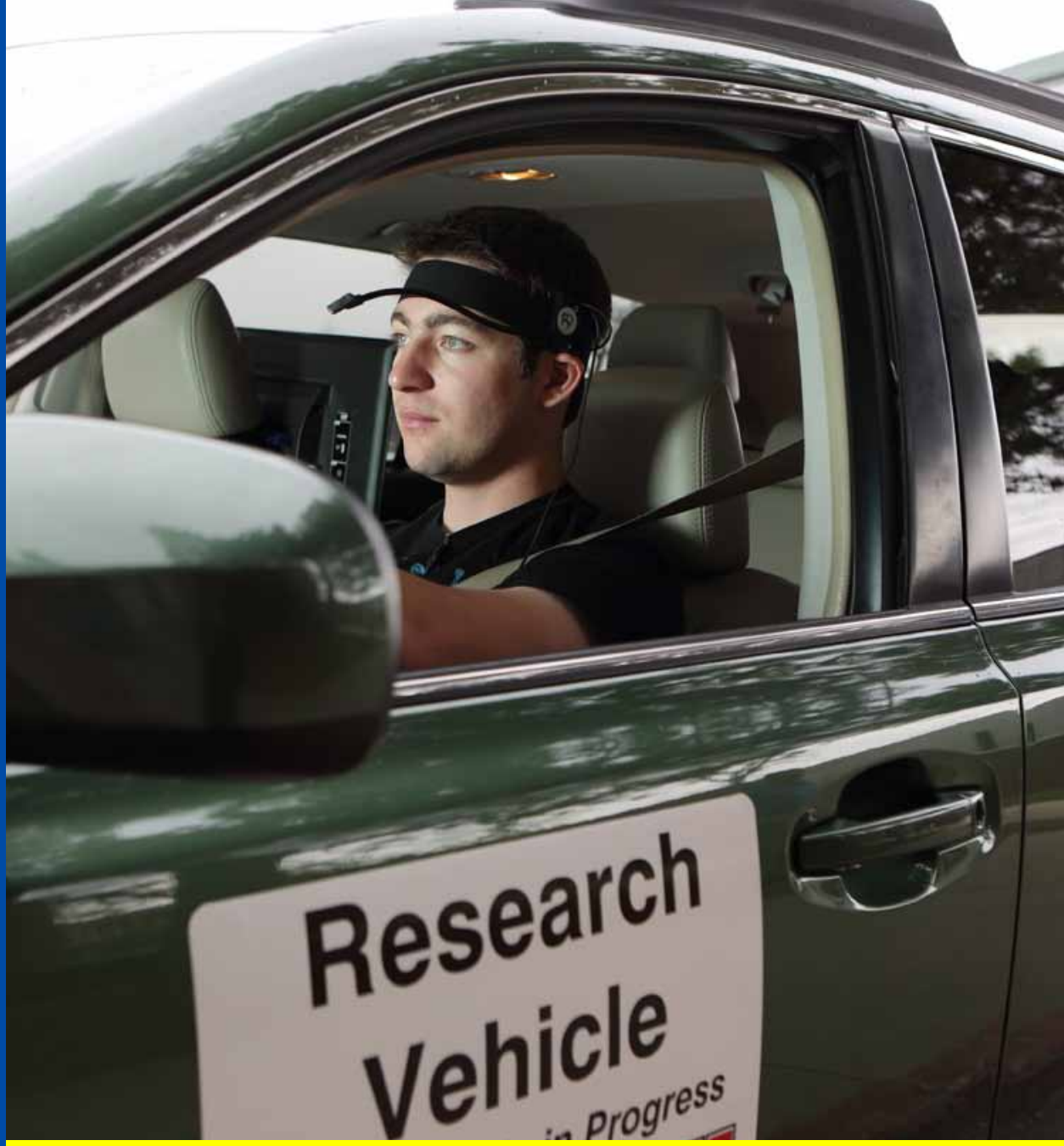
Car crashes rank among the leading causes of death in the United States.



Measuring Cognitive Distraction in the Automobile II: Assessing In-Vehicle Voice-Based Interactive Technologies

October 2014

AAA Foundation for Traffic Safety

607 14th Street, NW, Suite 201 | Washington, DC 20005 | AAAFoundation.org | 202-638-5944

Title

Measuring Cognitive Distraction in the Automobile II: Assessing In-Vehicle Voice-Based Interactive Technologies *(October 2014)*

Authors

David L. Strayer, Jonna Turrill, James R. Coleman, Emily V. Ortiz, & Joel M. Cooper
University of Utah

Acknowledgments

We acknowledge the assistance of Paul Atchley, Frank Drews, Jurek Grabowski, Donald Fisher, Peter Kissinger, Neil Learner, John Lee, Bruce Mehler, Daniel McGehee, David Sanbonmatsu, Jeanine Stefanucci, Brian Tefft, and Jason Watson, and for suggestions on improving the research described in this report.

About the Sponsor

AAA Foundation for Traffic Safety
607 14th Street, NW, Suite 201
Washington, DC 20005
202-638-5944
www.aaafoundation.org

Founded in 1947, the AAA Foundation in Washington, D.C. is a not-for-profit, publicly supported charitable research and education organization dedicated to saving lives by preventing traffic crashes and reducing injuries when crashes occur. Funding for this report was provided by voluntary contributions from AAA/CAA and their affiliated motor clubs, from individual members, from AAA-affiliated insurance companies, as well as from other organizations or sources.

Cover photo credit: Dan Campbell/AAA

Abstract

The goal of the current research was to measure and understand cognitive distraction stemming from voice-based technologies in the vehicle. Three controlled experiments evaluated 1) a baseline single-task condition, 2) issuing simple voice-based car commands, 3) listening to e-mail/text messages read by a "natural" pre-recorded human voice, 4) listening to e-mail/text messages read by a "synthetic" computerized text-to-speech system, 5) listening and composing replies to e-mail/text messages read by a "natural" voice, 6) listening and composing replies to e-mail/text messages read by a "synthetic" voice, 7) interacting with a menu-based system with perfect reliability, 8), interacting with a menu-based system with moderate reliability, and 9) using "hands-free" Siri to listen to and send text messages, update Facebook or Twitter status, and modify and review calendar appointments. Because each task allowed the driver to keep his or her eyes on the road and hands on the steering wheel, any impairment to driving must be caused by the diversion of attention from the task of operating the motor vehicle. We used a combination of primary-task, secondary-task, subjective, and psychophysiological indices to assess the mental workload of the driver using these voice-based technologies. The data extend the rating system for cognitive distraction developed by Strayer et al., (2013). The new ratings suggest that some voice-based interactions in the vehicle may have unintended consequences that adversely affect traffic safety.

Introduction

Background

Driver distraction, defined as "the diversion of attention away from activities critical for safe driving toward a competing activity" (Regan, Hallet, & Gordon, 2011; see also Engström, et al., 2013; Regan & Strayer, 2014), is increasingly recognized as a significant source of injuries and fatalities on the roadway. Indeed, a recent video analysis of crashes involving teen drivers found that 50 percent of the crashes involved driver distraction of one form or another (McGehee, in preparation). Other researchers have estimated that driver distraction and inattention account for somewhere between 25 percent and 75 percent of all crashes and near crashes (e.g., Dingus et al., 2006; Ranney, et al., 2000; Sussman, et al., 1985; Wang, Knipling, & Goodman, 1996).

Driver distraction can arise from visual/manual interference, for example when a driver takes his or her eyes off the road to interact with a device. Impairments also come from cognitive sources of distraction when attention is withdrawn from the processing of information necessary for the safe operation of a motor vehicle. In the latter case, the driver's eyes may be on the roadway and his or her hands on the steering wheel, but he or she may not be attending to the information critical to safe driving.

The National Highway Safety Traffic Administration (NHTSA) is in the process of developing voluntary guidelines to minimize driver distraction created by electronic devices. There are three phases to the NHTSA guidelines. The Phase 1 guidelines, entered into the Federal Register on March 15, 2012, address visual-manual interfaces for devices installed by vehicle manufactures. The Phase 2 guidelines, scheduled for release sometime in 2014, will address visual/manual interfaces for portable and aftermarket electronic devices. Phase 3 guidelines (forthcoming) will address voice-based auditory interfaces for devices installed in vehicles and for portable aftermarket devices.

In order to allow drivers to maintain their eyes on the driving environment, nearly every vehicle sold in the US and Europe can now be optionally equipped with a voice-based interface. Using voice commands, drivers can access functions as varied as voice dialing, music selection, GPS destination entry, and even climate control. Voice activated features may seem to be a natural development in vehicle safety that requires little justification. However, a large and growing body of literature cautions that auditory/vocal tasks may have unintended consequences that adversely affect traffic safety. What has become clear is that synthetic speech interactions can lead to surprisingly high levels of cognitive workload, well beyond that of natural conversation with another human (e.g., Strayer et al., 2013).

Cognitive distraction from voice-based interactions is difficult to assess because of the problems associated with observing what a driver's brain (as opposed to hands or eyes) is doing. Studies have found that when drivers divert attention to an engaging secondary task such as talking on a cellular phone, visual scanning is disrupted (Recarte & Nunes, 2000; Tsai et al., 2007; Victor, Harbluk, & Engström, 2005; Reimer, 2009), prediction of hazards is impaired (Taylor et al., 2013; Strayer et al., 2013), identification of objects and events in the driving environment is retarded (Strayer & Drews, 2007; Strayer, Drews, & Johnston, 2003), decision for action is altered (Cooper et al., 2009, Drews, Pasupathi, & Strayer, 2008), and appropriate reactions are delayed (Caird et al., 2008; Horrey & Wickens, 2006).

Phase I

In the first phase of our research, we developed a framework for assessing cognitive distraction in the vehicle (Strayer et al., 2013). The procedure involved comparing eight secondary-task conditions in three separate experiments. The first experiment served as a control in which participants performed the different tasks without the concurrent operation of a motor vehicle. In the second experiment, participants performed the same tasks while operating a high-fidelity driving simulator. In the third experiment, participants performed the tasks while driving an instrumented vehicle in a residential section of a city. We used a combination of primary-task, secondary-task, subjective, and psychophysiological indices of mental workload to develop a rating system of cognitive distraction where non-distracted single-task driving anchored the low-end (Category 1), and the mentally demanding Operation Span (OSPAN) task anchored the high-end (Category 5) of the scale. The workload ratings for the eight tasks are presented in Figure 1. In the figure, it is evident that some activities, such as listening to the radio or an audio book, were not very distracting. Other activities, such as conversing with a passenger or talking on a hand-held or hands-free cell phone, were associated with moderate/significant increases in cognitive distraction. Finally, there are in-vehicle activities, such as using a speech-to-text system to send and receive text or e-mail messages, which produced a relatively high level of cognitive distraction.

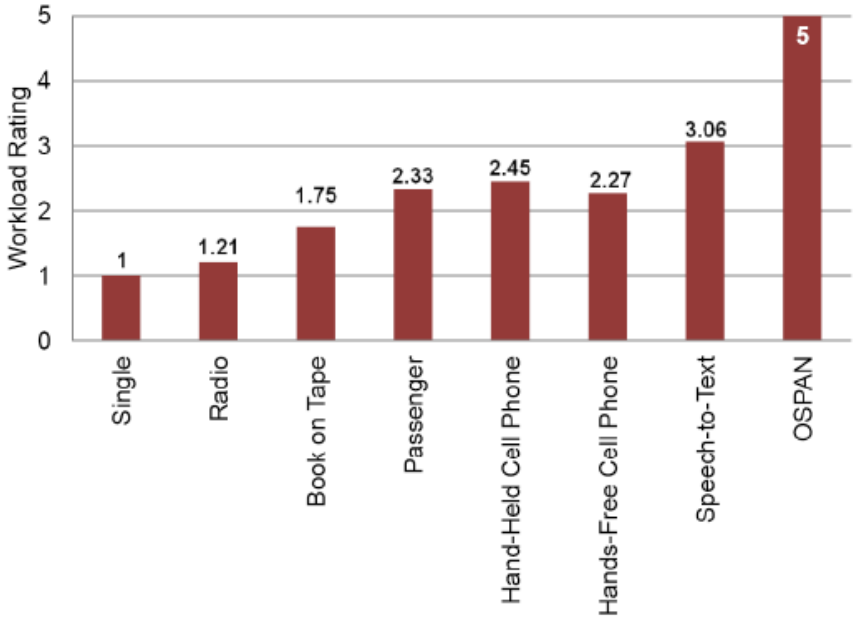


Figure 1. The cognitive distraction scale developed in Phase I of the research

The speech-to-text system that we evaluated read incoming text/e-mail messages using a commercially available computerized text-to-speech reader (NaturalReader 10.0) and a speech-recognition system with perfect reliability was implemented in which there was no requirement to review, edit, or correct garbled speech-to-text translations. Consequently, drivers did not need to take their eyes off the road or their hand off the steering wheel when making these voice-based interactions. Nevertheless, this condition received a Category-3 rating on the cognitive distraction scale, a level significantly higher than more traditional voice-based interactions on the cell phone. This highlights the need to better understand auditory/vocal interactions in the vehicle.

Research Objectives

The current research was designed to address four important issues related to voice-based interactions in the vehicle. First, what is the source of the workload associated with using speech-based e-mail/texting? This activity involves both listening to messages (primarily a comprehension task) and in some instances speaking to create a reply (involving both comprehension and speech production). How much of the increase in workload from a single-task baseline can be attributed to speech comprehension and how much to speech production? This issue has practical implications for the design of systems that allow the driver to listen to messages with or without the possibility of crafting a reply. In addition, current speech-based systems use "synthetic" computerized speech to read messages to the driver. How does the quality of the speech generated by the computer impact the driver's workload compared to situations where the message is delivered with a "natural" pre-recorded human voice? This issue also has practical implications because were meaningful differences between natural and synthetic speech observed, it would suggest that refinements to the text-to-speech technology might help to reduce cognitive workload when drivers make voice-based interactions in the vehicle. To test these issues, we employed a factorial design where natural vs. synthetic speech was crossed with conditions where the driver listened to messages without generating a reply vs. conditions where the driver listened and crafted a reply to messages where it was required.

A second question concerns the use of voice commands to control systems in the vehicle such as climate control and infotainment systems. Given that car commands are often short, scripted, and infrequent utterances, do they compare favorably with an activity like listening to an audio-book or talking on a cell phone? The use of voice commands in the vehicle is now ubiquitous in newer models and the user interface in the different systems is known to vary considerably. Does the mental workload associated with the different device manufacturers differ, or are they all essentially the same? If the systems do differ, are the differences negligible or are they striking?

Third, do menu-based systems that support navigation (e.g., locate the nearest ATM or gas station) incur a significant cost of concurrence and if so, how do they compare with other in-vehicle activities? Given that menu-based systems offer a limited selection of alternatives and a restricted set of responses, it is possible that the associated workload may be lower than the speech-to-text e-mail/texting system tested by Strayer et al., (2013). We also examined the extent to which the reliability of the menu-based interface affected distraction. In particular, we contrasted a menu-based system with perfect reliability with a system with moderate reliability, the latter being more representative of current in-vehicle systems. How does the reliability of the menu-based system impact cognitive workload? What factors could be adopted to reduce the working memory burden of the driver using these voice-based menu systems?

