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

Measuring Cognitive Distraction in the Automobile III:

A Comparison of Ten 2015 In-Vehicle Information Systems

October 2015

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 III: A Comparison of Ten 2015 In-Vehicle Information Systems. *(October 2015)*

Author

David L. Strayer, Joel M. Cooper, Jonna Turrill, James R. Coleman, and Rachel J. Hopman

University of Utah

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.

Executive Summary

This research examined the impact of In-Vehicle Information System (IVIS) interactions on the driver's cognitive workload. Two hundred fifty-seven subjects participated in a weeklong evaluation of the IVIS interaction in one of 10 different model-year 2015 automobiles. After an initial assessment of the cognitive workload associated with using the IVIS, participants took the vehicle home for five days and practiced using the system. At the end of the five days of practice, participants returned and the workload of these IVIS interactions was reassessed. The cognitive workload was found to be moderate to high, averaging 3.34 on a 5-point scale and ranged from 2.37 to 4.57. The workload was associated with the intuitiveness and complexity of the system and the time it took participants to complete the interaction. The workload experienced by older drivers was significantly greater than that experienced by younger drivers performing the same operations. Practice did not eliminate the interference from IVIS interactions. In fact, IVIS interactions that were difficult on the first day were still relatively difficult to perform after a week of practice. Finally, there were long-lasting residual costs after the IVIS interactions had terminated. The higher levels of workload should serve as a caution that these voice-based interactions can be cognitively demanding and ought not to be used indiscriminately while operating a motor vehicle.

Introduction

In order to allow drivers to maintain their eyes on the forward roadway, nearly every vehicle sold in the US and Europe can now be optionally equipped with an In-Vehicle Information System (IVIS). Using voice commands, drivers can access functions as varied as voice dialing, music selection, GPS destination entry, and even climate control. Voice activated features would seem to be a natural evolution in vehicle safety that requires little justification. Yet, a large and growing body of literature cautions that auditory/vocal tasks may have unintended consequences that adversely affect traffic safety (e.g., Bergen et al., 2014).

The National Highway Traffic Safety Administration (NHTSA) is in the process of developing voluntary guidelines to minimize driver distraction created by electronic devices in the vehicle. There are three planned 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 2015, will address visual/manual interfaces for portable and aftermarket electronic devices. Phase 3 guidelines will address voice-based auditory interfaces for devices installed in vehicles and for portable aftermarket devices.

Currently; however, there are no unified regulations regarding the use of wireless technology in the vehicle – the NHTSA Phase 1 guidelines are voluntary and it is unknown whether any of the currently available vehicles meet these guidelines. With the explosive growth in technology, the problem of driver distraction is poised to become much more acute.

Benchmarking Cognitive Distraction

Our prior research provided a benchmark for the cognitive workload associated with common in-vehicle activities (Strayer et al., in press; see also Cooper et al., 2014; and Strayer et al., 2013; Strayer et al., 2014). In our studies, we developed and validated a cognitive distraction scale based on converging operations from the laboratory, driving simulator, and using an instrumented vehicle driven in a residential section of Salt Lake City. Our research shows that the distraction potential can be reliably measured, that cognitive workload systematically varies as a function of the secondary task performed by the driver, and that some activities, particularly newer voice-based interactions in the vehicle, are associated with surprisingly high levels of mental workload.

We obtained workload ratings attributable to cognitive sources by comparing seven different concurrent tasks with a "single-task" condition where the drivers did not perform any concurrent secondary-task activity (Strayer et al., 2013). The seven tasks were listening to the radio, listening to a book on tape, talking to a passenger, talking on a hands-free cell phone, talking on a hand-held cell phone, interacting with a simple voice messaging system, and a cognitively demanding Operation Span (OSPAN) task that was used for calibration. In our distraction scale, the non-distracted single-task driving anchored the low-end (Category 1) and the mentally demanding OSPAN task anchored the high-end (Category 5) of the scale. Using this method, we found that 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, are associated with moderate increases in cognitive distraction. Finally, activities such as using a speech-to-text system to send and receive short text or e-mail messages produced a surprisingly high level of cognitive distraction.

The speech-to-text system that we evaluated in the laboratory is noteworthy because the speech-recognition portion of the system was perfectly reliable and there was no requirement to review, edit, or correct garbled translations. In our research protocol, perfect speech recognition was implemented using a "Wizard-of-Oz" paradigm (Kelley, 1983; Lee et al., 2001), in which the participant's speech was secretly entered into the computer by the experimenter with no transcription errors. 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 "best case" speech-to-text e-mail/text message system received a Category-3 rating on the cognitive distraction scale.

In our 2014 research (Strayer et al., 2014) we examined voice-based interactions in greater detail. We found that just listening to voice-messages without the possibility of generating a reply was associated with a cognitive workload rating comparable to that of conversing on a cell phone (i.e., Category-2). However, when drivers composed replies to these messages, the workload rating increased to a Category-3 rating on the cognitive distraction scale. Like our earlier testing, this laboratory-based system was perfectly reliable. We also found no systematic difference between the natural (i.e., human) and synthetic (i.e., computerized) delivery of the messages. 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 mental workload.

Our 2014 research also evaluated Apple's intelligent personal assistant, Siri, to send and receive text messages, update Facebook or Twitter, and to modify and review calendar appointments. To create a completely hands-free version of the interaction, a lapel microphone was clipped to the participant's collar and they activated Siri with the command "Hey Siri," at which point a researcher manually activated the device. Drivers neither looked at nor made physical contact with the iPhone during these interactions. Even so, the workload ratings for these interactions exceeded Category 4 on our workload scale. Moreover, there were two crashes in the driving simulator study when participants were using Siri.

The primary difference between our laboratory-based speech-to-text system and the Siri-based interactions was the reliability of the system (see also Strayer et al., in press). Siri was error-prone, producing different responses to seemingly identical commands. In other circumstances, Siri required exact phrasing to accomplish specific tasks and subtle deviations from that phrasing would result in a failure. Moreover, 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. For these reasons and others, voice-based interactions using an intelligent personal assistant, such as Siri, were significantly more mentally demanding than conversing on a cell phone.

Research Objectives and Experimental Overview

The current research addresses several important issues related to the assessment of cognitive workload in the vehicle. First, our prior research examined drivers who were in their mid-20's (e.g., the average age of participants in the Strayer et al. (2013) study was 23). This younger cohort tends to be more tech-savvy than an older population: it is unclear how demanding older drivers will find these voice-based interactions. This issue gains importance because drivers between the ages of 55 and 64 are the most likely to purchase new vehicles equipped with voice-command technology to control infotainment and other vehicle functions (Sivak, 2013). In fact, laboratory studies have documented substantially greater costs of multitasking for older adults (e.g., Hartley & Little, 1999; Kramer & Larish, 1996; McDowd & Shaw, 2000); therefore, it is likely that the workload scale developed in our prior research is a conservative estimate of the cognitive workload experienced by older drivers interacting with these voice-based systems.

Second, our prior research examined the driver's cognitive workload soon after they had been introduced to the vehicle, with minimal training (i.e., 15 minutes or less) using the vehicle and the IVIS. The old adage "practice makes perfect" suggests that extended practice with the IVIS may reduce or even eliminate the interference caused by these voice-based interactions. However, for practice to be effective the system needs to be intuitive and error free with a consistent mapping between input-output operations (e.g., Shiffrin & Schneider, 1977). Because many of the systems that are currently available tend to be complex and error prone, with inconsistent behavior (e.g., Cooper, Ingebretsen, & Strayer, 2014), there are limits on how much improvement can be expected with extended practice.

Our study recruited male and female drivers between the ages of 21 and 70 to participate in a weeklong evaluation of IVIS interactions in one of 10 different model-year 2015 automobiles. After familiarization with the vehicle, participants were trained on how to interact with the voice-based system to perform common IVIS tasks (e.g., dialing, radio tuning). After this initial orientation, they were tested on the IVIS interactions using the method that we developed to assess cognitive workload in the vehicle (e.g., Strayer et al., 2013). Participants then took the vehicle home for a five days and practiced interacting with the IVIS. At the end of five days of practice, participants returned and were retested on the cognitive workload of these same IVIS interactions. This allowed us to evaluate the effects of age and practice on these IVIS interactions.

Methods

Participants

Following approval from the Institutional Review Board, participants were recruited by word of mouth and flyers posted on the University of Utah campus. They were compensated $250 upon completion of the weeklong study. Data were collected from July 4th of 2014 through June 18th of 2015.

Two hundred fifty-seven subjects participated in the study (127 males, 130 females). The youngest was 21 and the oldest was 70 years old, with an average age of 44. Participants were recruited to provide a minimum of 4 male and 4 female licensed drivers in each of the three age groups, 21-34, 35-53, 54-70, for each of the 10 vehicles. An accounting of participants' gender and age group is provided in Table 1.

Table 1. Distribution of age and gender for each of the vehicles used in the experiment.

	Buick LaCrosse		Chevy Equinox		Chevy Malibu		Chrysler 200c		Ford Taurus	
Age Categories	M	F	M	F	M	F	M	F	M	F
21-34	4	4	4	4	4	4	4	5	4	4
35-53	4	4	5	4	5	5	4	5	5	5
54-70	5	4	5	4	4	5	4	5	4	4

	Hyundai Sonata		Mazda 6		Nissan Altima		Toyota 4Runner		VW Passat	
Age Categories	M	F	M	F	M	F	M	F	M	F
21-34	4	4	4	4	4	5	4	4	4	5
35-53	5	4	4	4	4	4	4	4	4	4
54-70	5	5	4	5	4	4	4	4	4	4

Prior to participation in the research, the University of Utah's Division of Risk Management ran a Motor Vehicles Record report on each prospective participant to ensure a clean driving history (e.g., no at-fault accidents in the past five years) and eligibility to be registered as a University driver. In addition, following University of Utah policy, each participant was required to complete a 20-minute online defensive driving course and pass the certification test. Participants reported between 5 and 55 years of driving experience with an average of 28 years. Additionally, participants reported driving an average of 160 miles per week. All participants were recruited from the greater Salt Lake area and spoke with a western US English dialect.

Materials and Equipment

Ten 2015 model year vehicles, equipped with automatic transmissions, were used in this research (see Appendix A for a complete breakdown of the different vehicles used in the study). In each vehicle, voice-based interactions with the IVIS were initiated with the press

of a button located on the steering wheel and ended either automatically or with a second press of the button, depending on the vehicle and function. Each of the ten vehicle-systems allowed drivers to complete contact calling and number dialing tasks through a Bluetooth paired smartphone.

Dual-Vision XC cameras, manufactured by Rosco Vision Systems, were installed in the vehicles by a qualified technician. Cameras were mounted under the rear view mirror, providing a view of the forward roadway and of the driver's face. An infrared illuminator was installed in each vehicle for nighttime video recording. The cameras also included an embedded GPS system. Cameras were set to automatically begin recording audio, video, and GPS data as soon as the vehicle ignition was turned on by the driver and to stop recording when the vehicle ignition was turned off. Video data were recorded at 3.5 frames per second at standard VGA resolution.

During the first day of the study (Session 1) and on the last day of the study (Session 2), participants wore a head-mounted Detection Response Task (DRT) device that was manufactured by Precision Driving Research. The DRT protocol for device placement and stimulus onset characteristics followed the specifications outlined in ISO WD 17488 (2015). The device consisted of an LED light mounted to a flexible arm that was connected to a headband, a micro-switch attached to the participant's left or right thumb (the switch was attached to the hand opposite that of the vehicle's steering wheel voice-activation button), and a dedicated microprocessor to handle all stimulus timing and response data. The light was positioned in the periphery of the participant's left eye (approximately 15° to the left and 7.5° above the participant's left eye) so that it could be seen while looking at the forward roadway but did not obstruct their view of the driving environment. The stimulus presentation configuration adhered to the ISO standard 17488 with red LED stimuli configured to flash every 3-5 seconds. Data was collected using an Asus Transformer Book T100s with quad-core Intel® Atom™ processors running at 1.33GHz.

