAE International.

Schindhelm, R. & Schmidt, E. (2015, September). Evaluation of the tactile detection response task in a laboratory test using a surrogate driving set-up. *IET Intelligent Transport Systems*, *9*, 683–689(6).

Sodhi, M., Reimer, B., Cohen, J., Vastenburg, E., Kaars, R., & Kirschenbaum, S. (2002). On-road driver eye movement tracking using head-mounted devices. In *Proceedings of the 2002 symposium on eye tracking research & applications* (pp. 61–68). ACM.

Stelmach, G. E. & Nahom, A. (1992, February). Cognitive-Motor Abilities of the Elderly Driver. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *34*(1), 53–65. doi:10.1177/001872089203400107

Strayer, D., Cooper, J., Turrill, J., Coleman, J., & Hopman, R. (2015). *Measuring Cognitive Distraction in the Automobile III: A Comparison of Ten 2015 InVehicle Information Systems.* AAA Foundation for Traffic Safety. Washington, DC.

Strayer, D. & Drew, F. A. (2004). Profiles in driver distraction: effects of cell phone conversations on younger and older drivers. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(4), 640–649.

Strayer, D. & 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.

Strayer, D., Turrill, J., Coleman, J., Ortiz, E., & Cooper, J. (2014, October). *Measuring Cognitive Distraction in the Automobile II: Assessing In-Vehicle Voice-Based Interactive Technologies*. AAA Foundation for Traffic Safety. Washington, DC.

Tijerina, L., Parmer, E., & Goodman, M. J. (1998). Driver workload assessment of route guidance system destination entry while driving: a test track study. In *Proceedings of the 5th its world congress* (pp. 12–16).

Treat, J. R. (1980). A study of precrash factors involved in traffic accidents. *HSRI Research Review*.

Tsimhoni, O., Smith, D., & Green, P. (2004, December). Address Entry While Driving: Speech Recognition Versus a Touch-Screen Keyboard. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(4), 600–610. doi:10.1518/hfes.46.4.600.56813

Ursin, H. & Ursin, R. (1979). Physiological indicators of mental workload. In *Mental workload* (pp. 349–365). Springer.

Van Elslande, P., Fouquet, K. et al. (2007). *Analyzing 'human functional failures' in road accidents* (tech. rep. No. 027763-TRACE). TRACE.

van Winsum, W., Martens, M., & Herland, L. (1999). *The effects of speech versus tactile driver support messages on workload driver behavior and user acceptance* (tech. rep. No. TM-99-C043).

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 & Prevention*, *42*(3), 852–858.

Victor, T. W., Engström, J., & Harbluk, J. L. (2009). Distraction assessment method based on visual behavior and event detection. In M. A. Regan, J. D. Lee, & K. Young (Eds.), *Driver distraction: theory, effects, and mitigation* (pp. 135–165). CRC Press.

Vilimek, R., Schäfer, J., & Keinath, A. (2013). Effects of Task and Presentation Modality in Detection Response Tasks. In D. Harris (Ed.), (pp. 177–185). Springer Berlin Heidelberg.

Wickens, C. D. (2008, June). Multiple Resources and Mental Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 449–455. doi:10.1518/001872008X288394

Wickens, C. D., Gordon, S. E., Lee, J. D., & Liu, Y. (2004). *An introduction to human factors engineering* (2nd). Pearson Prentice Hall.

Widyanti, A., Johnson, A., & de Waard, D. (2013). Adaptation of the rating scale mental effort (rsme) for use in indonesia. *International Journal of Industrial Ergonomics*, *43*(1), 70–76.

Young, K. & Regan, M. (2007). Driver distraction: a review of the literature. *Distracted driving. Sydney, NSW: Australasian College of Road Safety*, 379–405.

Young, R. A., Hsieh, L., & Seaman, S. (2013). The tactile detection response task: preliminary validation for measuring the attentional effects of cognitive load. In *Proceedings of 7th international driving symposium on human factors in driver assessment, training and vehicle design* (pp. 71–77).