Finally, advanced voice recognition systems such as Apple's Siri offer the potential for the driver to issue commands and queries using natural language. Does this sort of voice-based interface reduce the level of cognitive workload compared to the speech-to-text system evaluated by Strayer et al., (2013)? Perhaps the natural language interface is similar in cognitive demand to that of a cell phone conversation. Based upon discussion with technical staff at Apple, we customized Siri so that the system was completely hands- and eyes-free.

To create a completely hands-free version, a lapel microphone was clipped to the driver's collar and they activated Siri with the command "*Hello Siri,*" at which point a researcher manually activate the device. The driver neither looked at nor made physical contact with the iPhone during these interactions. Therefore, any differences from the single-task baseline provide a pure measure of the cognitive workload associated with use of the system.

In this report, we present the results from three experiments designed to systematically measure cognitive workload associated with voice-based interactive technologies in the automobile. The first experiment served as a control in which participants performed nine different tasks without the concurrent operation of a motor vehicle. In the second experiment, participants performed the same nine tasks while operating a high-fidelity driving simulator. In the third experiment, participants performed the nine tasks while driving an instrumented vehicle in a residential section of a city.

In each of the three experiment, the order of the nine tasks was counterbalanced and the tasks involved 1) a baseline single-task condition (i.e., no concurrent secondary task), 2) issuing simple voice-based car commands, 3) listening to e-mail/text messages read by a "natural" pre-recorded human voice, 4) listening to e-mail/text messages read by a "synthetic" computerized text-to-speech system, 5) listening and composing replies to e-mail/text messages read by a "natural" voice, 6) listening and composing replies to e-mail/text messages read by a "synthetic" voice, 7) interacting with a menu-based navigation system with perfect reliability, 8) interacting with a menu-based navigation system with moderate reliability, and 9) using "hands (and eyes)-free" Siri to listen to and send text messages, update Facebook or Twitter status, and modify and review calendar appointments. Note that condition 6 of the current research is identical to the speech-to-text condition used by Strayer et al., (2013), thereby providing a direct comparison between the two research projects.

It is important to note that each task allowed drivers to keep their eyes on the road and both hands on the steering wheel so that any impairment to driving must stem from cognitive sources associated with the diversion of attention from the task of operating the motor vehicle. Based upon prior research (Strayer et al., 2013), these tasks were hypothesized to reflect increasing levels of cognitive workload. The parallel construction of the experimental protocol allows a direct comparison with other in-vehicle secondary tasks (e.g., listening to a radio, listening to an audio-book, conversing on a cell phone, etc.)

In each of the three experiments described below, we used a combination of performance indices to assess mental workload including reaction time and accuracy in response to a peripheral light detection task (the DRT task: ISO, 2012), subjective workload measures from the NASA Task Load Index (NASA TLX: Hart & Staveland, 1988) and psychophysiological measures associated with either the electro-encephalographic (EEG) or electrocardiographic (ECG) activity of the participant. We also obtained primary-task measures of driving in experiments using the driving simulator and instrumented vehicle.

After describing the methods and results of each study in greater detail, we report both multivariate and meta-analytic analyses that integrate the different dependent measures across the three studies to provide an overall cognitive distraction metric for each of the voice-based interactive technologies in the vehicle. In particular, we used these data to augment the rating system of cognitive distraction developed by Strayer et al., (2013) where

non-distracted single-task driving anchored the low-end (Category 1) and the cognitively demanding OSPAN task anchored the high-end (Category 5) of the scale. The relative ranking compared to these anchors provides an index of the cognitive workload for that activity when concurrently paired with the operation of a motor vehicle.

Experiment 1: Baseline Assessment

Experiment 1 was designed to provide a baseline assessment of the nine tasks previously described. In this controlled assessment, participants were seated in front of a computer monitor that displayed a static fixation cross and they performed the conditions without the added task of driving. The objective was to establish the cognitive workload associated with each activity and to thereby predict the accompanied cognitive distraction from performing that activity while operating a motor vehicle.

Method

Participants: Forty-five participants (27 men and 18 women) from the University of Utah participated in the experiment. Participants ranged in age from 18 to 40 years, with an average age of 24.8 years. All reported having normal neurological functioning, normal or corrected-to-normal visual acuity, normal color vision (Ishihara, 1993), a valid driver's license, and English fluency. Participants' years of driving experience ranged from 2.5 to 24, with an average of 8.5 years. All of the participants owned a cellular phone and 87 percent reported that they used their phone regularly while driving. They were recruited via University-approved flyers posted on campus bulletin boards and via word of mouth within the community. Interested individuals contacted an e-mail address for further information and to schedule an appointment. Eligible participants reported a clean driving history (e.g., no at-fault accidents in the past five years).

Materials: Subjective workload ratings were collected using the NASA TLX survey developed by Hart and Staveland (1988). After completing each of the conditions, participants responded to each of the six items on a 21-point Likert scale ranging from "very low" to "very high." The questions in the NASA TLX were:

a) *How mentally demanding was the task?*
b) *How physically demanding was the task?*
c) *How hurried or rushed was the pace of the task?*
d) *How successful were you in accomplishing what you were asked to do?*
e) *How hard did you have to work to accomplish your level of performance*?
f) *How insecure, discouraged, irritated, stressed, and annoyed were you?*

Equipment: Microsoft PowerPoint 2013 was used to coordinate an interactive messaging service with text to speech features. Participants were given a short list of commands (i.e., *Repeat, Reply, Delete, Next Message,* and *Send*) that were used to control the messaging program. The PowerPoint program was controlled by the experimenter who reacted to the participants' verbal commands, mimicking a speech detection system with perfect fidelity.

Cellular service was provided by Sprint. The cellular phone was manufactured by Apple (Model iPhone 5) running iOS 6 or iOS 7 when the update became available. An Olympus

ME-15 Mono Lapel microphone was clipped to the participant's collar for a voice-controlled Siri messaging system. An iPhone headset microphone adapter was used to allow output from and input to the iPhone 5 when participants used auditory commands to interact with Siri. TEAC CD-X70i Micro Hi-Fi system speakers were used for the presentation of the audio for each of the conditions.

The peripheral detection response task (DRT) hardware and software were designed by Precision Driving Research, Inc. following ISO standards (ISO, 2012). Adopting the protocol used by Strayer et al., (2013), a red/green LED light was mounted on the participant's head via a headband. The light was adjusted to an average 15° to the left and 7.5° above the participant's left eye. Response reaction time was recorded with millisecond accuracy via a microswitch attached to participants' left thumb that was depressed in response to the green light.

Zephyr BioHarness 3 Heart Rate Monitors, which attach around the chest with a flexible strap, were used for 10 of the participants in this experiment. The BioHarness 3 collected measures of heart rate (e.g., Beats per Minute; BPM).

Hosted on a 32-bit research laptop, NeuroScan 4.5 software was used to collect continuous EEG for 10 of the participants in the experiment. The EEG was recorded using a NeuroScan32-electrode NuAmp amplifier. The EEG was filtered online with a DC notch filter (60 Hz) with a sample A/D rate of 250 Hz. The DRT software communicated with the NeuroScan system via a parallel port connection to create event markers associated with the continuous EEG. These event markers allowed for offline stimulus-locked analysis of the EEG recordings (i.e., the DRT stimuli (see below) were used to create time-locked ERPs). The EEG was first visually inspected for artifact and any sections with excessive noise from movement or electronic interference were removed. Next, the influence of blinks on the EEG was corrected using ocular artifact rejection techniques (Semlitsch, Anderer, Schuster, & Presslich, 1986) and the data was epoched 200ms before to 1200ms after the onset of the green target light. These epochs were then filtered with a band pass, zero phase shift filter of 0.01 to 12 Hz. Finally, events that exceeded an artifact rejection criterion of 100 μV were rejected and the remaining events were averaged to obtain one subject's average waveform for each condition in the experiment.

Procedure: Prior to their appointment time, participants were sent a general demographic survey. Upon arrival at the lab, participants read and signed the University of Utah IRB approved consent document.

Table 1. Dependent Measures Obtained in Experiment 1

DRT and NASA TLX measures were collected	DRT, NASA TLX, and Heart Rate measures were collected	DRT, NASA TLX, and EEG measures were
N = 25, 12 female	N = 10, 4 female	N = 10, 2 female

As shown in Table 1, 25 of the participants performed the experiment without any physiological recording equipment attached to their body. This group served as a control to ensure that the collection of physiological data did not alter the observed pattern of

behavioral data.[1] Ten of the participants served in the heart-rate group and they wore a Zephyr BioHarness 3 Heart Rate Monitor. These participants were given instructions on how to attach the device and the experimenter verified correct placement and recording of heart rate data.

Ten of the participants served in the EEG group and performed the study while wearing an EEG cap. The research team placed an EEG cap on the participant and ensured cap fit. Measuring EEG involved using a cap with built-in electrodes configured based upon the International 10–20 system (Jasper, 1958). Dry sponges (QuickCell™ cellulose-based electrodes manufactured by Compumedics) were placed in each electrode location in preparation for cap use. Saline was applied to the sponges so that they expanded to make contact with the surface of the participant's head, with all impedances below 10kΩ. A reference electrode was placed behind the left ear on the mastoid bone and electrode site FP1 served as the ground. Electrooculogram (EOG) electrodes were placed at the lateral canthi of both eyes (horizontal) and above and below the left eye (vertical) to track eye movements and record eye blinks for later data processing. Participants' field of view and normal range of motion were not impeded when wearing the EEG cap.

Participants were asked to complete nine different nine-minute conditions that are described below. These conditions were counterbalanced across participants using a Latin Square design. The participants were seated an average of 65cm from a computer screen displaying a fixation cross. Participants were asked to look forward and avoid making excessive head and eye movements during the completion of each task. Before each condition began, participants were familiarized with the procedures for interacting with the system and they were required to demonstrate proficiency before data collection for that condition commenced.

Table 2. Experimental Conditions Performed in an Order Counterbalanced

1	2	3	4	5	6	7	8	9
Single-task	Car Command	Natural Listen	Synthetic Listen	Natural Listen + Compose	Synthetic Listen + Compose	Menu High Reliability	Menu Low /Moderate Reliability	Siri-based Interactions

As shown in Table 2, the Single Task condition was selected to provide a baseline of cognitive workload (i.e., no concurrent secondary task). In the second condition (Car Command), participants generated verbal commands to alter their vehicle environment. Every 30-45 seconds, a short audio cue was played (e.g., "You are getting hot" or "You want to change the radio station"). The time interval between audio cues was randomized to minimize the predictability of the secondary task stimuli. Participants interpreted the cue and then stated a verbal command in response to the cue (e.g., "Turn AC on low" or "Tune radio to 88.3").

[1] Indeed, a preliminary analysis revealed that the DRT and the NASA TLX measures reported below were identical for the three cohorts, thereby establishing that the collection of physiological data did not alter the pattern of behavioral data.

In the third condition (Natural Listen), participants interacted with a simulated email/text messaging service. The system was fully automated with perfect speech recognition capability implemented using the "Wizard-of-Oz" paradigm (Kelley, 1983; Lee, Caven, Haake, & Brown, 2001; Strayer et al., 2013). Prior to beginning the condition, the participant was familiarized with the program's basic commands, which were: *Repeat*, *Delete*, and *Next Message*. The email and text messages and the system confirmations were pre-recorded using a high-fidelity female voice (author J.T.). Participants were asked to listen to the messages, but they were not allowed to compose or send messages in reply. The messages were designed to be representative of text/email messages that individuals receive on a regular basis from friends, family, coworkers, and service providers (i.e., spam). Message type and duration were equated in the four message system conditions (Conditions 3-6).

In the fourth condition (Synthetic Listen), participants interacted with the same system design as in the third. However, the messages and system confirmations were pre-recorded using a synthetic, computerized female voice, "Kate," from NeoSpeech (NeoSpeech, 2012). NeoSpeech was selected because of its superior synthetic speech generation capabilities. Prior to beginning the condition, the participant was familiarized with the program's basic commands, which were: *Repeat*, *Delete*, and *Next Message*. Participants were asked to listen to the messages, but were not allowed to compose or send messages in reply.

In the fifth condition (Natural Listen + Compose**)**, participants interacted with the same system design as in the second condition. Prior to beginning the condition, the participant was familiarized with the program's basic commands, which were: *Repeat*, *Reply*, *Delete*, *Next Message,* and *Send*. The messages and system confirmations were pre-recorded using the same high-fidelity female voice used in the third condition. Participants were asked to listen and then *compose* a response to messages that required a response.

The sixth condition (Synthetic Listen + Compose), was identical to the fourth except that the messages and system confirmations were pre-recorded using the same synthetic NeoSpeech female voice used in the fourth condition. Prior to beginning the condition, the participant was familiarized with the program's basic commands, which were: *Repeat*, *Reply*, *Delete*, *Next Message,* and *Send*. Participants were asked to listen and then *compose* a response to messages that required a response.

In the seventh condition (Menu System with High Reliability), participants interacted with a simulated infotainment/navigation system. They were instructed to navigate through an auditory menu system to select a grocery store, coffee shop, gas station, bank, or a restaurant location for their GPS system to use once they arrived at an unfamiliar location in the city. For the restaurants, participants were asked to use the system's auditory guides to listen to at least one of the available reviews and make dinner reservations if they liked that restaurant. The system indicated the possible commands to make the next selection. Prior to beginning the condition, participants were familiarized with the program's basic commands. Participants interacted with the program as if it were a fully automated system. As with conditions 3-6, perfect speech recognition capabilities were implemented using the "Wizard-of-Oz" paradigm.

In the eighth condition (Menu System with Moderate Reliability), participants interacted with a system designed in the same manner as that in the seventh condition. However, the

system randomly introduced comprehension and menu navigation errors. System errors occurred on average 7.8 (sd = 2.15) times during the nine-minute condition.

In the ninth condition, participants interacted with Siri via an Apple iPhone 5. To create a hands-free system, a lapel microphone was clipped to the participant's collar and the audio output was played through the external speakers. Participants were asked to interact with Siri to perform three tasks: listening to and sending text messages, updating Facebook or Twitter status, and modifying and reviewing calendar appointments. Participants were instructed to activate this version of "Eyes-Free" Siri by saying, "*Hello/Hi Siri.*" The researcher would then manually activate Siri to allow participants to state their command (i.e., the participant neither looked at nor made physical contact with the iPhone). Before starting the condition, the researcher demonstrated the use of each interaction, and then participants were required to demonstrate proficiency before data collection commenced. Participants were free to alternate between the three tasks in a self-paced order.