An auditory version of the OSPAN task, developed by Watson and Strayer (2010), was used to induce a high workload baseline during testing. This task required participants to recall single syllable words in serial order while solving mathematical problems. In the auditory OSPAN task, participants were asked to remember a series of two to five words that were interspersed with math-verification problems (e.g., given "[3 / 1] – 1 = 2?" – "cat" – "[2 x 2] + 1 = 4?" – "box" – RECALL, the participant should have answered "true" and "false" to the math problems when they were presented and recalled "cat" and "box" in the order in which they were presented when given a recall probe). In order to standardize presentation for all participants, a prerecorded version of the task was created and played back during testing.

Subjective workload ratings were collected using the NASA TLX survey developed by Hart and Staveland (1988). After completing each of the conditions (single-task, IVIS, and OSPAN, see below for details) in the experiment, participants responded to the NASA TLX survey consisting of six questions that used 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?

A study facilitator was assigned to each participant for the duration of the data collection session. Facilitators were trained to precisely administer the research procedure and adhered to a scripted evaluation protocol. Additionally, facilitators were responsible for ensuring the safety of the driver, providing in-car training, and delivering task cues to participants. All facilitators had a current driver's license and were over the age of 21.

Procedure

Before the study began, participants filled out an IRB approved consent form and a brief intake questionnaire to assess basic characteristics of phone and driving usage and experience. Participants were then familiarized with the controls of the instrumented vehicle, adjusted the mirrors and seat, and were informed of the conditions that would be completed while driving. The first portion of training involved an introduction to the DRT device. Participants were fitted with the device and were instructed on its functionality. Once comfortable with the general procedure, they were allowed to practice with the DRT device until they felt comfortable with its usage. In most cases, participants were comfortable with the functionality of the device within a couple of minutes. Participants then completed a three-minute orientation for each of the tasks in the IVIS condition and a three-minute orientation of the OSPAN task while the vehicle was parked. Participants were provided training on the functionality of the IVIS system and asked to complete a series of contact calling, number dialing, and radio tuning tasks until they reached proficiency. A practice loop within a parking lot was completed in order to familiarize the participant with the handling of the vehicle.

Next, participants completed one circuit around the 2.7-mile driving loop, located in the Avenues section of Salt Lake City, UT in order to become familiar with the route itself. The route provided a suburban/residential driving environment and contained seven all-way controlled stop signs, one two-way stop sign, and two stoplights. Given the restricted usage characteristics of the roadway, traffic remained relatively consistent during testing. After the practice drive, participants began the experimental portion of the study. In total, participants drove the vehicle for approximately 20 minutes before the initial data collection began.

Six tasks were given to participants during the IVIS condition of the study; each involved the use of the vehicle's unique voice-activated infotainment system. The tasks were initiated once participants reached pre-specified locations that were chosen to allow participants approximately 1.5 minutes to complete each task. If the participant was unable to complete a task before the next task was to begin, they were told to abandon that first task and move on to the new one.

All of the tasks in the IVIS condition began when participants pressed the voice activation button located on the steering wheel. Once initiated, each of the tasks was completed through auditory + vocal system interactions. System interactions were performed in a

fixed order and alternated between completing a phone calling task and a radio-tuning task. The tasks in the IVIS condition were as follows:

Task 1: "Call from your contacts Joel Cooper"
Task 2: "Tune your radio to 98.3 FM" once completed "Tune your radio to 1320 AM"
Task 2b (for the Nissan and Volkswagen vehicles): "Call from your contacts Chris Hunter"
Task 3: "Dial your own phone number"
Task 4: "Tune your radio to 1160 AM" once completed… "Tune your radio to 90.1 FM"
Task 4b (for the Nissan and Volkswagen vehicles): "Dial your own phone number"
Task 5: "Call from your contacts Amy Smith at work"
Task 6: "Dial your own phone number"

Participants were then familiarized with the specific requirements of the upcoming condition and were told that their task was to follow the route previously practiced while complying with all local traffic rules, including obeying a 25 mph speed limit. Throughout each of the three experimental conditions (single-task, IVIS, and OSPAN), the driver performed the DRT task. Any driving sections with turns were excluded from the DRT and video analyses to minimize the potential of a manual distraction confound.

At the conclusion of the first day of testing (Session 1), participants were given a logbook to document their interactions with the IVIS during the ensuing five days. Participants were encouraged to practice using the IVIS system on their own time with special emphasis given to contact calling, number dialing, and radio station selection. Finally, participants were instructed not to allow other drivers to use the vehicle; however, passengers were acceptable in order to match the driver's normal weekly pattern of driving. Once familiar with the journaling and instructions for the week, participants took the research vehicle home and began the practice portion of the study. Following the five-day practice interval, participants returned on the last day for evaluation (Session 2). The data collection protocol for Session 2 was identical to that of Session 1 except that the extensive IVIS training was no longer necessary.

Design

The core experimental design was a 3 (Age) x 10 (Vehicle) X 3 (Condition) x 2 (Session) Split-Plot Factorial. Age was a between subject factor and included three Age Groups: 21-34, 35-53, and 54-70.[1] Vehicle was also a between subjects factor and included ten 2015 model year vehicles: a Buick LaCrosse with IntelliLink, a Chevy Equinox with MyLink, a Chevy Malibu with MyLink, a Chrysler 200c with Uconnect, a Ford Taurus with Sync MyFord Touch, a Hyundai Sonata with Blue Link, a Mazda 6 with Connect, a Nissan Altima with NissanConnect, a Toyota 4Runner with Entune, and a Volkswagen Passat with Car-Net. Condition was a 3-level within-subjects factor (single-task, IVIS, and OSPAN conditions). Session was also a within-subjects factor and refers to the first day of testing (Session 1) and the last day of testing (Session 2) that were separated by five days of practice with the IVIS system. The three Conditions in each session were performed in a counterbalanced order across participants. Interactions with the IVIS involved 2 number dialing tasks, 2 contact calling tasks, and 4 radio tuning tasks, with the exception that

[1] The analyses reported below show the same pattern as when Age is treated as a continuous variable rather than a categorical variable.

participants driving the Nissan and Volkswagen vehicles completed 3 number dialing tasks and 3 contact calling tasks because these vehicles did not support radio tuning. Additionally, because the DRT analysis allowed for a differentiation between on-task performance (i.e., the time when participants were actively engaged in the IVIS interactions) and off-task performance (i.e., the period of time between IVIS tasks when the driver was not interacting with the IVIS, but rather was driving as in the single-task condition), Condition had 4 factors (single-task, IVIS off-task (i.e., IVIS-0), IVIS on-task (i.e., IVIS-1), and OSPAN) when assessing the effects of IVIS interactions on DRT performance.

Dependent Measures

Cognitive workload was determined by a number of performance measures. These were derived from the DRT task, subjective reports, and analysis of video recorded during the experiment.

DRT data were cleaned following procedures specified in ISO 17488 (2015). Consistent with the standard, all responses briefer than 100 msec or greater than 2500 msec were rejected for calculations of Reaction Time. Responses that occurred later than 2.5 seconds from the stimulus onset were coded as misses. Any DRT data collected around turns was flagged and removed from analysis. During testing of the IVIS interactions, trial engagement was flagged by the facilitator through a keyboard press which allowed the identification of segments of the IVIS condition when the participant was actively engaged in an activity (IVIS-1) or had finished that activity and was operating the vehicle without voice-based interactions (IVIS-0).

- DRT – MANOVA. An overall analysis that statistically combined the effects of Reaction Time and Hit Rate (See below).
- DRT –Reaction Time. Defined as the sum of all valid reaction times to the DRT task divided by the number of valid reaction times.
- DRT – Hit Rate. Defined as the number of valid responses divided by the total number of stimuli presented during each condition.
- DRT – Residual Costs. To evaluate the residual effects of secondary task interactions on DRT Reaction Time, performance in the off-task segments of the drive was sorted into 3-second bins relative to the time that the off-task interval began. For example, a DRT event occurring 5 seconds after the end of an IVIS interaction would be sorted into the second bin.

Following each drive, participants were asked to fill out a brief questionnaire that posed 8 questions related to the just completed task. The first 6 of these questions were from the NASA TLX; the final 2 assessed the intuitiveness and complexity of the IVIS interactions.

- Subjective – NASA TLX. Defined as the response on a 21-point scale for each of the 6 subscales of the TLX (Mental, Physical, Temporal, Performance, Effort, and Frustration).
- Subjective – Intuitiveness and Complexity. Defined as the response on a 21-point scale to questions on task intuitiveness (i.e., “how intuitive, usable, and easy was it to use the system”) and complexity (i.e., “how complex, difficult, and confusing was it to use the system”).

Task Completion Time, Glance Location, and Practice Frequency were derived from the video recordings. Task Completion Time and Glance Location were available for 214/257 participants, while video analysis of Practice Frequency was available for 180/257 participants. In all cases, frame-by-frame analysis was completed, sampling 2 frames per second. The reliability of the coding was assessed through an evaluation of the time-on-task data from the DRT and the coded videos. Results from this assessment indicated that the two sources showed a nearly identical pattern ($r = .96$).

- Video – Task Completion Time. Task completion time was defined as the time from the moment participants first pressed the voice activation button to the time that the same button was pressed to terminate a task, or in the case of radio tuning, the moment when the system accurately carried out the requested task. Task completion time reflects the average task duration across the 6 tasks in the IVIS condition.
- Video – Glance Location. Defined as the percentage of all visual glances that fell within the forward roadway, the dashboard region, or the right, left, and rear-view mirrors.
- Video – Practice Frequency. Defined as the count of IVIS voice interactions during the 5-day practice session where participants practiced using the voice assistant to call a contact, dial a number, tune the radio, or engage in other voice tasks.

Results

DRT

The DRT data reflect the response to the onset of the red light in the peripheral detection task. RT was measured to the nearest millisecond. Hit Rate was calculated based on a response to the red light, which was coded as a "hit", and non-responses to a red light, which were coded as a "miss." The RT and Hit Rate data for the DRT task are plotted as a function of Age X Condition in Figures 1 and 2, respectively. The data from the DRT task are also plotted as a function of Session X Condition in Figures 3 and 4, respectively. The data are broken down by active involvement in the IVIS condition, denoted by a suffix of "-1," (e.g., IVIS-1) or when participants were operating the vehicle without concurrent secondary-task interaction, denoted by a suffix of "-0" (e.g., IVIS-0).

MANOVA

The DRT data were first analyzed using a 3 (Age) X 10 (Vehicle)[2] X 4 (Condition) X 2 (Session) MANOVA that included both Reaction Time and Hit Rate as dependent variables.[3] The results of the MANOVA are presented in Table 2. There were significant main effects of Age, $F(4, 454) = 14.07$, $p < .001$, $\eta^2 = .110$; Condition, $F(6, 1362) = 164.86$, $p < .001$, $\eta^2 = .421$; and Session, $F(2, 226) = 48.61$, $p < .001$, $\eta^2 = .301$. In addition, Condition interacted with Age, $F(12, 1362) = 8.15$, $p < .001$, $\eta^2 = .067$; Vehicle, $F(54, 1362) = 1.53$, $p = .009$, $\eta^2 = .057$; and Session, $F(6, 1362) = 12.54$, $p < .001$, $\eta^2 = .052$. None of the other effects were significant.

Reaction Time

The reaction time data from the DRT were analyzed using a 3 (Age) X 10 (Vehicle) X 4 (Condition) X 2 (Session) ANOVA; the results of which are presented in Table 3. The analysis revealed significant main effects of Age, $F(2, 227) = 31.71$, $p < .001$, $\eta^2 = .218$; Condition, $F(3, 681) = 894.29$, $p < .001$, $\eta^2 = .798$; and Session, $F(1, 227) = 84.65$, $p < .001$, $\eta^2 = .272$. In addition, Condition interacted with Age, $F(6, 681) = 15.75$, $p < .001$, $\eta^2 = .122$; Vehicle, $F(27, 681) = 2.00$, $p = .002$, $\eta^2 = .074$; and Session, $F(3, 681) = 16.62$, $p < .001$, $\eta^2 = .068$. None of the other effects were significant.

[2] The Vehicle condition codes for all data collected in each vehicle. Thus, a significant effect of Vehicle would reflect general differences in performance associated with *driving* the vehicle and not differences in the IVIS interface. Differences in the IVIS interfaces are seen in the effect of Condition and the Condition by Vehicle interaction.

[3] A preliminary analysis that included Gender as a factor found that males responded, on average, 45 msec faster than females, ($p < .001$); however, Gender did not interact with any of the other factors (all p's > .200), hence we collapsed across this variable for all additional analyses.

Table 2. MANOVA results on DRT. A = Age, V = Vehicle, C = Condition, and S = Session.

	df_n	df_d	F	p	η^2
A	4	454	14.07	.001**	.110
V	18	454	1.38	.138	.052
AxV	36	454	1.09	.336	.080
C	6	1362	164.86	.001**	.421
CxA	12	1362	8.15	.001**	.067
CxV	54	1362	1.53	.009*	.057
CxAxV	108	1362	0.76	.968	.057
S	2	226	48.61	.001**	.301
SxA	4	454	1.16	.326	.010
SxV	18	454	0.88	.609	.034
SxAxV	36	454	1.04	.410	.076
CxS	6	1362	12.54	.001**	.052
CxSxA	12	1362	1.39	.164	.012
CxSxV	54	1362	0.89	.699	.034
CxSxAxV	108	1362	1.04	.384	.076

* $p < .05$, ** $p < .001$

Table 3. ANOVA results on Reaction Time. A = Age, V = Vehicle, C = Condition, and S = Session.

	df_n	df_d	F	p	η^2
A	2	227	31.71	.001**	.218
V	9	227	1.58	.121	.059
AxV	18	227	0.96	.500	.071
C	3	681	894.29	.001**	.798
CxA	6	681	15.75	.001**	.122
CxV	27	681	2.00	.002*	.074
CxAxV	54	681	0.89	.688	.066
S	1	227	84.65	.001**	.272
SxA	2	227	0.48	.621	.004
SxV	9	227	0.46	.900	.018
SxAxV	18	227	0.69	.820	.052
CxS	3	681	16.62	.001**	.068
CxSxA	6	681	1.13	.341	.010
CxSxV	27	681	0.76	.807	.029
CxSxAxV	54	681	0.94	.596	.069

* $p < .05$, ** $p < .001$

Hit Rate

The Hit Rate data from the DRT task were analyzed using a 3 (Age) X 10 (Vehicle) X 4 (Condition) X 2 (Session) ANOVA; the results are presented in Table 4. The analysis revealed significant main effects of Age, $F(2, 227) = 17.87$, $p < .001$, $\eta^2 = .136$; Condition, $F(3, 681) = 129.15$, $p < .001$, $\eta^2 = .363$; and Session, $F(1, 227) = 53.61$, $p < .001$, $\eta^2 = .191$. In addition, Condition interacted with Age, $F(6, 681) = 7.94$, $p < .001$, $\eta^2 = .065$; Vehicle, $F(27, 681) = 1.87$, $p = .005$, $\eta^2 = .069$; and Session, $F(3, 681) = 12.44$, $p < .001$, $\eta^2 = .052$. None of the other effects were significant.

Table 4. ANOVA results on Hit Rate. A = Age, V = Vehicle, C = Condition, and S = Session.

	df_n	df_d	F	p	η^2
A	2	227	17.87	.001**	.136
V	9	227	1.25	.264	.047
AxV	18	227	1.57	.069	.111
C	3	681	129.15	.001**	.363
CxA	6	681	7.94	.001**	.065
CxV	27	681	1.87	.005*	.069
CxAxV	54	681	0.82	.815	.061
S	1	227	53.61	.001**	.191
SxA	2	227	1.88	.155	.016
SxV	9	227	0.79	.628	.030
SxAxV	18	227	1.59	.065	.112
CxS	3	681	12.44	.001**	.052
CxSxA	6	681	1.76	.101	.015
CxSxV	27	681	0.99	.482	.038
CxSxAxV	54	681	1.12	.268	.081

* $p < .05$, ** $p < .001$

The Condition X Age interaction, (see Figures 1 and 2), indicates that the costs of the IVIS interactions were greater for older adults than for younger adults. RT increased with age by 18.2 % in the single-task condition and by 29.7% in the IVIS-1 condition. A similar analysis of Hit Rates found a decrease with age of 2.1% in the single-task condition and of 8.5% in the IVIS-1 condition.

The Condition X Session interaction, (see Figures 3 and 4), indicates that the effects of practice were more pronounced when participants were using the IVIS than when they were in the single-task condition. RT decreased with practice by 3.5 % in the single-task condition and by 9.0% in the IVIS-1 condition. A similar comparison on Hit Rates found an increase with practice of 1.4% in the single-task condition and of 5.7% in the IVIS-1 condition.

Figure 5 presents the average of z-transformed DRT data (i.e., a weighted average of Reaction Time and Hit Rate data) plotted as a function of Vehicle in the IVIS condition. For comparison, performance in the single-task and OSPAN conditions are also included in Figure 5. To better understand the Condition X Vehicle interactions reported in Tables 2-4, a between-subjects Analysis of Variance (ANOVA) was performed on the z-transformed data from the IVIS condition. This analysis revealed a significant effect of Vehicle, $F(9, 247) = 2.03$, $p = .037$. By contrast, a similar analysis on the z-transformed data from the single-task and OSPAN conditions failed to yield a significant effect of Vehicle, $F(9, 247) = 0.16$, $p = .320$ and $F(9, 247) = 1.04$, $p = .411$, respectively. Moreover, an Analysis of Covariance (ANCOVA) on the data obtained in the IVIS condition that held constant any performance differences in the single-task condition, also found a significant effect of the IVIS voice-based interaction, $F(9, 246) = 3.29$, $p < .001$, $\eta^2 = .107$. This pattern is important because it indicates that there were significant differences in DRT performance when our drivers were interacting with the IVIS, but there were no significant differences in DRT performance when they were just driving the vehicle.

Residual Costs

A surprising finding was that the off-task performance in the DRT task differed significantly from single-task performance. Given that drivers were not engaged in any secondary-task activities during the off-task portions of the drive, it suggests that there were residual costs that persisted after the IVIS interaction had terminated. Figure 6 presents the residual costs plotted as a function of the time since the IVIS interaction terminated. In Figure 6, "O" refers to performance in the OSPAN task and "S" refers to single-task performance. The filled circles reflect the average RT as a function of sorting bin and the solid blue line reflects the best-fitting power function describing the relationship between RT and bin:

$$f(x) = a * (x^{-.1878072}), \text{ where } a = \exp(6.691554), \text{ with } R^2 = .98.$$

Residual cost functions were also generated for each age group and they are plotted in Figure 7. In Figure 7, the effects of age are clearly evident as an intercept offset; however, the residual costs are very similar in duration across the three age groups.

$$\text{Younger-Age: } f(x) = a * (x^{-.1938970}), \text{ where } a = \exp(6.602465), \text{ with } R^2 = .97.$$

$$\text{Middle-Age } f(x) = a * (x^{-.1671658}), \text{ where } a = \exp(6.653588), \text{ with } R^2 = .98.$$

$$\text{Older-Age: } f(x) = a * (x^{-.1902559}), \text{ where } a = \exp(6.788466), \text{ with } R^2 = .94.$$

The residual costs took a significant amount of time to dissipate. In fact, the data indicate that off-task performance reflects a mixture of "single-task" performance and the persistent costs associated with the IVIS interactions from the immediately preceding on-task period. One way to contextualize these residual cost is to use logic underlying the workload scale developed by Strayer et al., (2013) to estimate, based solely on the DRT reaction time data, when the cognitive workload would reach a Category-4 level (approximately 6 seconds), when it would reach a Category-3 level (approximately 9 seconds), and when it would reach a Category-2 level (approximately 15 seconds). The residual costs are notable because of their magnitude, their duration, and the fact that they are obtained even when there is no active switch to perform another task. They appear to reflect the lingering act of disengaging from the cognitive processing associated with the IVIS task and fully reengaging attention to the driving environment. From a practical perspective, the data indicate that just because a driver terminates a call or text message does not mean that they are no longer impaired. Indeed, significant residual costs were observed for 27 seconds after the IVIS interaction had terminated. At the 25 MPH speed limit in our study, drivers would have traveled over the length of a three football field during this interval.

Subjective

Subjective assessments of workload were made using the NASA TLX and supplementary questions on the intuitiveness and complexity of the IVIS systems.

NASA TLX

The 6 scales of the NASA TLX were analyzed using a 3 (Age) X 10 (Vehicle) X 4 (Condition) X 2 (Session) ANOVA. The TLX data are plotted as a function of Condition in Figure 8, as a function of Session in Figure 9, and as a function of Age in Figure 10. The results of the ANOVA are presented in Table 5. There were significant main effects of Vehicle, $F(54, 1362) = 1.47$, $p = .016$, $\eta^2 = .055$; Condition, $F(12, 900) = 72.10$, $p < .001$, $\eta^2 = .490$; and

Session, $F(6, 222) = 28.51$, $p < .001$, $\eta^2 = .435$. In addition, Condition interacted with Age, $F(24, 1880) = 2.46$, $p < .001$, $\eta^2 = .032$; Vehicle, $F(108, 2724) = 1.60$, $p < .001$, $\eta^2 = .060$; and Session, $F(12, 900) = 3.36$, $p < .001$, $\eta^2 = .043$. The Session X Vehicle, $F(54, 1362) = 1.36$, $p = .045$, $\eta^2 = .051$, and the Session X Age X Vehicle interactions were also significant, $F(108, 1362) = 1.30$, $p = .025$, $\eta^2 = .094$. None of the other effects were significant.

Table 5. ANOVA results on the NASA TLX. A = Age, V = Vehicle, C = Condition, and S = Session.

	df_n	df_d	F	p	η^2
A	12	446	1.41	.159	.036
V	54	1362	1.47	.016*	.055
AxV	108	1362	1.04	.366	.076
C	12	900	72.10	.001**	.490
CxA	24	1880	2.46	.001**	.032
CxV	108	2724	1.60	.001**	.060
CxAxV	216	2724	1.10	.150	.081
S	6	222	28.51	.001**	.435
SxA	12	446	1.04	.441	.027
SxV	54	1362	1.36	.045*	.051
SxAxV	108	1362	1.30	.025*	.094
CxS	12	900	3.36	.001*	.043
CxSxA	24	1808	1.44	.076	.019
CxSxV	108	2724	1.99	.503	.038
CxSxAxV	216	2724	1.11	.137	.081

* $p < .05$, ** $p < .001$

Figure 11 presents the average of z-transformed TLX data plotted as a function of Vehicle in the IVIS condition. For comparison, performance in the single-task and OSPAN conditions is also included in Figure 11. A between-subjects ANOVA that compared the z-transformed data from the IVIS condition found a significant effect of Vehicle, $F(9, 247) = 3.08$, $p = .002$. A similar analysis on the z-transformed data found a significant effect of Vehicle in the single-task condition, $F(9, 247) = 1.96$, $p = .044$ (a post-hoc analysis found that the Mazda, Hyundai, and Nissan vehicles had higher NASA TLX workload ratings than the VW and Equinox), but not in the OSPAN condition $F(9, 247) = 1.21$, $p = .292$. An ANCOVA on the data from the IVIS condition that held constant the performance differences observed in the single-task condition, also found a significant effect of IVIS interaction, $F(9, 246) = 2.93$, $p = .003$, $\eta^2 = .097$. As with the DRT data reported above, this pattern is important because it indicates that there were significant differences in TLX performance when our drivers were interacting with the IVIS, over and above any differences when participants were just driving the vehicle.