In each of the conditions described above, participants also performed the DRT task (ISO, 2012). Following the protocol used by Strayer et al., (2013), the DRT task presented red or green lights every three-five seconds via a head-mounted device. Red lights were presented 80 percent of the time and green lights were presented 20 percent of the time. Both the color of the light and the interval between trials (e.g., 3-5 seconds) was randomized (i.e., this is a 20/80 oddball with stimuli presented in a Bernoulli sequence with an interstimulus interval of 3-5 seconds). Using a go/no-go design, participants were instructed to respond to the green light as quickly as they could by depressing a microswitch that was placed on the participants' left thumb, but to not respond to the red lights. The lights remained illuminated until a response was made or one second had elapsed.

Results

DRT: The DRT data reflect the manual response to the red and green lights in the peripheral detection task. The RT and accuracy data for the DRT task are plotted in Figures 2 and 3, respectively (Appendix). RT for correct responses (i.e., green light responses) was measured to the nearest msec. The accuracy data were converted to the non-parametric measure of sensitivity, A', where a response to a green light was coded as a "hit," non-responses to a red light were coded as a "correct rejection," non-responses to a green light were coded as a "miss," and responses to a red light were coded as a "false alarm" (Pollack & Norman, 1964)[2]. A repeated measures Analysis of Variance (ANOVA) found that RT increased across condition, $F(8, 352) = 26.10$, $p < .01$, partial $\eta^2 = .37$, and that A' decreased across condition, $F(8, 352) = 2.63$, $p < .01$, partial $\eta^2 = .06$.

NASA TLX: The data for the six NASA TLX subjective workload ratings are plotted in Figure 4 (Appendix). In each of the panels, the nine conditions are plotted across the abscissa and the 21-point Likert scale workload rating is represented on the ordinate, ranging from "very low," 1, to "very high," 21. A series of repeated measures ANOVAs found that NASA TLX ratings increased for mental workload, $F(8, 352) = 24.74$, $p < .01$, partial $\eta^2 = .36$; physical workload, $F(8, 352) = 26.29$, $p < .01$, partial $\eta^2 = .13$; temporal demand, $F(8, 352) = 8.66$, $p < .01$, partial η^2

[2] A' measures the average area under the receiver operating characteristic curve (Parasurman & Davies, 1984) and is computed as A' = 1.0-0.25*((p(false alarm)/p(hit)) + (1-p(hit))/(1- p(false alarm))).

= .16; performance, $F(8, 352) = 16.80$, $p < .01$, partial $\eta^2 = .28$; effort, $F(8, 352) = 24.18$, $p < .01$, partial $\eta^2 = .36$; and frustration, $F(8, 352) = 31.51$, $p < .01$, partial $\eta^2 = .42$.

ERPs: Figure 5 (Appendix) presents the grand average ERP waveforms obtained in Experiment 1 at the midline Parietal electrode site (Pz) that were time-locked to the onset of green lights in the DRT task. In the figure, the amplitude in microvolts is cross-plotted with time in msec. A close inspection reveals a well-defined P2-N2-P300 ERP component structure. We focused on the P300 component of the ERP because of its sensitivity to cognitive workload, and we measured both its peak latency and amplitude.

In Figure 6 (Appendix), P300 peak latency, measured as the point in time of maximum positivity in a window between 350 and 800 msec, is plotted for each of the conditions in the experiment. A repeated measures ANOVA found no significant main effect of condition on P300 latency, $F(8, 72) = 1.57$, p = ns, partial $\eta^2 = .15$. The P300 amplitude was quantified by computing the average area under the curve between 350 and 800 msec. Figure 7 (Appendix) plots P300 amplitude as a function of condition. A repeated measures ANOVA found a main effect of condition, $F(8, 72) = 2.39$, $p < .05$, partial $\eta^2 = .21$.

Heart Rate: The Zephyr Bioharness 3 contains a number of internal algorithms that were useful for this research. The internal clock of each heart rate monitor was used to identify the segment of heart data that corresponded to each condition. Once activated, heart rate monitors began automatically collecting data at 1 Hz. To calculate beats per minute (BPM) for each subject and condition, the average BPM were calculated after removing the first and last 30s of each condition's recording interval. Figure 8 (Appendix) plots heart rate in BPM as a function of condition. The effect of condition was not significant, $F(8, 72) = .29$, p = ns, partial $\eta^2 = .03$.

Discussion

Experiment 1 was designed to provide a baseline assessment of several voice-based activities. In this assessment, participants did not drive but were seated in front of a computer monitor that displayed a static fixation cross. Participants were fitted with a head-mounted DRT device and they completed each of the secondary tasks for nine minutes while simultaneously responding to green lights from the DRT device. After completing each of the nine tasks, subjective workload ratings were taken.

Clearly, not all in-vehicle voice-based interactions had the same level of cognitive workload. It is noteworthy that the voice-based interactions that were evaluated were pure measures of cognitive workload in that the tasks did not require the participant to move their hands or divert their eyes from computer screen. The results from the different measures had a good correspondence and help lay the foundation for extending the metric of cognitive workload developed by Strayer et al., (2013). As the cognitive workload associated with performing secondary tasks increases, the cognitive distraction associated with performing that activity while operating a motor vehicle should increase (i.e., driving performance should be adversely affected by in-vehicle cognitive workload).

Experiment 2: Driving Simulator

The goal of Experiment 2 was to extend the findings from Experiment 1 to operating a high-fidelity driving simulator. Given the increase in cognitive workload associated with performing the respective secondary tasks, we expected that measures of driving performance would be adversely affected. The driving simulator used a car following scenario on a multilane highway with moderate traffic. Participants followed a lead vehicle that braked aperiodically throughout the scenario and, in addition to the measures collected in Experiment 1, we also collected brake reaction time and following distance, as these variables associated with the primary task of driving have been shown in earlier research to be sensitive to cognitive distraction (Caird et al., 2008; Horrey & Wickens, 2006).

Method

Participants: Forty-one participants (21 men and 20 women) from the University of Utah participated in the experiment. Participants ranged in age from 18 to 40, with an average age of 25.2 years. All reported normal neurological functioning, normal or corrected-to-normal visual acuity, normal color vision (Ishihara, 1993), a valid driver's license, and English fluency. Participant's years of driving experience ranged from 2.5 to 24, with an average of nine years. All participants owned a cellular phone and 84 percent reported that they used their phone regularly while driving. They were recruited via University-approved flyers posted on campus bulletin boards and via word of mouth within the community. Interested individuals contacted an e-mail address for further information and to schedule an appointment. Eligible participants reported a clean driving history (e.g., no at-fault accidents in the past five years).

Equipment: In addition to the equipment used in Experiment 1, the present study used a fixed-base high fidelity driving simulator (made by L-3 Communications) with high-resolution displays providing a 180-degree field of view. The dashboard instrumentation, steering wheel, gas, and brake pedals were from a Ford Crown Victoria sedan with an automatic transmission. The simulator incorporated vehicle dynamics, traffic-scenario, and road-surface software to provide realistic scenes and traffic conditions. All other equipment was identical to Experiment 1.

Procedure: The procedures used in Experiment 1 were also used in Experiment 2, with the following modifications. In Experiment 2, we used nine counterbalanced simulated car-following scenarios in which participants drove on a multilane freeway with moderate traffic (approximately 1500 vehicles/lane/hour). Participants followed a pace car that would apply its brakes aperiodically. Participants were instructed not to change lanes to pass the pace car, and were asked to maintain a two-second following distance behind the pace car. Participants were given a five-minute practice session to familiarize themselves with the driving simulator. In the practice session, participants were trained to follow a lead vehicle on the highway at a two-second following distance, braking whenever they saw the lead vehicle's brake lights illuminate. If they fell more than 25 meters behind the lead vehicle, a bell sounded, cueing them to shorten their following distance. The bell was not used once the experimental testing commenced.

Table 3. Dependent Measures Obtained in Experiment 2		
Driving Performance, DRT, and NASA TLX measures were collected	Driving Performance, DRT, NASA TLX, and Heart Rate were collected	Driving Performance, DRT, NASA TLX, and EEG were collected
N = 21, 12 female	N = 10, 5 female	N = 10, 3 female

As shown in Table 3, 21 of the participants performed the experiment without any physiological recording equipment attached to their body. Ten of the participants served in the heart-rate group, and they wore a Zephyr BioHarness 3 Heart Rate Monitor. Ten of the participants served in the EEG group and performed the study while wearing an EEG cap.[3]

Results

Driving Performance Measures: Figure 9 (Appendix) presents the *Brake Reaction Time* (RT) measured as the time interval between the onset of the pace car's brake lights and the onset of the participant's braking response (i.e., a 1% depression of the brake pedal). Figure 10 (Appendix) presents the *Following Distance*, measured as the distance between the rear bumper of the pace car and the front bumper of the participant's car at the moment of brake onset. A repeated measures ANOVA found that both RT, $F(8, 320) = 4.26$, $p < .01$, partial $\eta^2 = .10$, and following distance increased across condition, $F(8, 320) = 2.15$, $p < .05$, partial $\eta^2 = .05$. A subsidiary linear mixed model analysis that held following distance constant found that brake RT increased as a function of condition over and above any compensatory effects associated with following distance, $F(8, 5426) = 5.15$, $p < .01$ (see Figure A-3, Appendix). These data establish that performing in-vehicle activities that differ in their attentional requirements have differential effects on driving performance (i.e., the greater the cognitive workload associated with a subsidiary in-vehicle activity, the greater the cognitive distraction).

DRT: The RT and accuracy data for the DRT task are plotted in Figures 2 and 3, respectively. A repeated measures ANOVA found that RT increased across condition, $F(8, 320) = 23.09$, $p < .01$, partial $\eta^2 = .37$, and that A' decreased across condition, $F(8, 320) = 3.58$, $p < .01$, partial $\eta^2 = .08$.

NASA TLX: The data for the six NASA TLX subjective workload ratings are plotted in Figure 4. The subjective workload ratings increased systematically across the conditions. A series of repeated measures ANOVAs found that NASA TLX ratings increased for mental workload, $F(8, 320) = 28.32$, $p < .01$, partial $\eta^2 = .42$; physical workload, $F(8, 320) = 10.88$, $p < .01$, partial $\eta^2 = .21$; temporal demand, $F(8, 320) = 14.42$, $p < .01$, partial $\eta^2 = .27$; performance, $F(8, 320) = 8.20$, $p < .01$, partial $\eta^2 = .17$; effort, $F(8, 320) = 23.99$, $p < .01$, partial $\eta^2 = .38$; and frustration, $F(8, 320) = 40.60$, $p < .01$, partial $\eta^2 = .50$.

[3] A preliminary analysis revealed that the DRT and the NASA TLX measures reported below were identical for the three cohorts, thereby establishing that the collection of physiological data did not alter the pattern of behavioral data.

ERPs: EEG was recorded and analyzed in Experiment 2 using the same protocol as that of Experiment 1. The resulting ERPs are plotted in Figure 11 (Appendix). As with Strayer et al., (2013), the ERPs were degraded as we moved from the laboratory to the driving simulator due to the increased biological noise from eye/head/body movements and electronic noise from the driving simulator. P300 peak latency, measured as the point in time of maximum positivity in a window between 400 and 800 msec, is plotted for each of the conditions in the experiment. Figure 6 presents the P300 latency as a function of condition. A repeated measures ANOVA found no significant main effect of condition of P300 latency $F(8, 72) = 1.49$, p = ns, partial $\eta^2 = .14$. As in Experiment 1, the P300 amplitude, presented in Figure 7, were quantified by computing the average area under the curve between 400 and 800 msec. A repeated measures ANOVA of the P300 area under the curve found no effect of condition, $F(8, 72) = 1.05$, p = ns, partial $\eta^2 = .10$.

Heart Rate: Heart Rate was recorded in Experiment 2 using the same protocol as that of Experiment 1. Figure 8 plots heart rate in BPM as a function of condition. The effect of condition was not significant, $F(8, 72) = .61$, p = ns, partial $\eta^2 = .06$.

Discussion

Experiment 2 replicated and extended the pattern obtained in Experiment 1. Importantly, the increases in cognitive workload resulted in systematic changes in driving performance compared to non-distracted driving. In particular, brake reaction time to imperative events in the driving simulator systematically increased as a function of the cognitive workload associated with performing the different in-vehicle activities. This pattern held even when controlling for the increased following distance drivers adopted in these conditions. The P300 data also replicate our earlier reports of suppressed P300 activity when comparing single-task and hands-free cell phone conditions (Strayer & Drews, 2007).

It is worth considering the pattern of data had participants protected the driving task at the expense of the other in-vehicle activities. In such a case, we would expect that the primary task measures of driving would be insensitive to secondary-task workload. Instead, we show that the mental resources available for driving are inversely related to the cognitive workload of the concurrent secondary task. Thus increasing the cognitive workload of the in-vehicle secondary tasks resulted in systematic increases in cognitive distraction.

Experiment 3

The purpose of Experiment 3 was to establish that the patterns obtained in the laboratory and driving simulator generalize to the operation of an instrumented vehicle on residential roadways. This comparison is important because the consequences of impaired driving in the city are different from that of a driving simulator. Participants drove an instrumented vehicle in a residential section of a city while concurrently performing the nine conditions used in Experiments 1 and 2. If the findings generalize, then there should be a good correspondence between the results of Experiment 3 and those of Experiments 1 and 2.

Method

Participants: Forty participants (23 men and 17 women) from the University of Utah participated in the experiment. Participants ranged in age from 20 to 39, with an average age of 26.1 years. All had normal neurological functioning, normal or corrected-to-normal visual acuity, normal color vision (Ishihara, 1993), a valid driver's license, and English fluency. Participants' years of driving experience ranged from 2 to 24, with an average of 9.9 years. All participants owned a cellular phone and 89 percent reported that they used their phone regularly while driving. They were recruited via University-approved flyers posted on campus bulletin boards and via word of mouth within the community. Interested individuals contacted an e-mail address for further information and to schedule an appointment. The Division of Risk Management Department at the University of Utah ran a Motor Vehicles Record (MVR) report on each prospective participant to ensure participation eligibility based on a clean driving history (e.g., no at-fault accidents in the past five years). In addition, following University policy, each prospective participant was required to complete a University-devised 20-minute online defensive driving course and pass the certification test.

Equipment: In addition to the equipment used in Experiment 1, Experiment 3 used an instrumented 2010 Subaru Outback. The vehicle was augmented with four 1080p Microsoft LifeCam USB cameras that captured the driving environment and participants' facial features. All other equipment was identical to Experiment 1.

Procedure: The procedures used in Experiment 1 were also used in Experiment 3, with the following modifications: prior to their appointment time, participants were sent the University of Utah IRB approved informed consent document, general demographic surveys, and instructions for completing the 20-minute online defensive driving course and the certification test.

Before beginning the study, the driver was familiarized with the controls of the instrumented vehicle, adjusted the mirrors and seat, and was informed of the tasks to be completed while driving. The participant drove around a parking lot in order to become familiar with the handling of the vehicle. Next, participants drove one circuit on a 2.7-mile loop in the Avenues section of Salt Lake City, UT in order to become familiar with the route itself (see Appendix). The route provided a suburban driving environment and contains seven all-way controlled stop signs, one two-way stop sign, and two stoplights. A research assistant and an experimenter accompanied the participant in the vehicle at all times. The research assistant sat in the rear and the experimenter sat in the front passenger seat and had ready access to a redundant braking system and notified the driver of any potential roadway hazards. Participants were familiarized with each condition while stopped on the side of the road.