Intuitiveness

Participants were also asked to rate how intuitive, usable, and easy it was to use the IVIS. Figure 12 presents the intuitiveness ratings on a 21-point scale where 1 reflected "not at all" and 21 reflected "very much." A 3 (Age) by 10 (Vehicle) X 2 (Session) split-plot ANOVA found that intuitiveness varied as a function of Vehicle, $F(9, 227) = 4.55$, $p < .001$, $\eta^2 = .153$. None of the other effects were significant (all other p's > .14).

Complexity

Participants were also asked to rate how complex, difficult, and confusing it was to use the IVIS. Figure 13 presents the complexity ratings on a 21-point scale where 1 reflected "not at all" and 21 reflected "very much." A 3 (Age) by 10 (Vehicle) X 2 (Session) split-plot ANOVA found that complexity ratings varied as a function of Age (i.e., older adults found the IVIS interactions to be more complex), $F(2, 227) = 6.21$, $p = .002$, $\eta^2 = .052$ and Vehicle, $F(9, 227) = 4.82$, $p < .001$, $\eta^2 = .160$. None of the other effects was significant (all other p's > .07).

Video Analysis

Three performance measures were derived from analysis of the video. These were: Task Completion Time, Glance Location, and Practice Frequency.

Task Completion Time

Task completion time is plotted in Figure 14**.** The data were analyzed using a 3 (Age) X 10 (Vehicle) repeated measures ANOVA. As can be seen in the figure, the time to complete the task varied as a function of Vehicle, $F(9, 165) = 22.56$, $p < .001$, $\eta^2 = .552$. The main effect of Age, $F(1, 165) = 2.72$, $p = .069$, $\eta^2 = .032$ was not significant; however, the Age X Condition interaction was, $F(18, 165) = 2.09$, $p = .008$, $\eta^2 = .108$. This interaction indicates that older adults tended to have more difficulty with the more demanding IVIS interactions than younger adults. Planned comparisons revealed that participants took longer to perform the IVIS tasks with the Nissan than with the Mazda and VW (which did not differ), and that task completion time was greater for these three vehicles than the rest of the vehicles (which did not differ from each other).

Glance Location

The percentage of time that drivers spent looking forward, down, and scanning mirrors was analyzed using a 3 (Age) X 10 (Vehicle) X 3 (Condition) X 2 (Session) X 3 (Glance Location) repeated measures ANOVA. The results of the ANOVA are presented in Table 6. Glance Location is plotted as a function of Condition in Figure 15. There was a significant main effect of Glance Location, $F(2, 412) = 1247$, $p < .001$, $\eta^2 = .868$, and the Glance Location X Condition interaction was also significant, ($F(4, 824) = 10.81$, $p < .001$, $\eta^2 = 057$. None of the other effects were significant.

Table 6. ANOVA results on the Glance Location. A = Age, V = Vehicle, C = Condition, and S = Session, G = Glance Location. Note that Glance Location sums to 100% for each of the conditions

	df_n	df_d	F	p	η^2
A	1	207	.148	.701	.001
V	1	207	.014	.840	.000
C	2	414	1.04	.354	.005
S	1	207	.903	.343	.004
G	2	414	1362	.000**	.868
AxC	2	414	.343	.785	.001
AxS	1	207	2.38	.124	.011
VxC	2	414	1.18	.309	.006
VxS	1	207	1.56	.214	.007
CxS	2	414	2.46	.087	.012
GxA	2	414	.303	.738	.001
GxV	2	414	1.31	.272	.006
GxC	4	828	12.5	.000**	.057
GxS	2	414	1.30	.273	.006

* $p < .05$, ** $p < .001$

A simplified 3 (Glance Location) X 3 (Condition) repeated measures ANOVA was conducted on the data presented in Figure 15. Both the main effect of Glance Location $F(2, 856) = 12617$, $p < .001$, $\eta^2 = .983$, and the Glance Location X Condition interaction were significant, $F(4, 856) = 52.9$, $p < .001$, $\eta^2 = .198$. Performing the voice tasks with the IVIS led to a reduction in the glance time to the mirrors and forward roadway with a corresponding increase in glance time to the dashboard displays. Similarly, performing the OSPAN task led to a reduction in the glance time to mirrors and dashboard displays with a corresponding increase in glance time to the forward roadway. Given that the primary task was to drive the vehicle and that the secondary tasks were primarily cognitive in nature, it is not surprising that drivers maintained their eyes on the forward roadway the majority of the time.

Practice Frequency

The frequency of practice was coded from the video recordings. On average, participants completed a total of 21.8 (SD = 19.3) voice-based tasks during the five days that they had the vehicle. As shown in Figure 16, the age of the participant did not affect the amount of practice with the IVIS voice systems. Participants gained the most practice with the music selection task, followed by the contact-calling task, then the number dialing task. The practice data were analyzed using a 3 (Age) x 4 (Practiced Item: Contact Call, Number Dial, Music Selection, Other) ANOVA. The main effect of Practiced Item was significant, $F(3,522) = 41.1$, $p < .001$, but neither the main effect of Age nor the Age X Practiced Item interaction were significant.

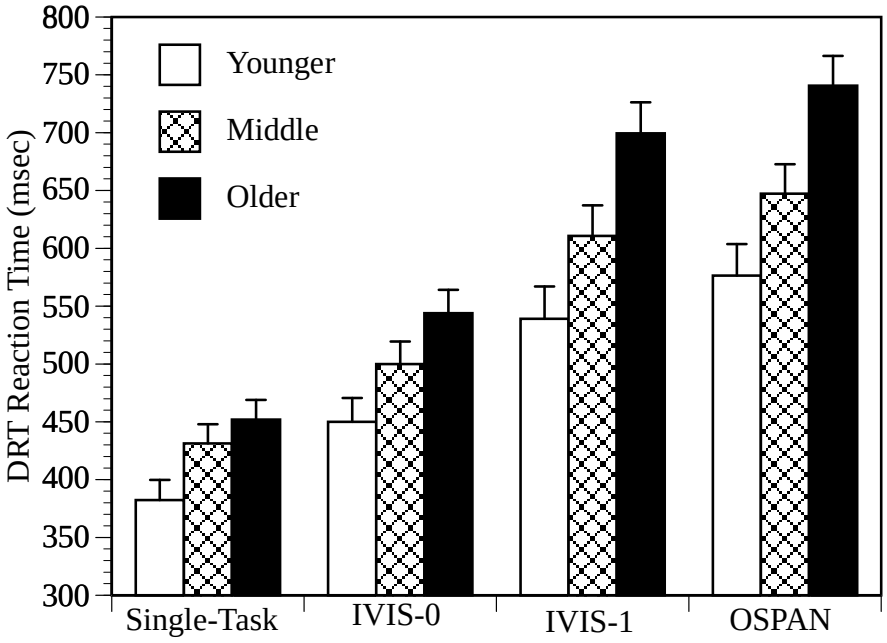

Figure 1. Mean DRT reaction time (in msec) for the single-task, IVIS-0 ("off-task"), IVIS-1 ("on-task"), and OSPAN conditions. The data are plotted for younger, middle, and older-age groups. Error bars reflect the 95% confidence interval around the point estimate.

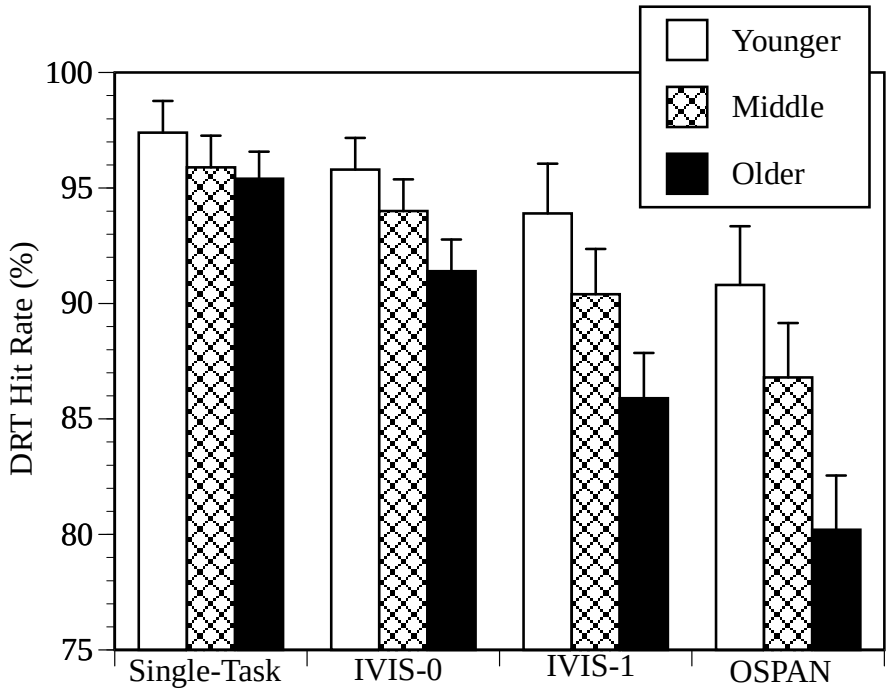

Figure 2. Mean DRT Hit Rate (an accuracy measure expressed as a percentage and computed by determining the number of valid responses divided by the total number of responses) for the single-task, IVIS-0 ("off-task"), IVIS-1 ("on-task"), and OSPAN conditions. The data are plotted for younger, middle, and older-age groups. Error bars reflect the 95% confidence interval around the point estimate.

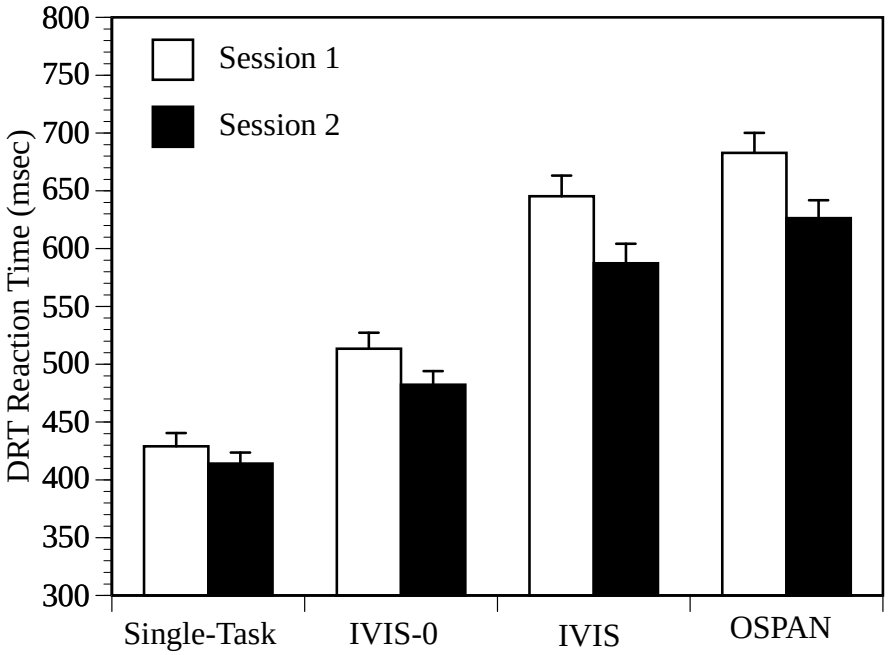

Figure 3. Mean DRT reaction time (in msec) for the single-task, IVIS-0 ("off-task"), IVIS-1 ("on-task"), and OSPAN conditions. The data are plotted for the first testing day (Session 1) and the last testing day (Session 2). Error bars reflect the 95% confidence interval around the point estimate.