The driver's task was to follow the route defined above while complying with all local traffic rules, including a 25 mph speed restriction. If drivers exceeded 25 mph, they were reminded of this restriction by the research team. Throughout each condition, the driver completed the DRT. Each condition lasted approximately 10 minutes, which was the average time required to make one loop around the track. Safety directions were reiterated before each driving condition. At the conclusion of the study, participants returned to the Behavioral Sciences building where the participants were compensated for their time and debriefed.

Table 4. Dependent Measures Obtained in Experiment 3	
Driving Performance, DRT, and NASA TLX were collected	Driving Performance, DRT, NASA TLX, and Heart Rate were collected
N = 20, 10 female	N = 20, 7 female

As shown in Table 4, 20 of the participants performed the experiment without any physiological recording equipment attached to their body. Another 20 participants served in the heart-rate group, and they wore a Zephyr BioHarness 3 Heart Rate Monitor.[4]

Results

DRT: The RT and accuracy data for the DRT task are plotted in Figures 2 and 3, respectively. A repeated measures ANOVA found that RT increased across condition, $F(8, 312) = 22.83$, $p < .01$, partial $\eta^2 = .37$, and that A' decreased across condition, $F(8, 312) = 8.00$, $p < .01$, partial $\eta^2 = .17$.

NASA TLX: The data for the six NASA TLX subjective workload ratings are plotted in Figure 4. The subjective workload ratings increased systematically with condition. A series of repeated measures ANOVAs found that NASA TLX ratings increased for mental workload, $F(8, 312) = 34.84$, $p < .01$, partial $\eta^2 = .47$; physical workload, $F(8, 312) = 7.78$, $p < .01$, partial $\eta^2 = .17$; temporal demand, $F(8, 312) = 16.19$, $p < .01$, partial $\eta^2 = .29$; performance, $F(8, 312) = 11.12$, $p < .01$, partial $\eta^2 = .23$; effort, $F(8, 312) = 33.18$, $p < .01$, partial $\eta^2 = .46$; and frustration, $F(8, 312) = 36.39$, $p < .01$, partial $\eta^2 = .48$.

Physiological measures: Heart Rate was recorded in Experiment 3 using the same protocol as that of the prior studies. Figure 8 plots heart rate in BPM as a function of condition. The effect of condition was not significant, $F(8, 152) = .84$, p = ns, partial $\eta^2 = .04$.

Discussion

Experiment 3 replicated and extended the findings from the prior experiments in several important ways. Most importantly, they document that the patterns observed in the controlled laboratory setting of Experiment 1 and in the driving simulator setting of Experiment 2 generalize to what was observed with the instrumented vehicle in a naturalistic setting.

General Discussion

The patterns observed in the three experiments reported in this report are strikingly consistent, establishing that lessons learned in the laboratory and driving simulator are in good agreement with studies of cognitive distraction on the roadway. In each case, they document a systematic increase in cognitive workload as participants performed different in-vehicle activities. The data for the three studies were entered into a MANOVA to determine how cognitive workload changed across condition for the three experiments. For

[4] A preliminary analysis revealed that the DRT and the NASA TLX measures reported below were identical for the two cohorts, thereby establishing that the collection of physiological data did not alter the pattern of behavioral data.

the sake of clarity, we focused our analyses based upon secondary and subjective assessments because these measures were identical across the three experiments. Obviously, there were no primary-task driving measures in Experiment 1 and the measures of brake reaction time and following distance obtained in the simulator were not available in the instrumented vehicle.

A MANOVA performed on the secondary-task DRT data revealed a significant effect of condition, $F(16, 108) = 27.16$, $p < .01$, partial $\eta^2 = .80$, experiment, $F(4, 246) = 35.78$, $p < .01$, partial $\eta^2 = .37$, and a condition X experiment interaction, $F(32, 218) = 2.02$, $p < .01$, partial $\eta^2 = .23$. Further analysis found a main effect of condition such that RT increased, $F(8, 984) = 68.13$, $p < .01$, partial $\eta^2 = .36$, and A' decreased, $F(8, 984) = 12.84$, $p < .01$,partial $\eta^2 = .09$ across condition. In addition, RT increased, $F(2, 123) = 28.08$, $p < .01$, partial $\eta^2 = .31$ and A' decreased, $F(2, 123) = 84.07$, $p < .01$, partial $\eta^2 = .57$, from Experiment 1 to 3. On the whole, there is good agreement across experiments; however, the laboratory- and simulator-based studies would appear to provide a more conservative estimate of the impairments to driving associated with in-vehicle technology use.

A MANOVA performed on the subjective workload ratings revealed a significant effect of condition, $F(48, 76) = 15.10$, $p < .01$, partial $\eta^2 = .90$, of experiment, $F(12, 238) = 1.85$, $p < .05$, partial $\eta^2 = .08$; and a condition X experiment interaction, $F(96, 154) = 1.38$, $p < .05$, partial $\eta^2 = .46$. Across experiments, main effects of condition were obtained for mental workload, $F(8, 984) = 84.32$, $p < .01$, partial $\eta^2 = .41$; physical workload, $F(8, 984) = 23.50$, $p < .01$, partial $\eta^2 = .16$; temporal demand, $F(8, 984) = 35.86$, $p < .01$, partial $\eta^2 = .23$; performance, $F(8, 984) = 33.90$, $p < .01$, partial $\eta^2 = .22$; effort, $F(8, 984) = 78.15$, $p < .01$,partial $\eta^2 = .39$; and frustration, $F(8, 984) = 107.14$, $p < .01$, partial $\eta^2 = .47$. The NASA TLX measures also differed across Experiment 1 to 3 for mental workload, $F(2, 123) = 7.63$, $p < .01$, partial $\eta^2 = .11$; physical workload, $F(2, 123) = 7.06$, $p < .01$,partial $\eta^2 = .10$; temporal demand, $F(2, 123) = 5.54$, $p < .01$, partial $\eta^2 = .08$; effort, $F(2, 123) = 9.40$, $p < .01$,partial $\eta^2 = 13$; and frustration, $F(2, 123) = 4.85$, $p < .01$, partial $\eta^2 = .07$, but not for performance ($p > .08$). On the whole, the subjective workload measures were in agreement across six sub-scales, nine conditions, and three experiments. In particular, there was a consistent increase in subjective workload ratings from conditions 1-9 and also a systematic increase in subjective workload ratings from Experiments 1-3.

In the main, moving from the laboratory to the driving simulator to the instrumented vehicle increased the intercept of the cognitive workload curves, and similar condition effects were obtained for the different dependent measures. This experimental cross-validation establishes that the effects obtained in the simulator generalize to on-road driving. In fact, our measures in Experiment 1 were remarkably consistent with those obtained in Experiment 3, suggesting that there may be occasions where the added complexity, expense, and risk of on-road study are unnecessary. Moreover, the similarity of the primary, secondary, subjective, and physiological measures provides convergence in our workload assessments. It is noteworthy that these tasks allowed drivers to maintain their eyes on the road and their hands on the wheel. That is, these in-vehicle activities are cognitively distracting to different degrees.

One finding that merits further discussion is that Heart Rate did not reach statistical significance in any of the three studies. There are at least three general reasons this may be. The first is that Heart Rate may simply not be sensitive to workload. That is, changes in

cognitive load may not be associated with changes in Heart Rate. This seems unlikely, however, given that a number of recent driving studies have found associations between cognitive load and characteristics of cardiovascular functioning, especially Heart Rate (See Lenneman & Backs, 2009; Mehler, Reimer, & Coughlin, 2012; Reimer et al., 2011). A second possible explanation is that a third variable could have obscured the sensitivity of Heart Rate. This masking variable could have been related to the study design, the data collection protocol, the hardware used for data collection, or the algorithms used to compute heart rate. While all of these possibilities are difficult to completely rule out, the fact that the same basic hardware and data collection protocols have been successfully used in other research (see Cooper, Ingebretsen, and Strayer, 2014) makes the possibility of a masking variable unlikely. A final potential explanation is that the effect of heart rate may be relatively small and the sample size of the current studies may not have been sufficient to detect an effect. Given that experiments 1 and 2 had just 10 subjects each, while experiment 3 had just 20, this explanation seems reasonable. Indeed, the effect size estimates across each of the experiments ranged from just $\eta 2 = .03$ to $\eta 2 = .06$. By way of comparison, using the exact same hardware, the effect size estimate of Heart Rate in the Cooper, Ingebretsen, & Strayer paper (2014) was $\eta 2 = .146$. In order to find a significant effect with these small sample sizes, the effect size of Heart Rate would have needed to be considerably larger. Given all these factors and considerations, the likeliest explanation for not finding an effect of cognitive task demands on Heart Rate is that it was simply not as sensitive as the DRT or subjective measures. By no means does this conclusion rule out the potential utility of Heart Rate as a complementary measure of cognitive load; it does, however, suggest that the expected effect size of Heart Rate is likely relatively small, indicating that relatively robust sample sizes may be needed to effectively utilize the measure.

Toward a Standardized Scale of Cognitive Distraction

The primary goal of the current research was to assess cognitive distraction associated with performing voice-based interactions while operating a motor vehicle. Because the different dependent measures are on different scales (e.g., msec, meters, amplitude, etc.), each was transformed to a standardized score. This involved Z-transforming each of the dependent measures to have a mean of 0 and a standard deviation of 1 (across the experiments and conditions), and the average for each condition was then obtained. The standardized scores for each condition were then summed across the different dependent measures to provide an aggregate measure of cognitive distraction. Finally, the aggregated standardized scores were scaled such that the non-distracted single-task driving condition anchored the low-end (Category 1), and the OSPAN task anchored the high-end (Category 5, see Strayer et al., 2013) of the cognitive distraction scale. For each of the other tasks, the relative position compared to the low and high anchors provided an index of the cognitive workload for that activity when concurrently performed while operating a motor vehicle. The four-step protocol for developing the cognitive distraction scale is listed below.

> Step 1: For each dependent measure, the standardized scores across experiments, conditions, and subjects were computed using $Z_i = (x_i - X) / SD$, where X refers to the overall mean and SD refers to the pooled standard deviation.

Step 2: For each dependent measure, the standardized condition averages were computed by collapsing across experiments and subjects (see Table 5 for the standardized condition averages for each dependent measure).

Step 3: The standardized condition averages across dependent measures were computed with an equal weighting for physical, secondary, subjective, and physiological metrics. Table 5 lists the 14 dependent measures that were used in the standardized condition averages separated in grey by the metric of which they are subordinate. The measures within each metric were also equally weighted. For example, the secondary task workload metric comprised an equal weighting of the measures DRT-RT and DRT-A'. Note that A' and P300 amplitude were inversely coded in the summed condition averages. Figure 12 (Appendix) presents the average effect size of the difference between single-task and each of the remaining conditions using the pooled SD.

Step 4: The standardized mean differences were range-corrected so that the non-distracted single-task condition had a rating of 1.0 and the OSPAN task had a rating of 5.0.[5]

$$X_i = (((X_i - \min) / (\max - \min)) * 4.0) + 1$$

The cognitive distraction scale presented in Figure 13 below ranges from 1.0 for the single-task condition and 5.0 for the OSPAN task (Strayer et al., 2013; see Figure 14 [Appendix] for a side-by-side comparison of the results of Phase I and Phase II). Issuing simple car commands had a rating of 1.88, whereas listening to e-mail/text messages increased the cognitive workload to an average of 2.18. When participants were allowed to compose short messages in response to e-mail/text messages, the workload increased to an average of 3.08. The workload associated with a menu-based navigation (e.g., locate the nearest ATMs) was 2.83 when there was perfect speech translation, and rose to 3.67 when the errors in translation were introduced into the system. Finally, the Siri-based interactions using an eyes-free, hands-free interface had the highest workload ratings observed in Phase II, with a rating of 4.15.

The goal of this research was to be comprehensive, using a variety of driving environments and an inclusive set of dependent measures. Using the standardized values for each dependent measure provided in Table 5 (Appendix), it is possible to use steps 3 and 4 to alter the contribution of dependent measures from the current study. For example, it is straightforward to modify steps 3 and 4 to exclude physiological measures to see their impact on the cognitive distraction scale. Moreover, provided two common anchor points (e.g., single-task driving and OSPAN), other investigators could easily extend the workload scale to an entirely different set of driving conditions, secondary tasks, and dependent measures. However, on a cautionary note, it is inappropriate to post-hoc "cherry-pick" dependent measures to create a desired outcome that is not representative of the overall pattern in the data.

[5] Note that there are two conditions in common between Strayer et al., (2013) and the current research. First, the single-task baseline conditions are identical in the two studies. Second, the speech-to-text condition from Strayer et al., (2013) is identical to the synthetic listen + compose condition. These two anchor points served to calibrate the range-correction algorithm affording ready comparison across the two studies.

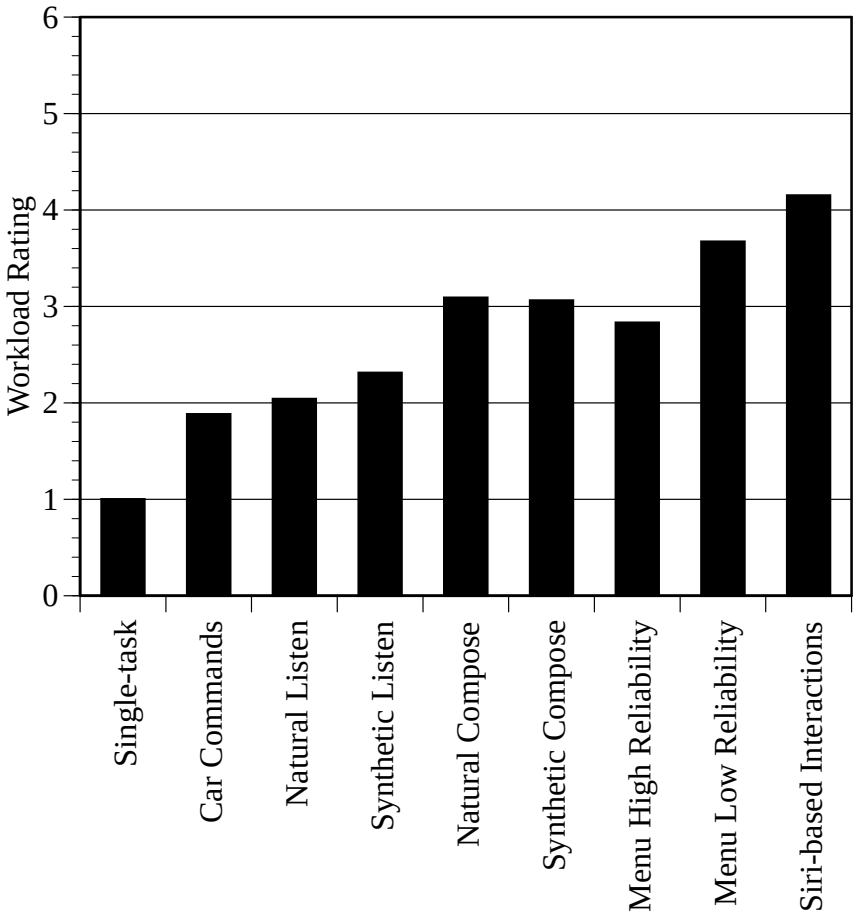


Figure 13. Cognitive workload scale

Takeaways from Phase II

The data from the current research can be used to address four important issues related to voice-based interactions in the vehicle. First, what is the basis of the impairments stemming from the use of speech-based e-mail/texting? This question directly bears on how one might make the voice-based interactions less distracting. For example, does the quality of the speech impact cognitive workload? Is there a difference between just listening to messages compared with listening and replying to the messages? Second, how do simple car commands compare on the workload metric? Given that car commands are short simple utterances, do they compare favorably with an activity like listening to an audio-book or talking on a cell phone? Moreover, do the different interfaces used by different original equipment manufacturers (OEMs) differ, or are they all essentially the same? Third, do menu-based systems that support navigation (e.g., locate the nearest ATM or gas station) incur a workload cost? Given that menu-based systems offer a limited selection of offerings and a restricted set of responses, it is possible that the associated workload may be lower than the speech-to-text e-mail/texting system tested by Strayer et al., (2013). To what extent does the reliability of the menu-based interface affect distraction? Finally, advanced voice recognition systems such as Siri offer the potential for the driver to issue commands and queries using

natural language. Does this sort of voice-based interface reduce the level of cognitive workload compared to the speech-to-text system evaluated by Strayer et al., (2013)?