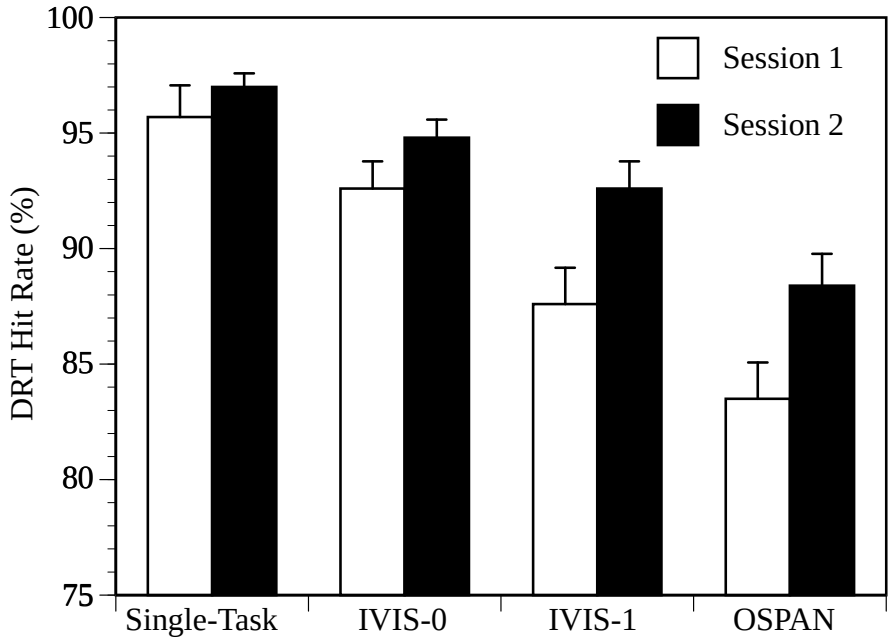

Figure 4. Mean DRT Hit Rate (an accuracy measure expressed as a percentage and computed by determining the number of valid responses divided by the total number of responses) for the single-task, IVIS-0 ("off-task"), IVIS-1 ("on-task"), and OSPAN conditions. The data are plotted for the first testing day (Session 1) and the last testing day (Session 2). Error bars reflect the 95% confidence interval around the point estimate.

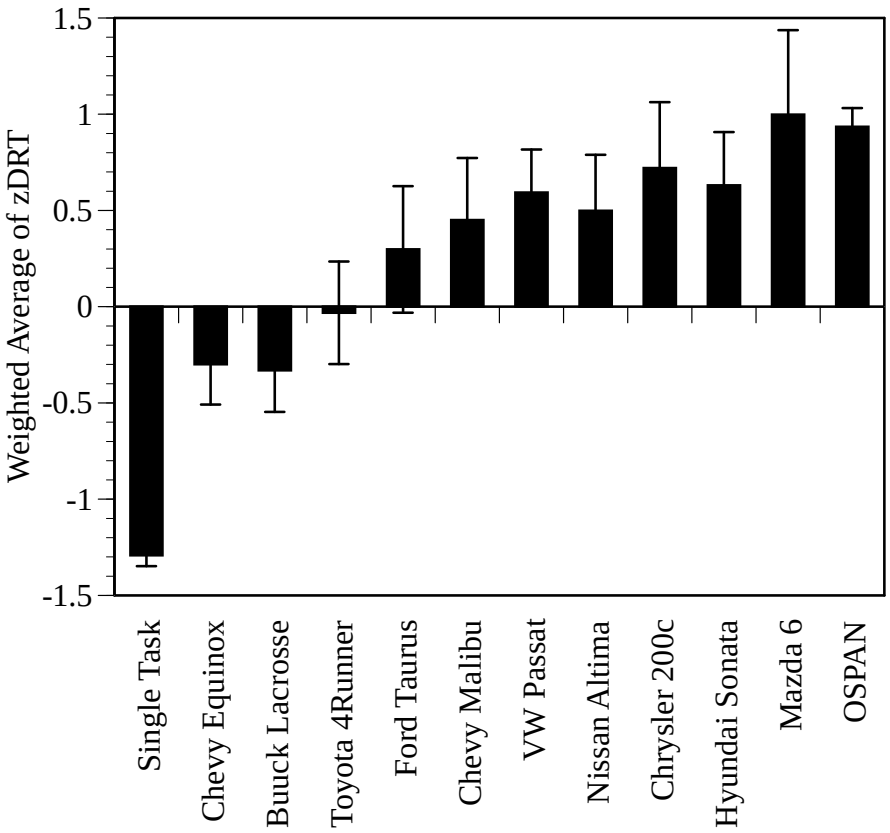

Figure 5. Weighted average of the z-transformed DRT data (i.e., DRT Reaction Time and DRT Hit Rate) plotted as a function of Vehicle in the IVIS condition. Error bars reflect the 95% confidence interval around the point estimate.

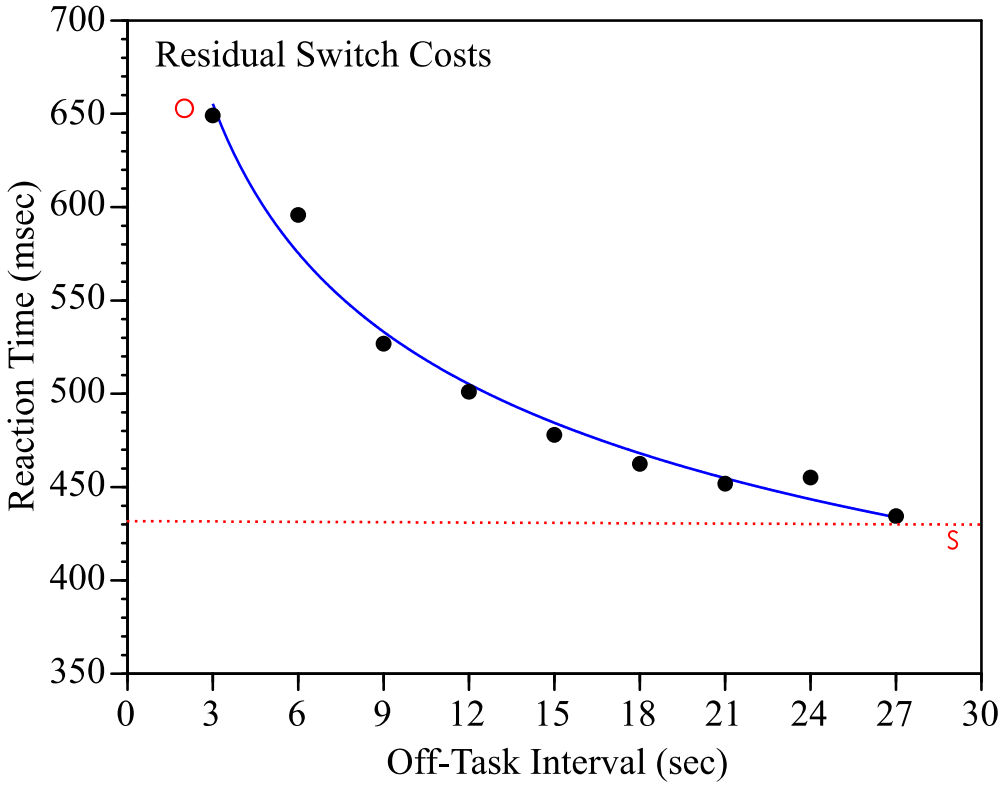

Figure 6. Residual switch costs in transitioning from on-task to off-task performance. The red "O" indicates average OSPAN RT from the DRT task. The red "S" indicates the average single-task RT from the DRT task. Off-task performance is distributed into 3-second intervals (relative to when the on-task activity terminated). The blue line represents the best fitting power function relating transition from on-task to single-task levels of performance. The dotted red line represents the critical t-value for significant differences from the single-task condition. Residual switch costs were significantly different from the single-task baseline up to 27 seconds after the on-task interval had terminated.

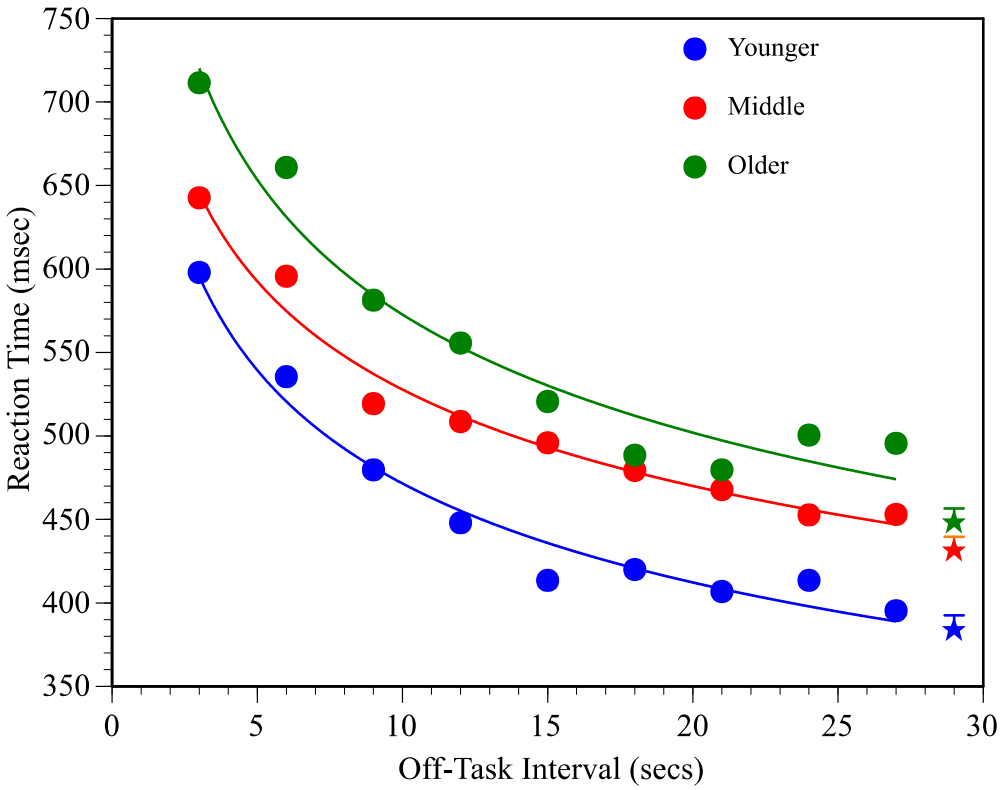

Figure 7. Residual switch costs for the three age groups in transitioning from on-task to off-task performance. The filled-stars indicate the average single-task RT for each group in the DRT task. Off-task performance is distributed into 3-second intervals (relative to when the on-task activity terminated). The solid lines represents the best fitting power function relating transition from on-task to single-task levels of performance for younger (blue lines), middle (red lines), and older- adults (green lines).

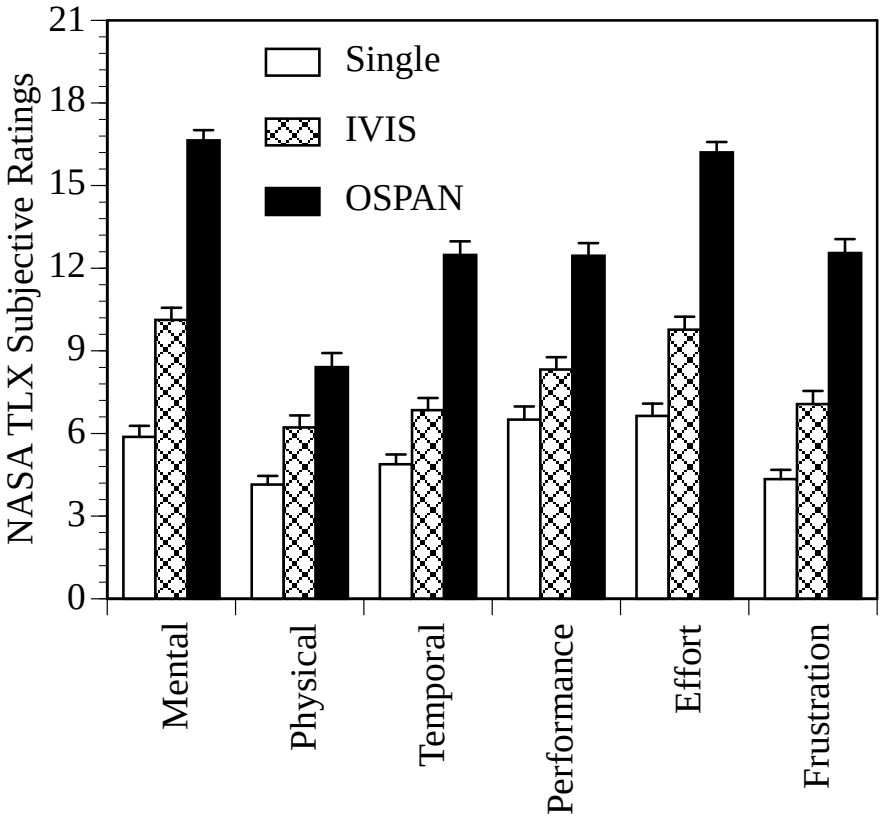

Figure 8. Mean NASA TLX ratings for the six sub-scales in the single-task, IVIS, and OSPAN conditions. Error bars reflect the 95% confidence interval around the point estimate.

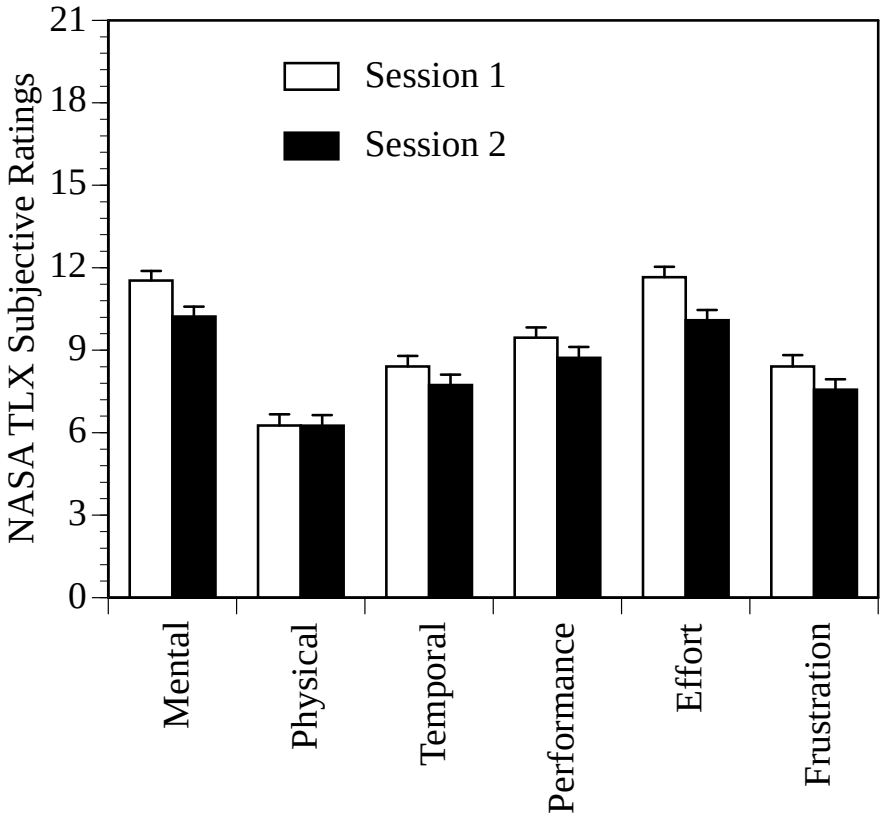

Figure 9. Mean NASA TLX ratings for the six sub-scales for the first testing day (Session 1) and the last testing day (Session 2). Error bars reflect the 95% confidence interval around the point estimate.

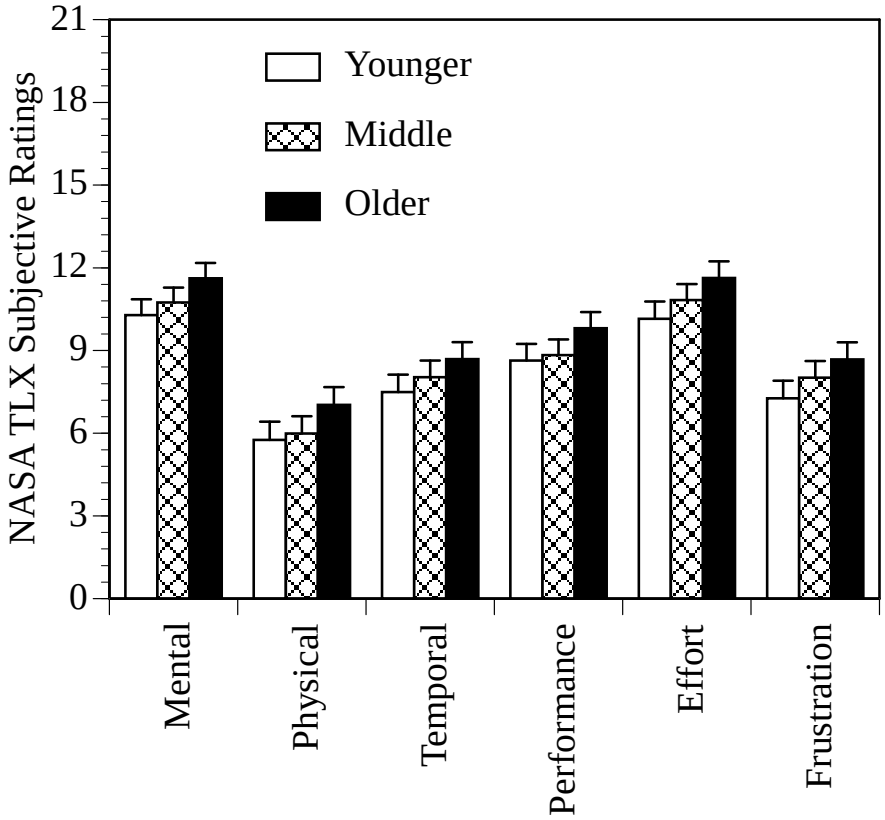

Figure 10. Mean NASA TLX ratings for the six sub-scales in the younger, middle, and older-age groups. Error bars reflect the 95% confidence interval around the point estimate.

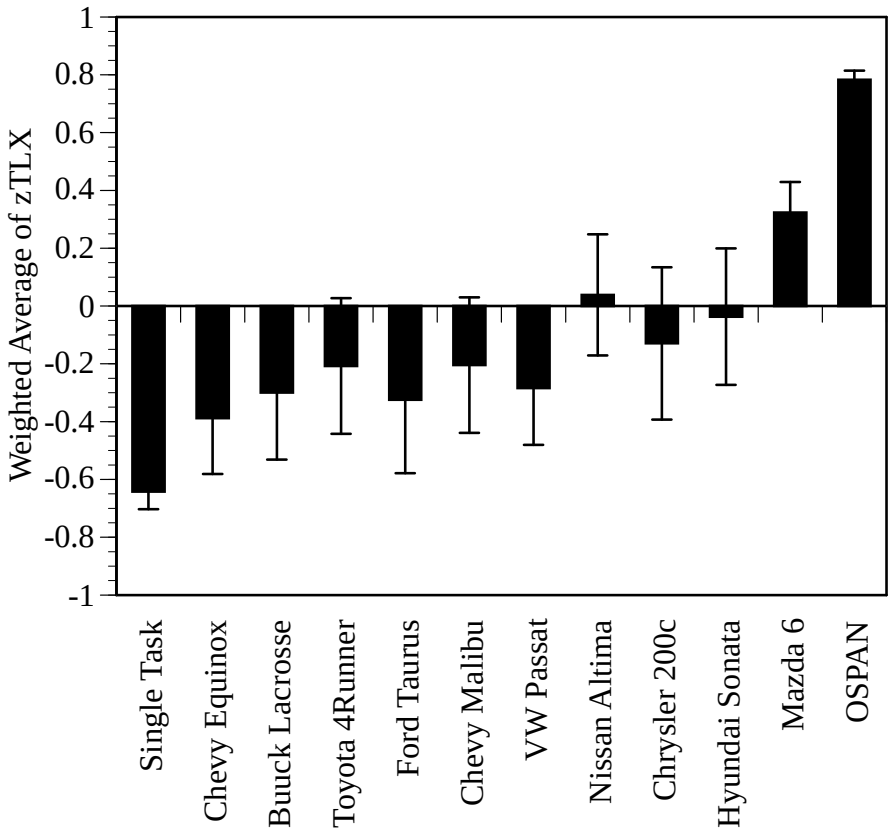

Figure 11. Weighted average of the z-transformed TLX data plotted as a function of Vehicle in the IVIS condition. Error bars reflect the 95% confidence interval around the point estimate.

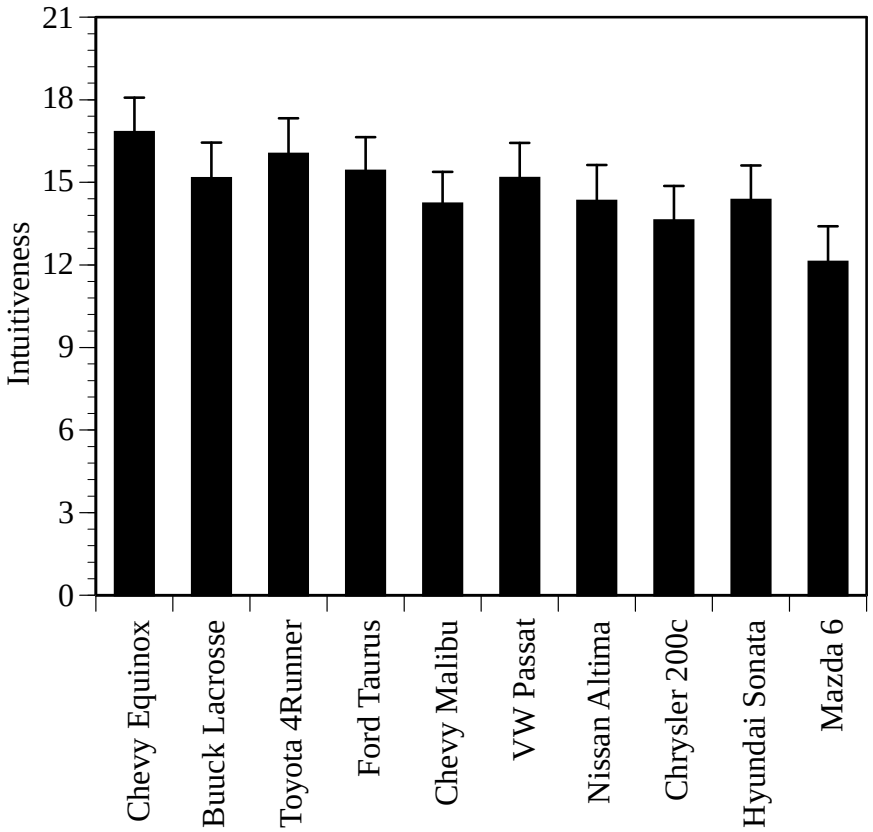

Figure 12. Mean ratings of intuitiveness (i.e., "how intuitive, usable, and easy was it to use the system") for the different IVIS systems on a 21-point scale where 1 reflected "not at all" and 21 reflected "very much." Error bars reflect the 95% confidence interval around the point estimate.