What makes voice-based interactions distracting?

In our Phase I report (Strayer et al., 2013), we found that interacting with a speech-to-text email/text messaging system with perfect speech recognition capabilities resulted in a driver workload rating of category 3. This finding was replicated in the Synthetic Listen + Compose condition in the current study. Importantly, the experimental protocol herein was structured such that we could decompose the speech-to-text interactions into a 2 X 2 factorial design in which the audio messages were delivered either by a pre-recorded human voice or a computerized synthetic voice (Natural vs. Synthetic), and the participants' interactions with the system involved either listening to e-mail/text messages or listening and composing replies to the messages (using a Wizard-of-Oz perfect speech recognition system).

Inspection of Figures 12 and 13 clearly indicates that there was a large effect of composition that accounted for approximately 45 percent of the increase in workload compared to the single-task baseline, and that the task of selecting and listening to the audio messages contributed approximately 55 percent to the increased cognitive workload. Importantly, there was no systematic difference between the natural and synthetic speech conditions. This conclusion was verified with a subsidiary 2 (Speech Quality: Natural vs. Synthetic) X 2 (Voice Interaction: Listen vs. Compose) MANOVA using DRT and NASA TLX measures that were in common for all participants in the four conditions. The MANOVA revealed a significant effect of Voice Interaction, $F(8, 116) = 19.5$, $p < .01$, partial $\eta^2 = .57$, but neither the effect of Speech Quality nor any of the higher-order interactions were significant. This latter finding suggests that there is little to be gained by improving the quality of the synthetic speech, at least with regard to the driver's cognitive workload. It is noteworthy, however, that prior research (Harbluk 2005; Jamson et al., 2004; Lee et al., 2001; Raney, Harbluk, & Noy, 2005) found that the quality of the synthetic speech was associated with increased mental workload. The difference between the current research and the prior research likely reflects the improvements in the quality of computerized speech technology over the last decade. Notice also that just listening to messages without the possibility of generating a reply was associated with an average cognitive workload rating of 2.17, a level that is comparable to the workload associated with conversing on a cell phone (Strayer et al., 2013).

How distracting are simple car commands?

Our research also examined the impact of simple car commands on cognitive workload. In this case, participants were requested to issue a voice-based command to change the infotainment system (e.g., change radio station) or adjust the climate control (e.g., raise or lower the HVAC). These commands were short, simple commands and we used Wizard-of-Oz technology so that the commands were received with perfect fidelity. Car commands were issued once every 30-45 seconds with the remaining time similar to the single-task baseline. With these simple car commands, the cognitive workload associated with this interaction was 1.88, ranking close to listening to an audio-book (Strayer et al., 2013).

As part of our ongoing research, we also conducted a companion research project, described in detail in Cooper & Strayer (in press), that evaluated the cognitive demands of simple

auditory/vocal vehicle interactions using five 2013 and one 2012 model year OEM voice-based systems. In this investigation, 36 participants completed a series of voice-based radio tuning and phone dialing tasks while driving on a variant of the course used in Experiment 3. Each of the participants drove the six vehicles on the nine-minute loop, and they were periodically instructed to dial a 10-digit number, call a contact from the contact list, change the radio station, or play a song from a pre-inserted CD. All of the interactions took place using a bluetooth hands-free voice system that was activated with the touch of a button on the steering wheel. The OEM systems evaluated in this research were: a Ford equipped with MyFord Touch, a Chevrolet equipped with MyLink, a Chrysler equipped with Uconnect, a Toyota equipped with Entune, a Mercedes equipped with COMAND, and a Hyundai equipped with Blue Link. For comparison purposes, mental workload was also assessed during single-task and OSPAN baseline drives.

Across these eight conditions (6 OEM systems, single-task, and OSPAN conditions), measures of cognitive workload were derived from reaction time, psychophysiological, and subjective workload metrics. Reaction time and accuracy measures were obtained using the DRT based head-mounted device used in Experiments 1-3. Heart rate was collected from all participants using the same procedure as Experiments 1-3. Finally, subjective workload measures were obtained using the NASA TLX. The resulting workload ratings are presented in Figure 15 (Appendix), alongside the ratings from Strayer et al., 2013 and the workload ratings obtained from the current study.

Inspection of Figure 15 indicates that that there are striking differences in the cognitive demand incurred through voice interactions with different OEM voice-based systems. In the best case, we found that radio tuning and voice/contact dialing using the Toyota's Entune imposed a modest level of cognitive workload (1.70), a level comparable to that obtained with the audio-book condition from Strayer et al., (2013) and from the car command condition in the current experiment. In the worst case, those same activities using Chevy's MyLink, imposed high levels of cognitive workload (3.70), a level only surpassed by the Siri-based interactions and the OSPAN task. It is important to note that these systems were evaluated in a counterbalanced order using the same set of secondary tasks under the same driving conditions given the same level of practice with each of the infotainment systems. Moreover, there were no differences in workload between the six vehicles in single-task driving conditions, indicating that the different ratings can be attributed directly to the cognitive interactions associated with the different OEM voice-based systems and not to differences in the workload associated with driving the different vehicles. Perhaps not surprisingly, one of the most critical elements of workload appeared to be the duration of the interaction. This element was driven by the verbosity of the system, the number of steps required to execute an action, and the number of comprehension errors that arose. For the secondary infotainment tasks selected for this analysis, Toyota's Entune system required the least amount of time-on-task while Chevrolet's MyLink required the most.

How distracting are structured menu-based interactions?

Menu-based systems offer the possibility of structuring the number of items in the list, reminding the driver of the alternatives, and restricting the set of potential responses (e.g., select from the following four options). As such, the working memory burden should be

reduced and the potential speech options limited. Even so, the high-reliability menu system evaluated in the current set of studies was associated with a level of cognitive workload similar to that obtained with the speech-to-text system from Strayer et al., (2013) and the current Synthetic/Natural compose + listen conditions. Recall that the high-reliability menu-based interactions had no translation errors and therefore represent a best-case scenario. In the low-reliability system, with random errors introduced into the translation, the workload rose to 3.67, a level substantially higher than what was observed in the speech-to-text condition of Strayer et al., 2013. When even the best-case scenario is associated with a relatively high level of workload, it suggests that the menu-based approach should be used with caution. For example, based on the limits of working memory capacity, the number of items in any given menu shouldn't exceed four or five, and great care should be given to considerations of the usability of the system and the reliability of speech recognition, as workload increased systematically with declines in subjective ratings of usability.

How distracting are natural-language interfaces?

Siri-based interactions involved using natural language to send and receive text messages, update Facebook or Twitter, and modify and review calendar appointments. To create a completely hands-free version, a lapel microphone was clipped to the participant's collar and they activated Siri with the command "*Hello Siri,*" at which point a researcher manually activate the device. The participant neither looked at nor made physical contact with the iPhone during these interactions. Even so, the workload ratings exceeded category 4 on our workload scale – the highest ratings that we have observed for any task short of OSPAN. Moreover, there were two crashes in the simulator study when participants used Siri (the only other crash we observed was when participants used the menu-based systems).

To understand the workload rating associated with interacting with Siri, it is first useful to consider what is *not* causing the effect. The high level of workload is not due to visual/manual interference. Participants never looked at nor touched the iPhone during the session; in fact, the experimenter performed all manual interaction with Siri. As such, this indicates that the impairments were cognitive in nature, associated with the allocation of attention to the task. The high level of workload is also not due to the quality of vocal input or audio output. Participants wore a lapel microphone that allowed them to speak in a normal voice and the audio was played clearly over stereo speakers in the lab, simulator, or car. Our current research also indicates that the quality of the synthetic speech was not a major contributor to the effect. As such, this suggests that the impairment was not attributable to input/output operations. However, as depicted in Figures A1 and A2 (Appendix), Siri had the lowest rating of intuitiveness and the highest rating of complexity of any of the conditions we tested.

With regard to Siri, it is also useful to contrast it with the "best case" natural listen + compose condition, which was rated at 3.08, and used Wizard-of-Oz technology to achieve perfect speech recognition. Siri scored more than a full point higher on the workload rating scale (4.15), and this likely reflects the added complexity when the voice-recognition system is less than perfect. Siri can learn about accents and other the characteristics of the user's voice, so it is possible that with extended practice the workload ratings might improve. Common issues involved inconsistencies in which Siri would produce different responses to seemingly identical commands. In other circumstances, Siri required exact phrases to

accomplish specific tasks, and subtle deviations from that phrasing would result in a failure. When there was a failure to properly dictate a message, it required starting over since there was no way to modify/edit a message or command. Siri also made mistakes such as calling someone other than the desired person from the phone contact list. Some participants also reported frustration with Siri's occasional sarcasm and wit.

These and other idiosyncrasies resulted in an overly complex interaction, and it is possible that improvements to the software design will address some of these issues. There are other voice-recognition systems (e.g., Google Now, Microsoft Cortana, etc.) that were not tested in the current evaluation. It is possible that these systems differ in cognitive workload, resulting in variability much the same as what we observed with in-vehicle car commands, but additional research will be required to verify this. Even so, it is unlikely that the ratings of these voice-recognition systems would drop below 3, the level we obtained with a perfect speech recognition system.

Caveats and Limitations

The cognitive distraction scale provides a comprehensive analysis of several of the cognitive sources of driver distraction. The scale does not directly measure visual/manual sources of distraction, although changes in visual scanning associated with cognitive workload are included in the metric. Moreover, there is not a comprehensive mapping of cognitive distraction to on-road crash risk. However, it is reasonable to assume that there would be a monotonic relationship between cognitive distraction and crash risk.

Our research examined participants between the ages 18 to 40 with a mean age of 25.3 and 9.2 years of driving experience. It is likely that older drivers may experience greater levels of workload because they find the task of driving more attention demanding due to capacity and processing speed declines with senescence (Salthouse 1996; Strayer & Drews, 2004; West, 1996). Consequently, it is likely that the workload estimates provide a conservative estimate of the workload experienced by older drivers when they interact with the same in-vehicle systems.

Summary and Conclusions

Measuring cognitive distraction has proven to be the most difficult of the three sources of distraction to assess because of the problems associated with observing what a driver's brain (as opposed to hands or eyes) is doing. The current research used a combination of primary, secondary, subjective, and physiological measures to assess cognitive distraction across a variety of voice-based in-vehicle activities. We established that there are significant impairments to driving that stem from the diversion of attention from the task of operating a motor vehicle. The data suggest that voice-based interactions in the vehicle may have unintended consequences that adversely affect traffic safety.

References

Angell, L., Auflick, J., Austria, P. A., Kochhar, D., Tijerina, L., Biever, W., Diptiman, T., Hogsett, J., & Kiger, S. (2006). *Driver workload metrics task 2 final report.* Washington, DC: DOT HS 810 635.

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

Cooper, J. M., Vladisavljevic, I., Medeiros-Ward, N., Martin, P. T., & Strayer, D. L. (2009). Near the tipping point of traffic stability: An investigation of driving while conversing on a cell phone in simulated highway traffic of varying densities. *Human Factors, 51*, 261- 268.

Cooper, J.M., Ingebretsen, H., & Strayer, D.L. (2014). Mental workload of common voice-based vehicle interactions. *AAA Foundation for Traffic Safety.*

Dingus, T. A., Klauer, S. G., Neale, V. L., Petersen, A., Lee, S. E., Sudweeks, J., Knipling, R. R. (2006). *The 100-car naturalistic driving study: phase II -- Results of the 100-car field experiment.* Washington, DC: DOT HS 810 593.

Drews, F. A., Pasupathi, M., & Strayer, D. L. (2008). Passenger and cell-phone conversation during simulated driving. *Journal of Experimental Psychology: Applied, 14*, 392-400.

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

Fisher, D. L., & Strayer, D. L. (2014). Modeling situation awareness and crash risk, *Annals of Advances in Automotive Medicine, 5, 33-39.*

Gugerty, L. (1997). Situation awareness during driving: explicit and implicit knowledge in dynamic spatial memory. *Journal of Experimental Psychology: Applied, 3*, 42-66.

Harbluk, J. L., Noy, Y. I., Trbovich, P. L., & Eizenman, M. (2007). An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance. *Accident Analysis & Prevention*, 372-379.

Harbluk, L., & Noy, I. (2002). *The impact of cognitive distraction on driver visual behaviour and vehicle control.* Canada: Ergonomics Division, Road Safety Directorate and Motor Vehicle Regulation Directorate.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock, & N. Meshkati, *Human Mental Workload.* Amsterdam: North Holland Press.

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

Ishihara, S. (1993). *Ishihara's test for color-blindness.* Tokyo: Kanehara.

ISO. (2012). Road vehicles -- Transport information and control systems -- Detection-Response Task (DRT) for assessing selective attention in driving. ISO TC 22 SC 13 N17488 (Working Draft). *Under development by Working Group 8 of ISO TC22, SC 13.*

Jasper, H. A. (1958). The ten-twenty system of the International Federation. *Electroencephalography and Clinical Neurophysiology, 10*, 371-375.

Kass, S. J., Cole, K. S., & Stanny, C. J. (2007). Effects of distraction and experience on situation awareness and simulated driving. *Transportation Research Part F, 10*, 321-329.