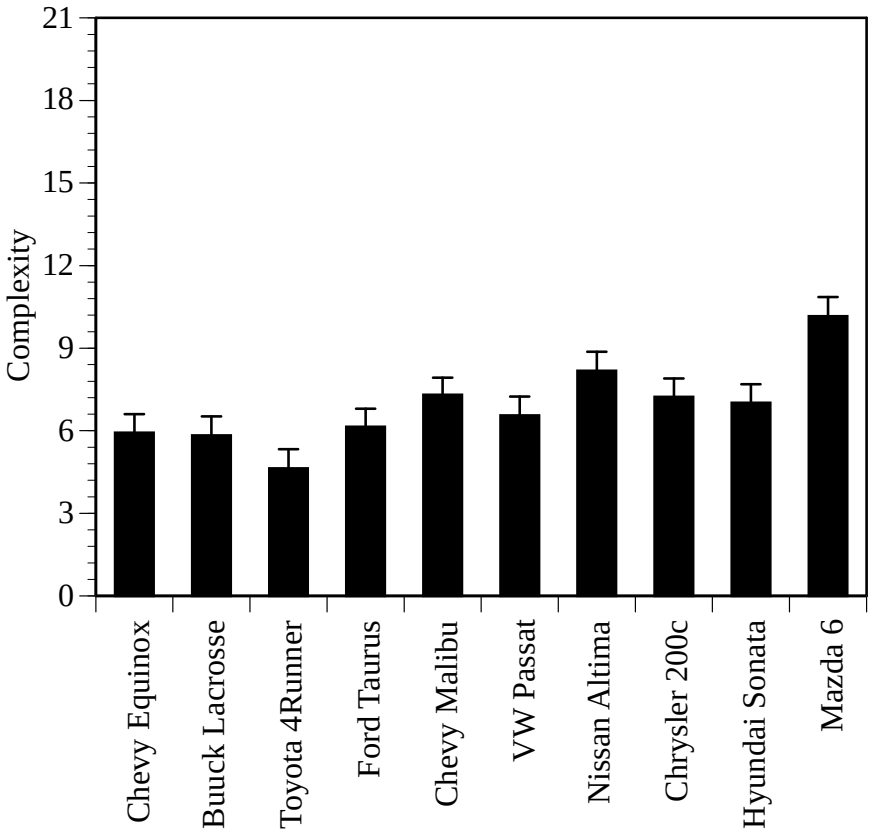

Figure 13. Mean ratings of complexity (i.e., "how complex, difficult, and confusing was it to use the system") for the different IVIS systems on a 21-point scale where 1 reflected "not at all" and 21 reflected "very much." Error bars reflect the 95% confidence interval around the point estimate.

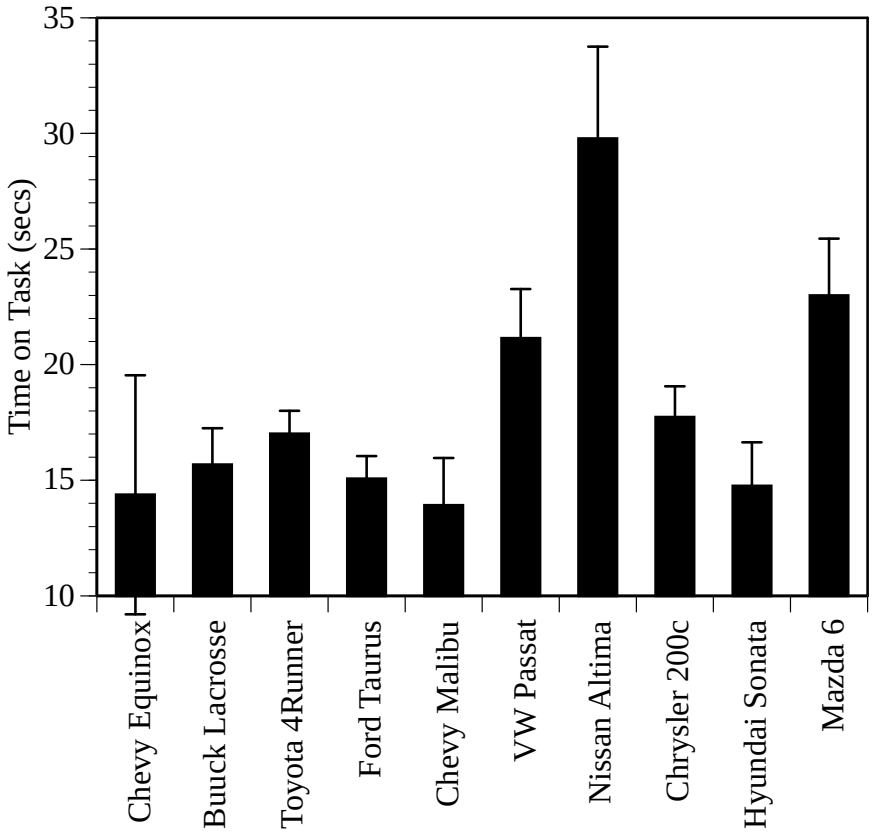

Figure 14. Mean time to complete the IVIS interactions for each vehicle. Error bars reflect the 95% confidence interval around the point estimate.

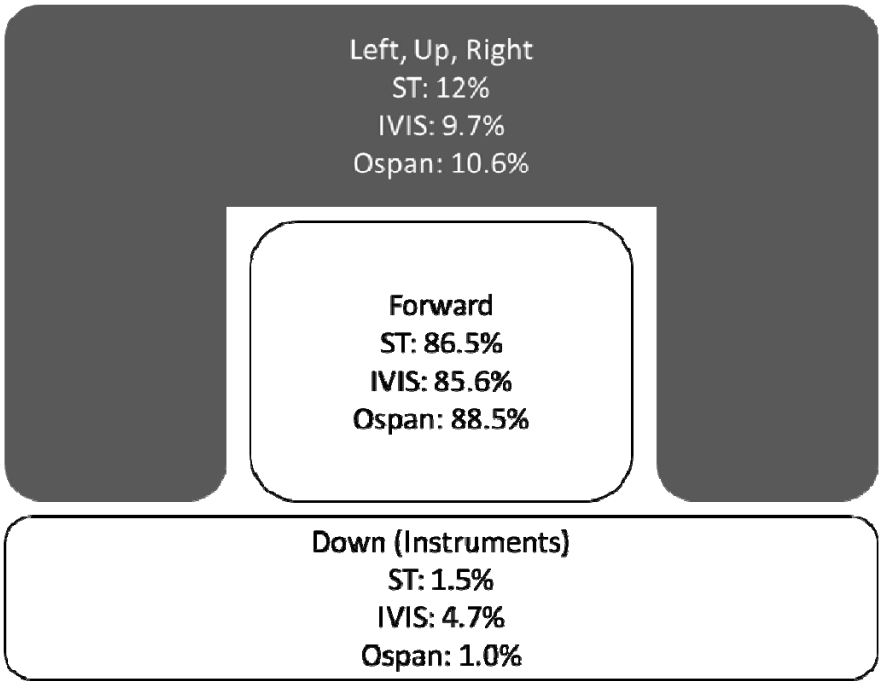

Figure 15. The distribution of glances to the forward roadway, instruments, and mirrors, broken down by Single Task (ST), IVIS, and OSPAN conditions.

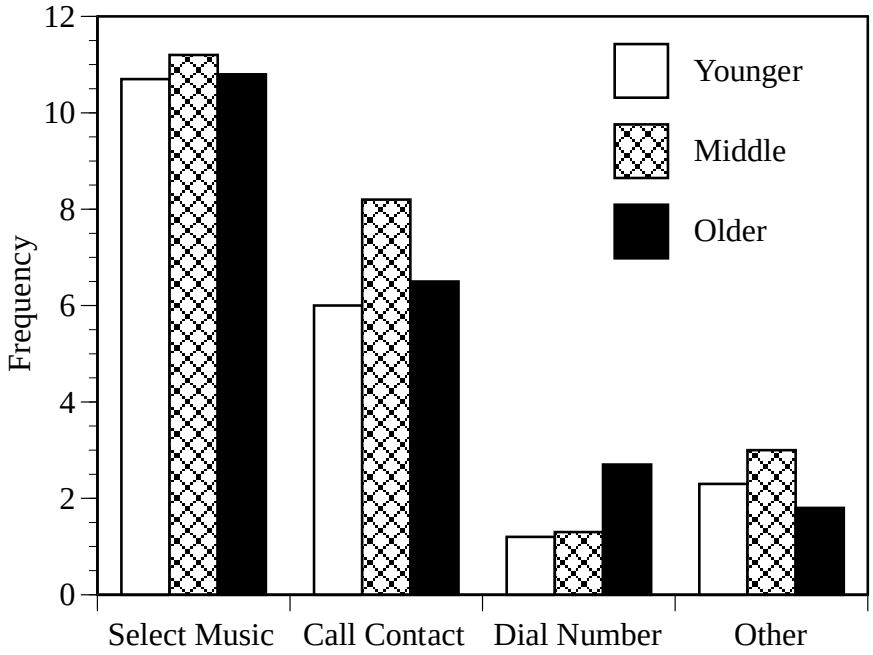

Figure 16. The mean number of interactions observed during the five days of practice. The data are plotted for younger, middle, and older-age groups.

The Cognitive Distraction Scale

A primary objective of the current research was to compare the cognitive workload associated with IVIS interactions in 10 different vehicles as drivers of different ages completed common IVIS voice-based tasks (e.g., voice dialing, music selection, etc.). Because the different dependent measures collected in this research were recorded on different scales, each was transformed to a standardized score. This involved Z-transforming the 2 DRT measures and the 6 NASA TLX measures to have a mean of 0 and a standard deviation of 1. The standardized scores were then weighted and summed to provide an aggregate measure of cognitive distraction. Weighting was equally assigned to the DRT and TLX such that each accounted for 50% of the collective rating. 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) 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 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 subjects.

Step 3: The standardized averages were computed with an equal weighting for secondary (i.e., DRT performance), and subjective (i.e., NASA TLX performance) metrics. The measures within each metric were also equally weighted. For example, the secondary task workload metric was comprised of an equal weighting of the measures DRT-RT and DRT-Hit Rate.

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

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

The cognitive workload scale for the different conditions is presented in Table 6 and Figure 17. By definition, the single-task condition had a rating of 1.0 and the OSPAN condition had a rating of 5.0. The rating for the different IVIS interactions varied considerably across vehicles, from a low rating of 2.37 to a high rating of 4.57. One method for determining how the systems compared is to compute the difference between the workload ratings for adjacent systems. For example, does the pair-wise comparison between the Chevy Equinox and Buick LaCrosse differ significantly? Under this method, the Chevy Equinox system had a significantly higher workload rating than the single-task condition, the Buick LaCrosse system did not differ from the Chevy Equinox system, the Toyota 4Runner system had a significantly higher workload rating than the Buick LaCrosse, the Ford Taurus system did not differ from the Toyota 4Runner, the Chevy Malibu had a significantly higher workload rating than the Ford Taurus, the VW Passat system did not differ from the Chevy Malibu, the Nissan Altima system did not significantly differ from the VW Passat, the Hyundai Sonata system did not differ from the Nissan Altima, the Chrysler 200c system did not differ from the Hyundai Sonata, and the Mazda 6 system had a significantly higher workload rating than the Chrysler 200c. Finally, the Mazda 6 system had a significantly lower workload rating than the OSPAN condition.