Kelley, J. F. (1983). An empirical methodology for writing user-friendly natural language computer applications. *Proceedings of ACM SIG-CHI '83 Human Factors in Computing Systems* (pp. 193-196). Boston: New York, ACM.

Kramer A. F., Sirevaag E. J., Braun R. *(*1987*).* A psychophysiological assessment of operator workload during simulated flight missions. *Human Factors, 29*, 145–160.
Lee, J. D. (2004). Simulator Fidelity: How low can you go? *48th Annual Meeting of the Human Factors and Ergonomics Society.* Santa Monica, CA.
Lee, J. D., Caven, B., Haake, S., & Brown, T. L. (2001). Speech-based interactions with in-vehicle computers; The effect of speech-based e-mail on drivers' attention and roadway. *Human Factors, 43*, 631-640.
Lenneman, J. K., & Backs, R. W. (2009). Cardiac autonomic control during simulated driving with a concurrent verbal working memory task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*.
McCarley, J. S., Vais, M., Pringle, H., Kramer, A. F., Irwin, D. E., & Strayer, D. L. (2004). Conversation disrupts scanning and change detection in complex visual scenes. *Human Factors, 46*, 424-436.
McCarthy, G., & Donchin, E. (1981). A metric for thought: A comparison of P300 latency and reaction time. *Science*, *211*, 77-80.
Mehler, B., Reimer, B. & Coughlin, J.F. (2012). Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: an on-road study across three age groups. *Human Factors, 54,* 396-412.
NeoSpeech (2012). NeoSpeech Text-to-Speech voices [computer software]. Santa Clara, CA. Available from http://www.neospeech.com/.
NHTSA. (2012). Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehcile Electronic Devices. Department of Transportation. Docket No. NHTSA-2010-0053.
Parasuram, R., & Davies, D. R. (1984). Varieties of Attention. Academic Press.
Pollack, I.,& Norman, D. A. (1964). Non-parametric analysis of recognition experiments. Psychonomic Science,1 ,125-126.
Ranney, T., Mazzae, E., Garrott, R., & Goodman, M. (2000). NHTSA Driver Distraction Research: Past, Present and Future [online].
Recarte, M. A., & Nunes, L. M. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Psychology: Applied, 6*, 31-43.
Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of Experimental Psychology: Applied*, 119-137.
Regan, M. A., Hallett, C. & Gordon, C. P. (2011). Driver distraction and driver inattention: Definition, relationship and taxonomy. *Accident Analysis and Prevention*, Vol. 43, pp1771-1781.
Regan, M. A., & Strayer, D. L. (2014). Towards an understanding of driver inattention: taxonomy and theory, *Annals of Advances in Automotive Medicine, 58*, 5-13.
Reimer, B. (2009). Impact of cognitive task complexity on drivers' visual tunneling. *Transportation Research Record: Journal of the Transportation Research Board, 2138*, 13-19.
Reimer, B., Mehler, B., Coughlin, J. F., Roy, N., & Dusek, J. A. (2011). The impact of a naturalistic hands-free cellular phone task on heart rate and simulated driving performance in two age groups. *Transportation research part F: traffic psychology and behaviour*, *14*(1), 13-25
Reimer, B., Mehler, B., Wang, Y., & Coughlin, J.F. (2012). A field study on the impact of variations in short term memory demands on drivers' visual attention and driving performance across three age groups. *Human Factors, 54,* 454-468.

Reimer, B., Mehler, B. L., Pohlmeyer, A. E., & Coughlin, J. F. (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.

Sanbonmatsu, D. M., Strayer, D. L., Medeiros-Ward, N., & Watson, J. M. (2013). Who multi-tasks and why? Multi-tasking ability, perceived multi-tasking ability, impulsivity, and sensation seeking. *PLOS ONE, 8*, 1-8.

Semlitsch, H. V., Anderer, P., Schuster, P., & Presslich, O. (1986). A solution for reliable and valid reduction of ocular artifacts, applied to the P300 ERP. *Psychophysiology*, *23*, 695-703.

Sirevaag, E. J., Kramer, A. F., Coles, M. G., & Donchin, E. (1989). Resource reciprocity: An event-related brain potentials analysis. *Acta Psychologica*, *70*, 77-97.

Sirevaag, E. J., Kramer, A. F., Wickens, C. D., Reisweber, M., Strayer, D. L., & Grenell, J. H. (1993). Assessment of pilot performance and mental workload in rotary wing aircraft. *Ergonomics, 9*, 1121-1140.

Sodhi, M., & Reimer, B. (2002). Glance analysis of driver eye movements to evaluate distraction. *Behavior Research Methods, Instruments, & Computers, 34*, 529-538.

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

Strayer, D. L., & Drews, F. A. (2007). Cell-phone-induced driver distraction. *Current Directions in Psychological Science, 16*, 128-131.

Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone induced failures of visual attention during simulated driving. *Journal of Experimental Psychology: Applied, 9*, 23-52.

Sussman, E. D., Bishop, H., Madnick, B., & Walker, R. (1985). Driver inattention and highway safety. *Transportation Research Record, 1047*, 40-48.

Taylor, T., Pradhan, A. K., Divekar, G., Romoser, M., Muttart, J., Gomes, R., Pollatsek, A., & Fisher, D. L. (2013). The view from the road; the contribution of on-road glance-monitoring technologies to understanding driver behavior. *Accident Analysis and Prevention*.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory & Language, 28*, 127-154.

Tsai, Y., Viirre, E., Strychacz, C., Chase, B. & Jung, T. (2007). Task performance and eye activity: Predicting behavior relating to cognitive workload. *Aviation, Space, and Environmental Medicine, 5,* B176-b185.

Victor, T. W., Harbluk, J. L., & Engström, J. A. (2005). Sensitivity of eye-movement measures to in-vehicle task difficulty. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*, 167-190.

Wang, J. -S., Knipling, R. R., & Goodman, M. J. (1996). The role of driver inattention in crashes: New statistics from the 1995 Crashworthiness Data System. *40th Annual Proceedings of the Association for the Advancement of Automotive Medicine*, (pp. 377-392). Vancouver Canada.

Wickens, C., Kramer, A., Vanasse, L., & Donchin, E. (1983). Performance of concurrent tasks: A psychophysiological analysis of the reciprocity of information-processing resources. *Science*, *221*, 1080-1082.

Appendix

Table 5. Standardized scores for each dependent measure. Note that the primary task measures of Brake RT and following distance (FD) were collected in Experiment 2, Glances at hazards were collected in Experiment 3, the secondary-task measures of DRT-RT and A' were collected in Experiments 1-3, the NASA TLX subjective workload measures of mental workload, physical workload, temporal demand, performance, effort, and frustration were collected in Experiments 1-3, and the physiological measures of P3 Latency (P3 Lat.) was collected in Experiments 1 and 2 and P3 Area measures were obtained in Experiments 1-2. Heart rate measures were collected in Experiments 1-3.

	Single-task	Car Command	Natural Listen	Synthetic Listen	Natural Listen + Compose	Synthetic Listen + Compose	Menu High Reliability	Menu Low Reliability	Siri-based Interactions
Brake RT	-.383	-.227	.043	-.118	-.048	.140	.220	.241	.132
FD	-.248	-.049	.017	-.060	-.028	.097	.135	.178	-.042
DRT-RT	-.660	-.214	-.117	-.168	.226	.141	.066	.175	.542
DRT-A'	.300	.119	.093	.050	-.077	-.064	-.020	-.076	-.325
Mental	-.869	-.444	-.395	-.172	.403	.362	.015	.367	.733
Physical	-.415	-.237	-.224	-.033	.108	.125	.027	.258	.390
Temporal	-.570	-.479	-.192	-.120	.259	.234	-.018	.325	.561
Performance	-.330	-.262	-.259	-.223	.012	.022	-.183	.528	.696
Effort	-.682	-.423	-.402	-.284	.236	.155	-.060	.551	.908
Frustration	-.673	-.392	-.407	-.276	.050	-.044	-.219	.858	1.100
P3 Lat.	-.376	.085	-.518	.056	.217	.061	.083	.267	.126
P3 Area	.390	.321	.142	-.046	-.056	.005	-.043	-.225	-.489
Heart Rate	.008	-.042	-.036	-.035	.090	-.001	-.086	-.010	.112

Figures Referenced in Text

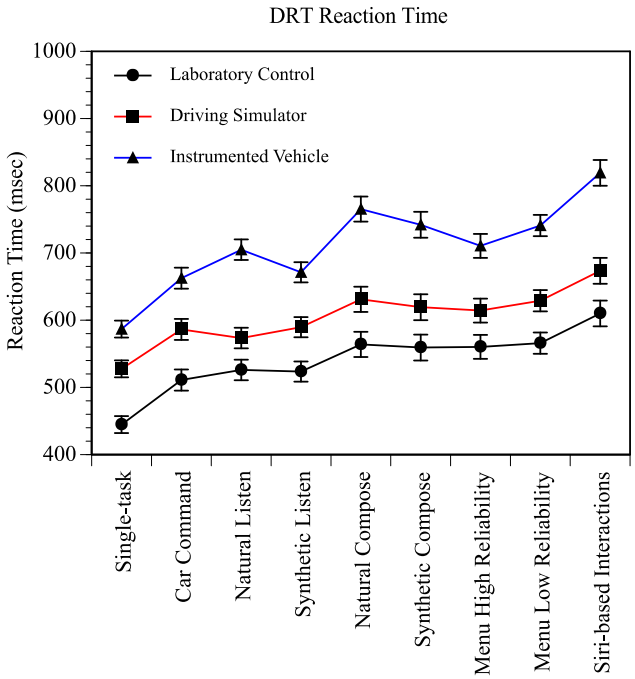


Figure 2. DRT RT (Experiments 1-3)

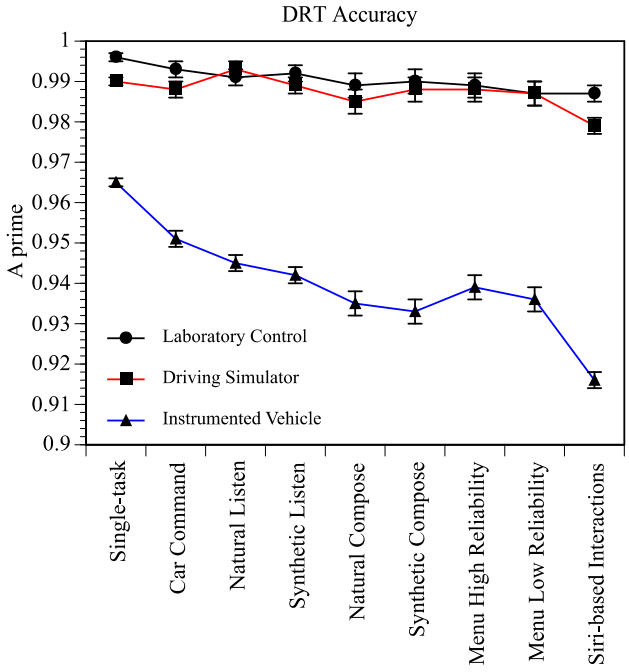


Figure 3. DRT A' (Experiments 1-3)

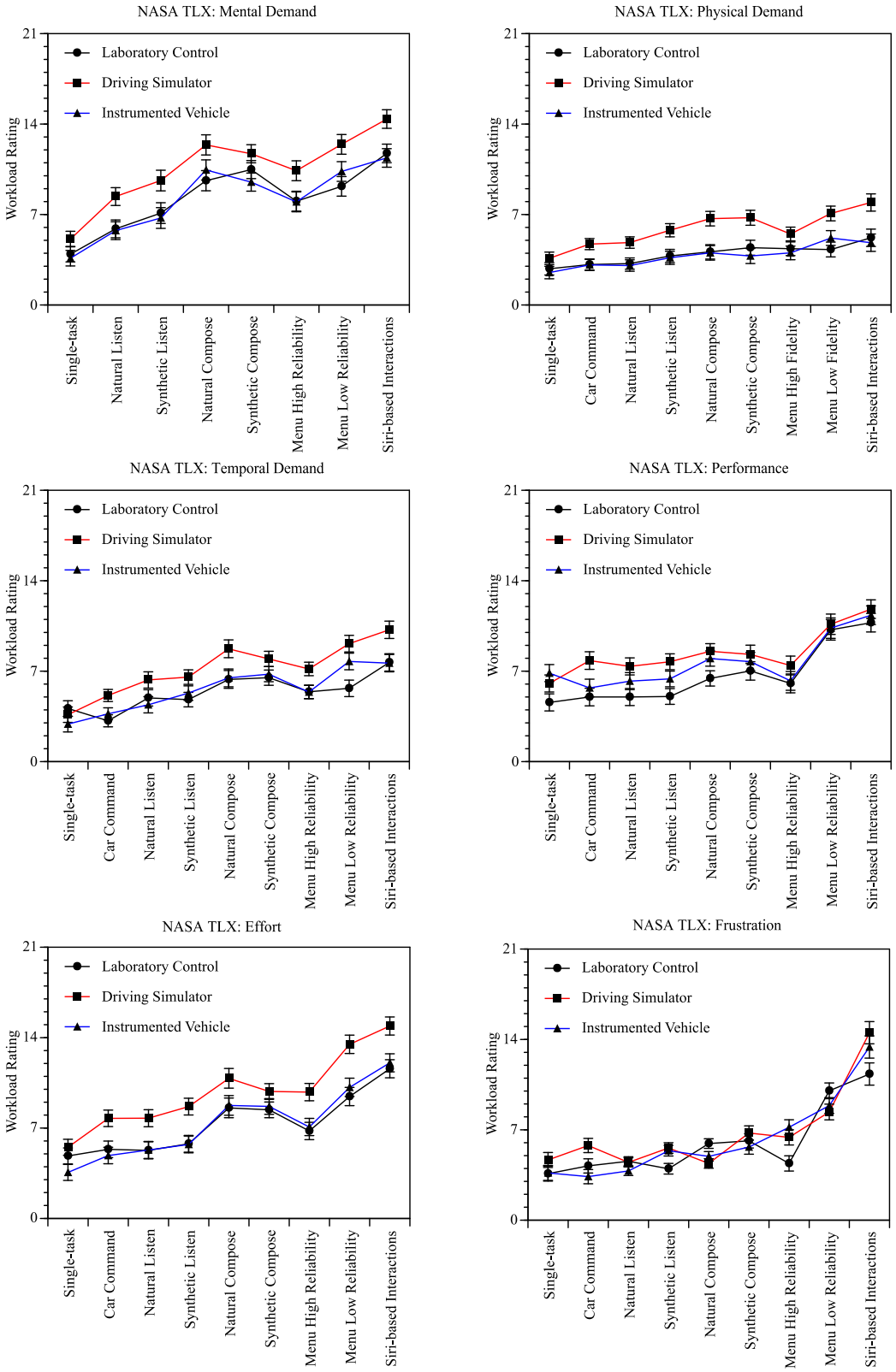


Figure 4. NASA-TLX (Experiments 1-3)

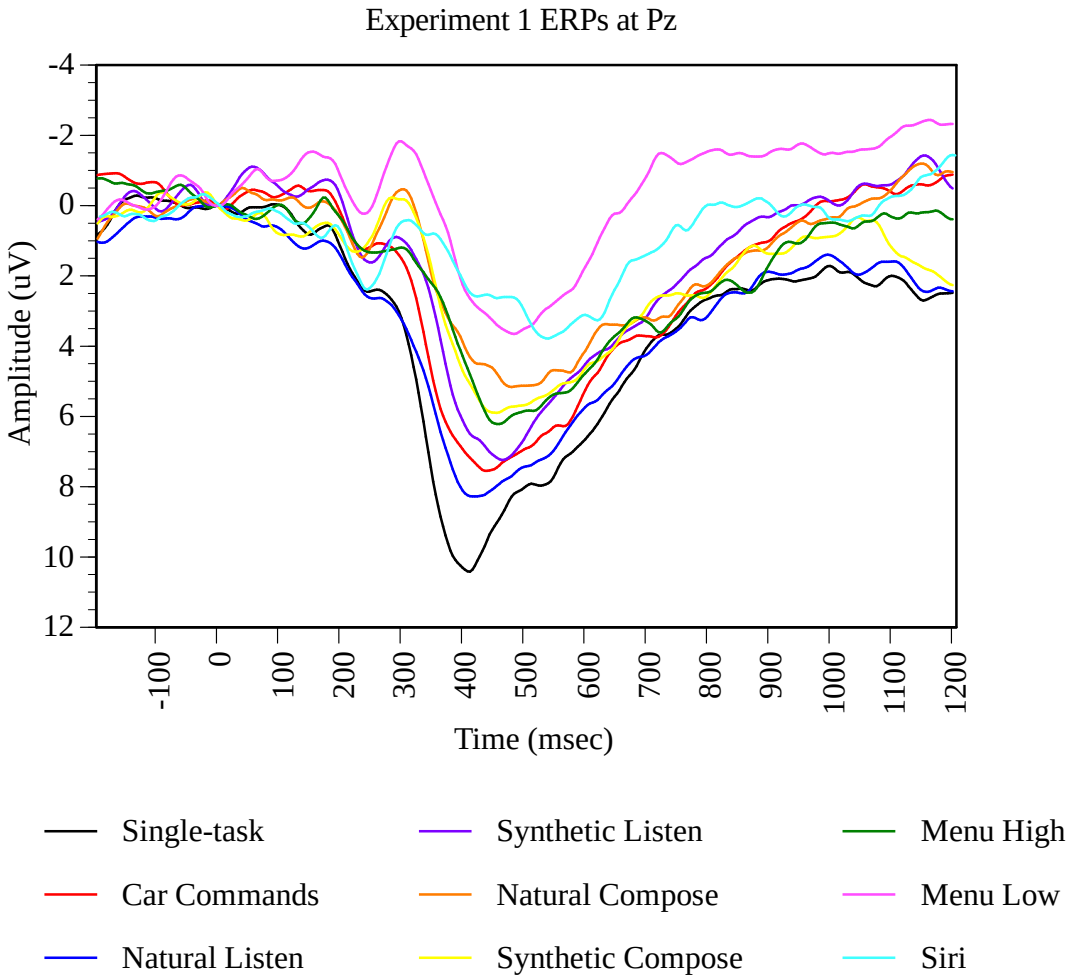


Figure 5. ERPs (Experiment 1)

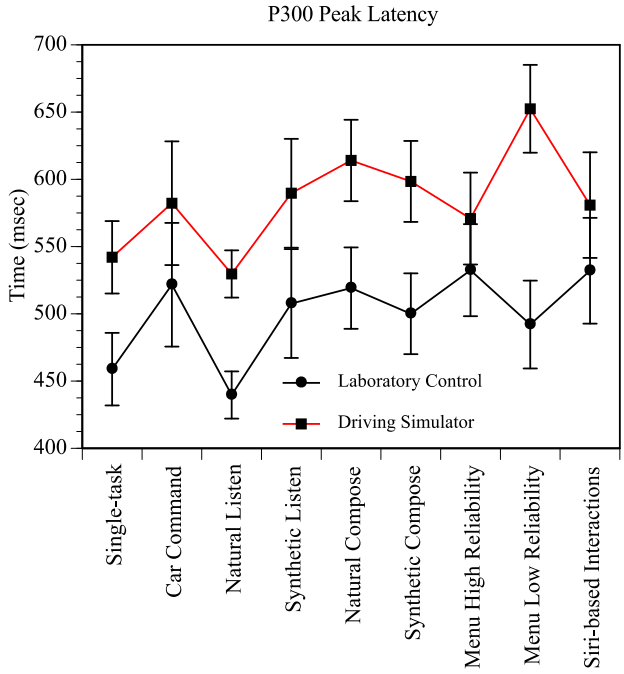


Figure 6. P3 Latency (Experiments 1-2)

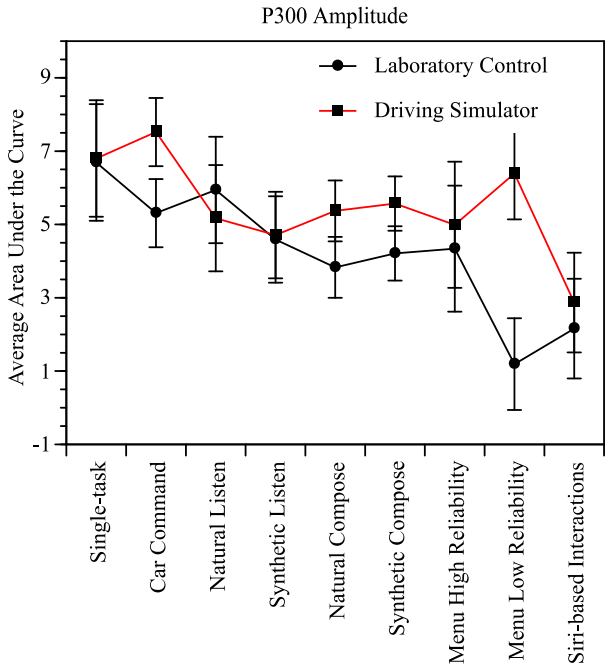


Figure 7. P3 Amplitude (Experiments 1-2)

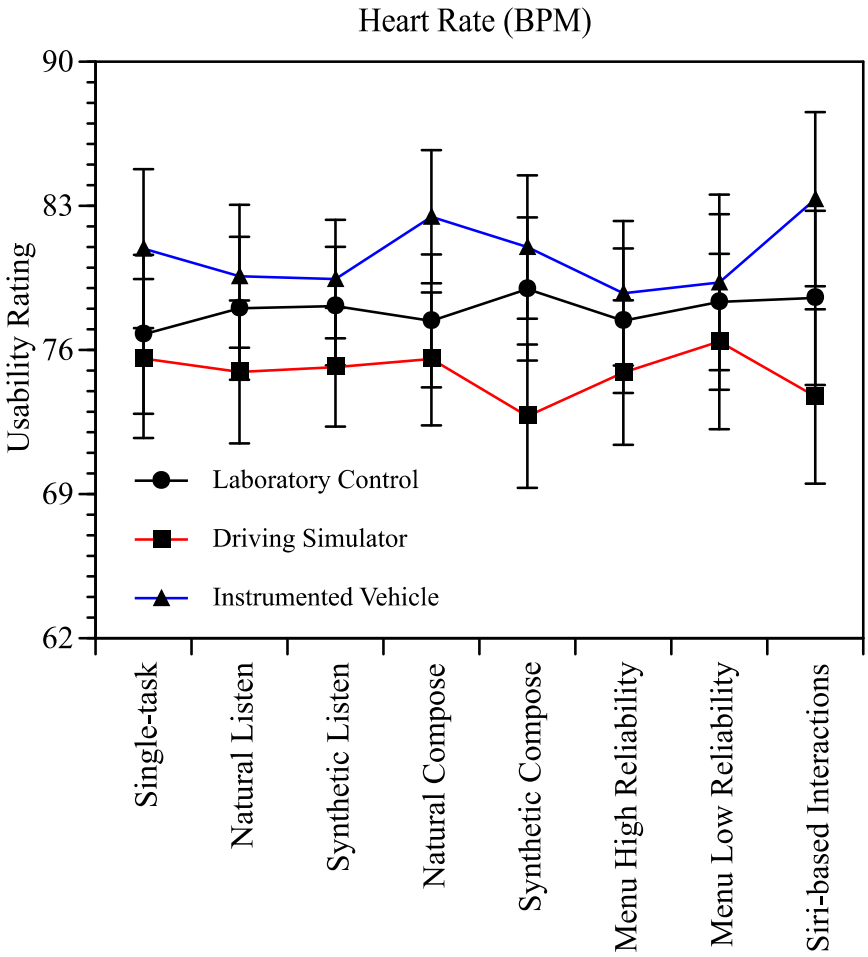


Figure 8. Heart Rate – Beats per Minute (Experiments 1-3)

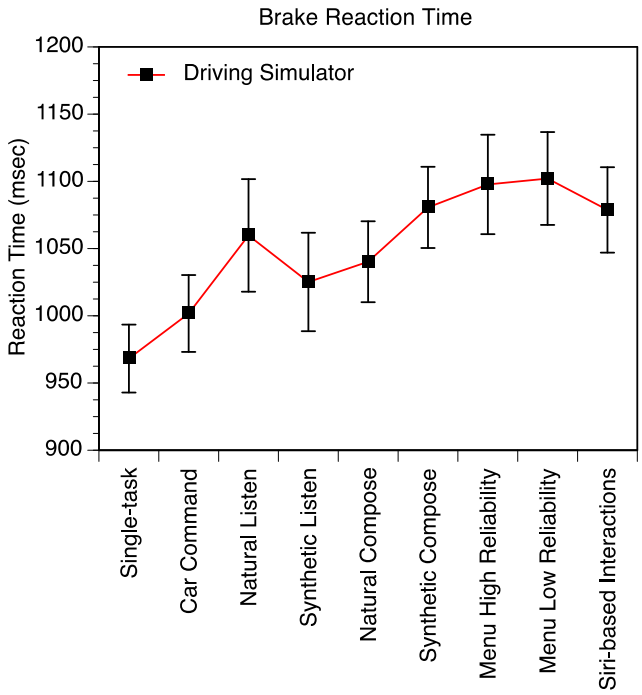


Figure 9. Brake RT (Experiment 2)

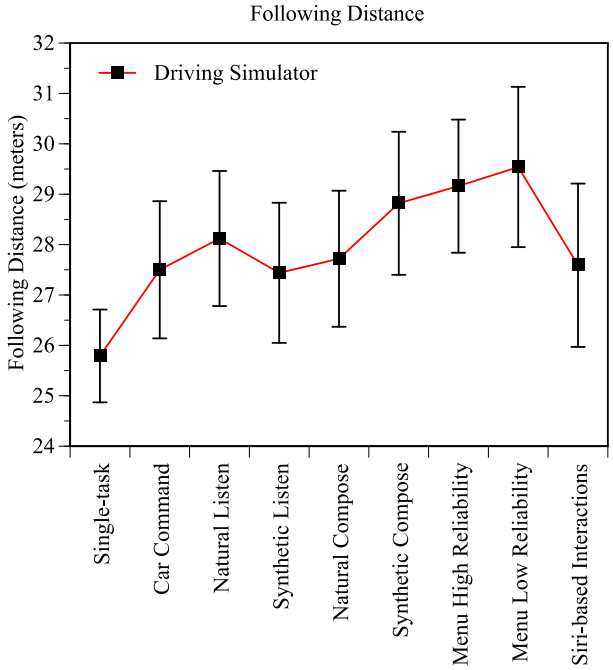


Figure 10. Following Distance (Experiemnt 2)

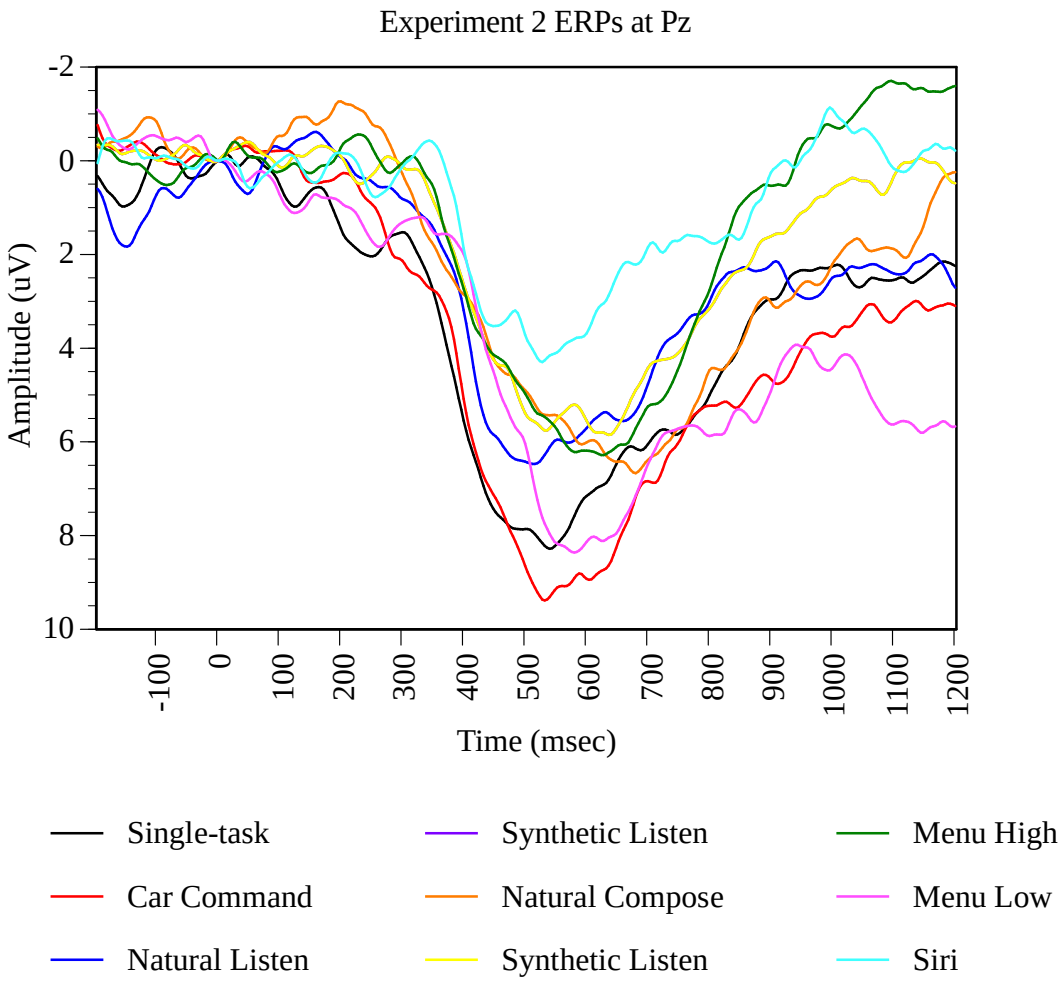


Figure 11. ERPs (Experiment 2)

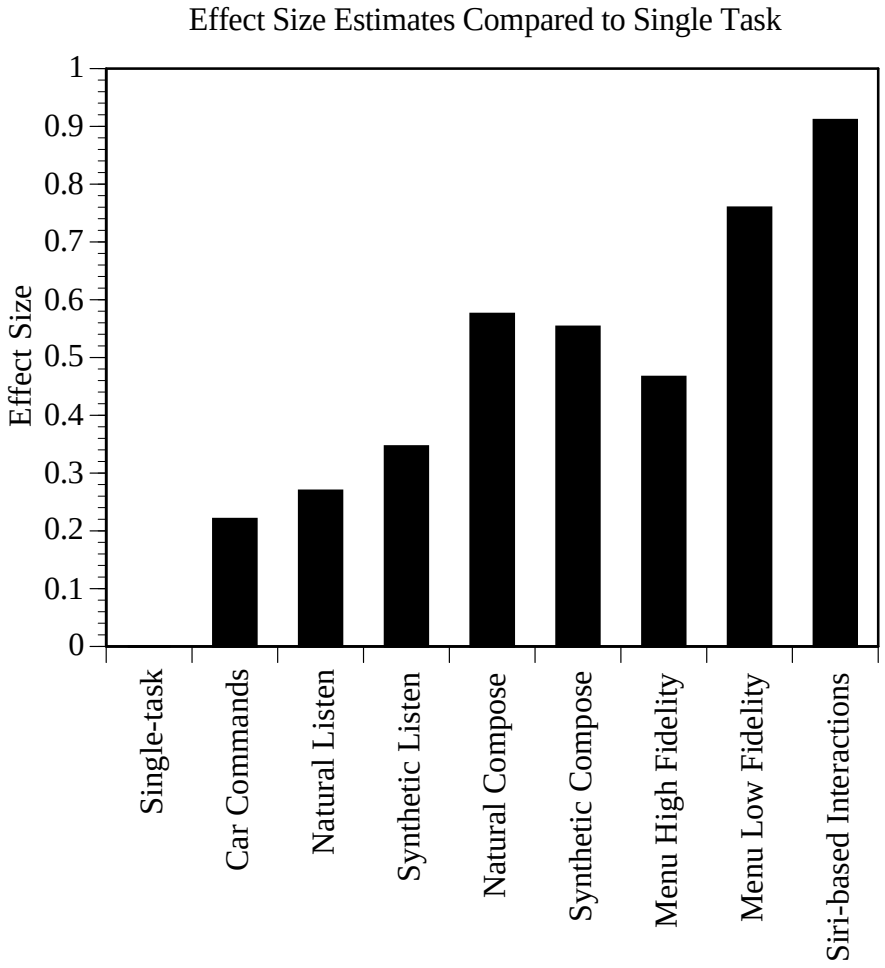


Figure 12. Effect size estimates compared to single-task

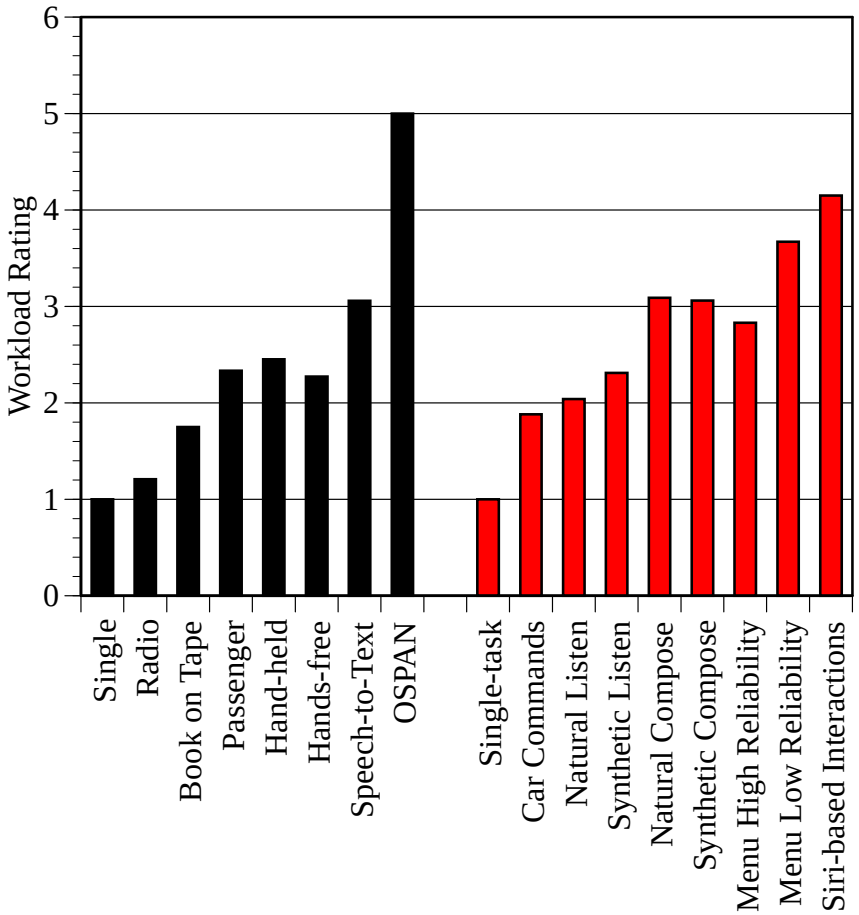


Figure 14. Workload scale for Strayer et al., (2013) (black bars) and the current research (red bars)

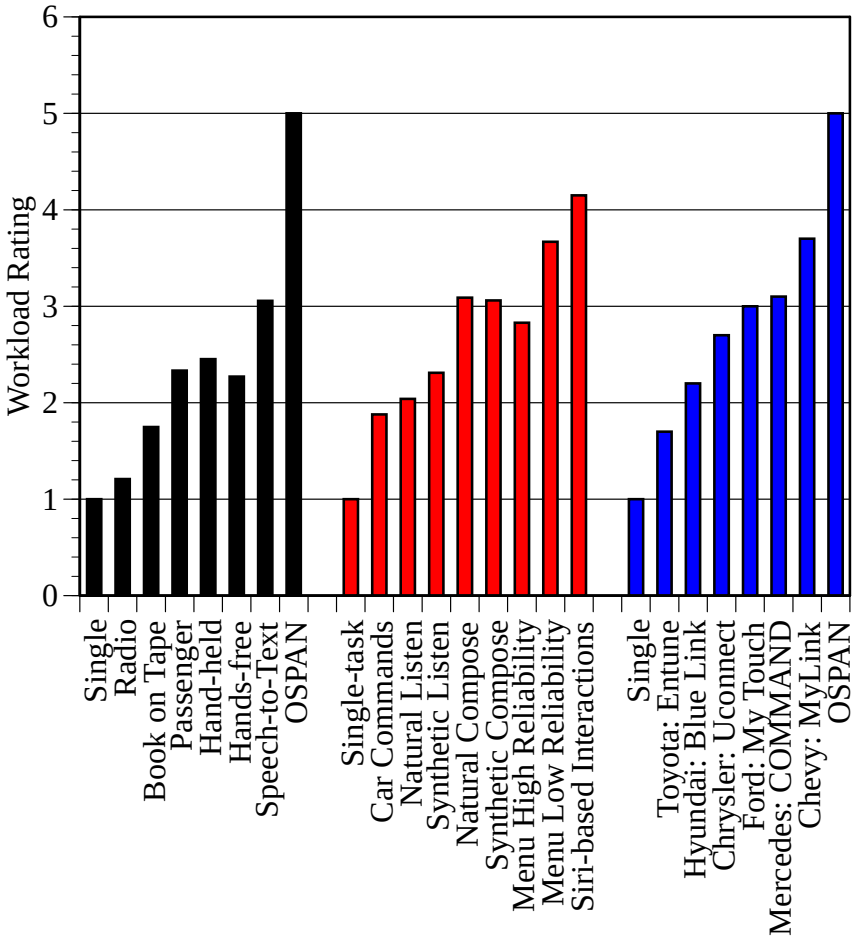


Figure 15. Workload scale for Strayer et al., (2013) (black bars), the current research (red bars), and the companion research using OEM voice-based systems (blue bars)

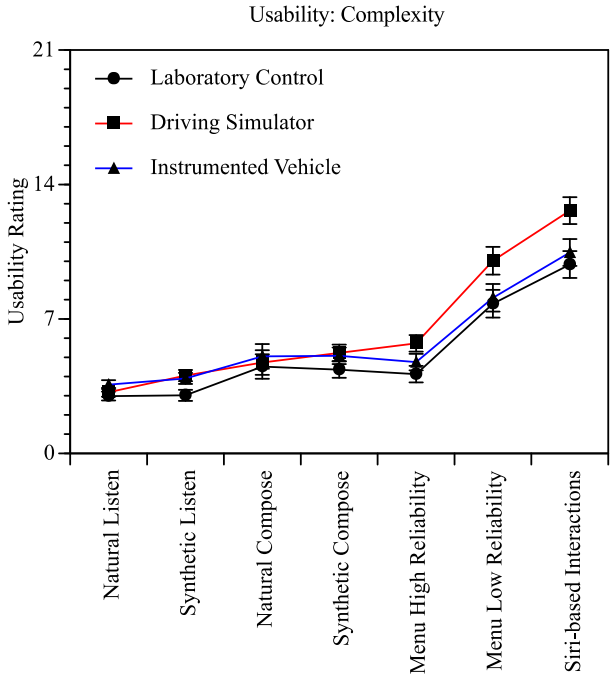


Figure A1. Intuitiveness ratings on a 21-point scale for the nine conditions

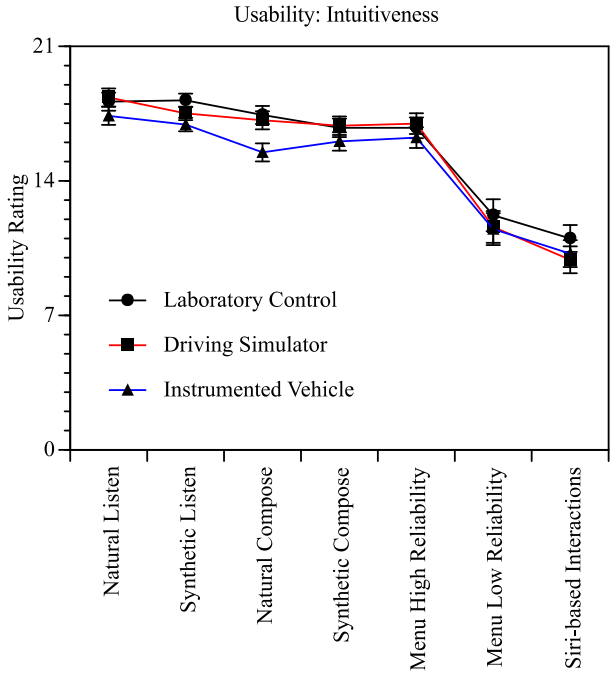


Figure A2. Complexity ratings on a 21-point scale for the nine conditions

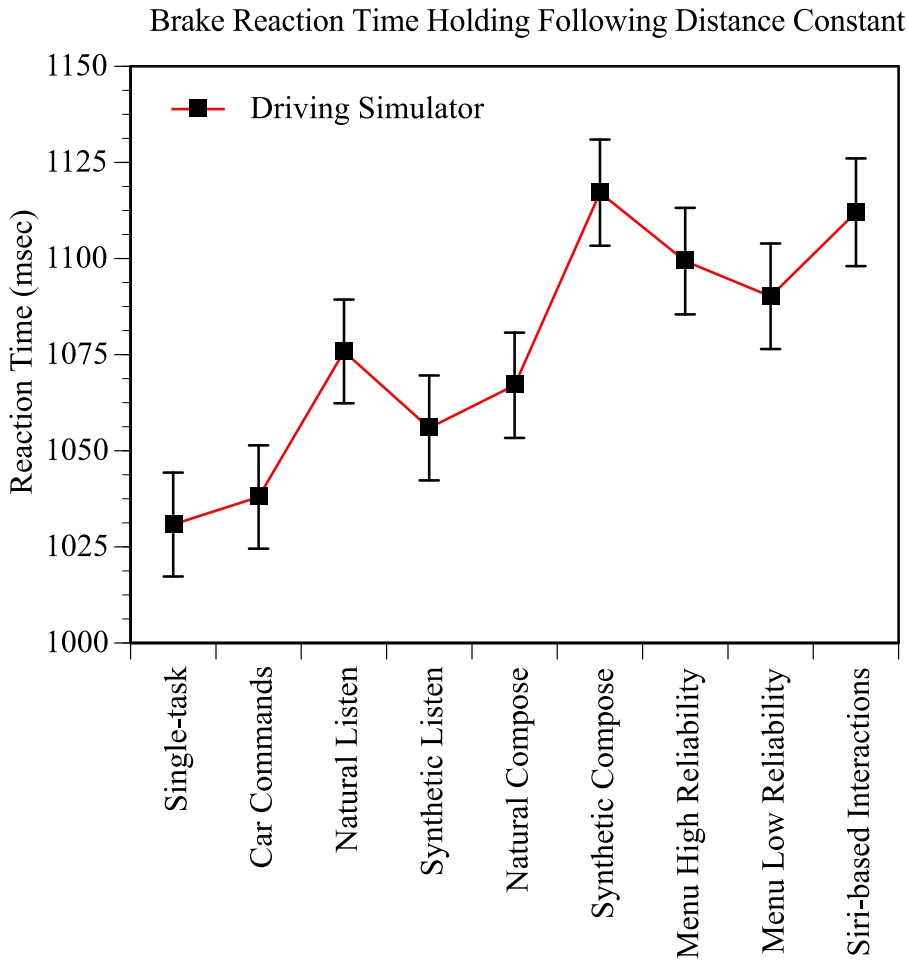


Figure A3. Brake Reaction Time Holding Constant Following Distance

Route Description for Experiment 3

- 2nd and 3rd Avenues in the Avenues of Salt Lake City, UT between U and E Streets.
- The total distance of the route is 2.7 miles, with each side of the route being 1.3 miles. The route is a suburban driving environment.
- 3rd Avenue consists of two-way traffic with a bike lane on each side and street parking off to either side. There are 4 stop signs and 1 stoplight. All are four way, meaning that traffic coming from all directions are to come to a stop.
- 2nd Avenue consists of two-way traffic with street parking off to either side. There are 5 stop signs and 1 stoplight. Four of the 5 stop signs are all-way controlled stops.
- Due to the wide nature of the streets, visibility is clear at all intersections.
- The average time to complete one loop is 8.5 minutes.
- Stop Signs
  - 3rd Ave. and N, K, I, and D Streets
  - 2nd Ave. and I, L, N, R, and U Streets (U St. is not all-way controlled)
- Stop Lights
  - 3rd Ave. and E Street
  - 2nd Ave. and E Street