Table 6. The cognitive workload scale for the IVIS interactions.

Vehicle	Workload Rating	Std. Error
Single Task	1.00	0.09
Chevy Equinox	2.37	0.27
Buick Lacrosse	2.43	0.24
Toyota 4Runner	2.86	0.28
Ford Taurus	3.09	0.25
Chevy Malibu	3.39	0.27
VW Passat	3.46	0.28
Nissan Altima	3.71	0.28
Chrysler 200c	3.77	0.28
Hyundai Sonata	3.77	0.27
Mazda 6	4.57	0.27
OSPAN	5.00	0.09

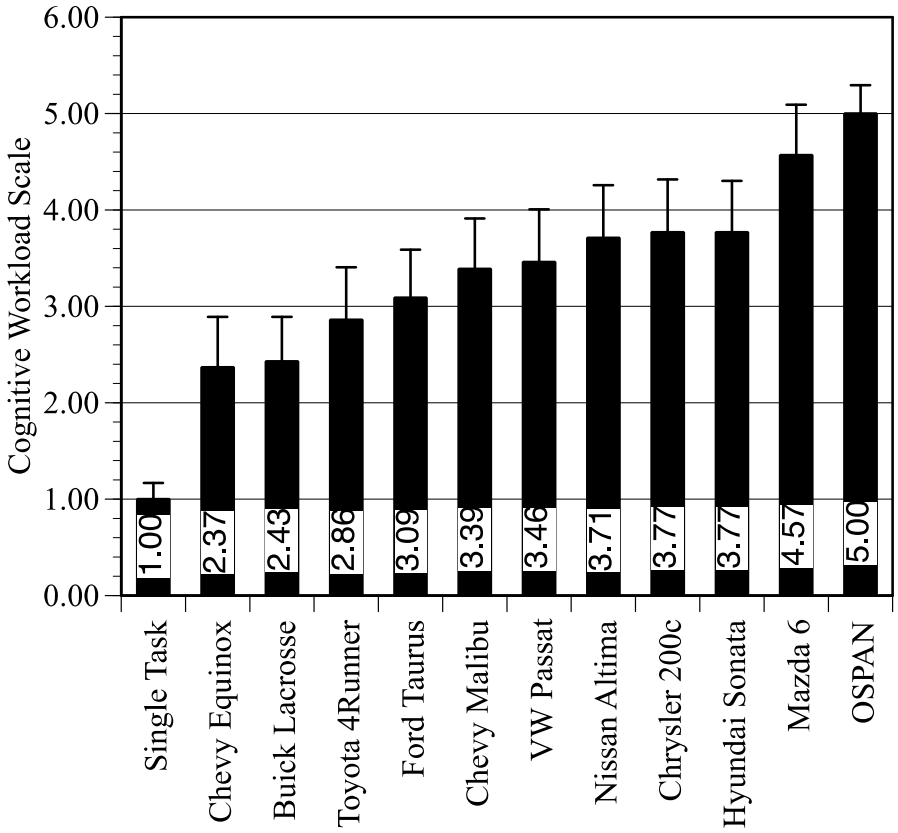

Figure 17. The cognitive workload scale for the IVIS interactions compared to single-task (category 1) and OSPAN (category 5). Error bars reflect the 95% confidence interval around the point estimate.

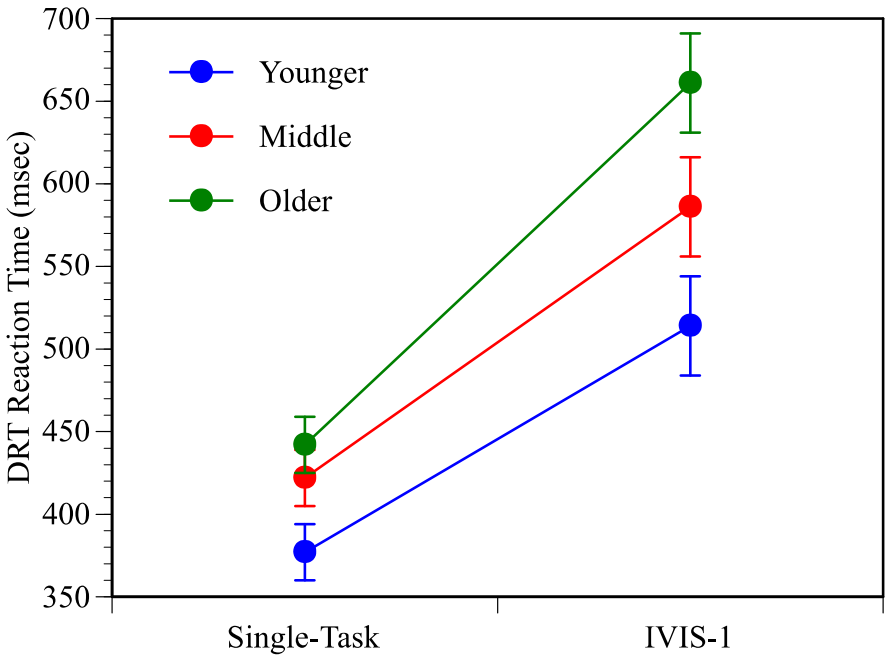

Figure 18. The DRT reaction time for single-task and IVIS-1 conditions after five days of practice. The data are plotted for younger, middle, and older-age groups. Error bars reflect the 95% confidence interval around the point estimate. This figure illustrates the classic Age-Complexity pattern, where age-related differences grow with task complexity. Moreover, it is clear that substantial costs are associated with the IVIS interactions after five of practice. Hence, older adults exhibit greater costs with the IVIS interactions and practice does not eliminate the costs (for any age group).

Discussion

The objective of the current research was to examine the impact of IVIS interactions on the cognitive workload experienced by drivers across the age range. We selected voice-based tasks that could be performed with no visual component and only a minimal button press to initiate the interaction. As such, they were primarily cognitive in nature (i.e., aside from the initial button press on the steering wheel, there was no requirement for visual or manual interaction). We explored several interrelated questions concerning the cognitive workload of these voice-based tasks. First, how demanding are these IVIS interactions? How do they compare to other common in-vehicle activities such as talking on a cell phone? Does the workload differ for the different vehicles? If they differ, what is the basis for the difference? Second, laboratory studies have found that older adults exhibit greater costs when multitasking. Do these age-related differences hold for real-world interactions while operating a motor vehicle? Third, does practice eliminate any age-related or vehicle-related differences in cognitive workload? If it does, how much practice is necessary? We address these issues in the following paragraphs.

First, using the IVIS to complete common tasks (e.g., voice dialing, contact calling, and radio tuning) was associated with a significant increase in the cognitive workload of the driver compared to the single-task condition. The overall workload ratings associated with IVIS interaction averaged 3.34 on our 5-point scale and ranged from 2.37 to 4.57; this reflects a moderate to a high level of cognitive workload. These cognitive workload ratings were associated with the intuitiveness and complexity of the IVIS and the time it took participants to complete the interaction. Systems that scored lower in cognitive workload were rated as being more intuitive, less complex, and it took participants a shorter time to complete the IVIS interactions. By contrast, systems that were higher in cognitive workload were rated as being less intuitive, more complex, and it took participants longer to complete the IVIS interactions. Importantly, our analyses were able to dissociate the differential workload associated with operating the vehicle (i.e., in the single-task condition) from the workload associated with IVIS interactions. We performed ANCOVAs that held constant single-task performance and found significant effects of the IVIS interaction. That is, the cognitive workload ratings are associated with the IVIS and not the operation of the vehicle.

Second, the cognitive workload experienced by older drivers performing these IVIS interactions was significantly greater than that experienced by younger drivers. This was revealed in the significant Condition X Age interactions, wherein performance differences between younger and older participants were amplified in the IVIS condition. For example, the age-related difference in RT in the single-task condition was 18.2%. This age-related difference grew to 29.7% in the on-task segments of the IVIS condition. The age-related difference in Hit Rates also grew from 2.1% in the single-task condition to 8.5% in the on-task segments of the IVIS condition.[4] This pattern was also found in a more fine-grained analysis that was restricted to the single-task condition and on-task segments of the IVIS (i.e., IVIS-1) after five days of practice (see Figure 18). In this targeted analysis, there

[4] These data rule out speed-accuracy tradeoffs as an explanation of the age-related differences in IVIS interaction. For both RT and accuracy measures, older adults performance was impaired to a greater extent than that of younger adults.

again was a Condition X Age interaction, $F(4, 454) = 7.35$, $p < .001$, $\eta^2 = .061$. The age-related difference in RT in the single-task condition was 17.2%. This age-related difference grew to 28.6% in the on-task IVIS condition. The age-related difference in Hit Rate also grew from 1.7% in the single-task condition to 11.3% in the IVIS condition. In essence, the age-related differences that were observed in the single-task condition doubled when participants interacted with the IVIS. Older adults also rated the IVIS interactions as being more complex. These findings are in line with the *Age-Complexity Hypothesis* (Cerella, 1985; Cerella, Poon, & Williams, 1980) that posits that age-related differences are amplified as the complexity of the task increases. The findings are important because drivers between the ages of 55 and 64 are the most frequent purchasers of new vehicles (Sivak, 2013). The voice-based systems found in many of these new vehicles are likely to induce high levels of cognitive workload for this cohort.

Third, practice improved performance for all conditions; however, the practice effects were greater as the task complexity increased. This was revealed in the Condition X Session interactions, where the effects of practice were more pronounced in the on-task IVIS condition than in the single-task condition. For example, RT decreased with five days of practice by 3.5% in the single-task condition and by 9.0% in the on-task segments of the IVIS condition. A similar comparison of Hit Rates found an increase with practice of 1.4% in the single-task condition and of 5.7% in the on-task IVIS condition. However, even after five days of practice, there were still large costs associated with IVIS interactions. A fine-grained analysis that focused on performance after five days of practice still found large differences between the single-task condition and on-task segments of the IVIS condition, $F(2, 226) = 336.17$, $p < .001$, $\eta^2 = .748$. Compared to the single-task condition, RT increased by 41.8% and Hit Rates decreased by 8.5% when participants performed IVIS interactions (cf. Figure 18).

Practice effects for all of human learning are known to be negatively accelerated (i.e., the *Power Law of Learning*), such that the biggest improvements occur early in training (Newell & Rosenberg, 1981; see also Heathcote, Brown, & Mewhort, 2000). This implies that any additional practice with IVIS interactions will have diminishing returns compared to what was observed after five days of practice. It appears that the impairments from using the IVIS cannot be practiced away. Moreover, neither the Condition X Session X Vehicle interactions (all p's .482), nor the Condition X Session X Age X Vehicle interactions (all p's .137) were significant. This is important because it indicates that the relative ordering of the IVIS systems was not altered with practice. IVIS interactions that were easy on the first day were also easy after five days of practice, and those IVIS interactions that were difficult on the first day were relatively difficult to perform after five days of practice.

Vehicle Differences

Our findings indicated that there were significant differences in the cognitive workload of the IVIS systems. The Chevy Equinox system had the lowest rating on the cognitive workload scale and the Mazda 6 system had the highest rating on the cognitive workload scale. Interestingly, the Chevy Equinox system rated highest (i.e., best) on intuitiveness, had one of the lowest ratings on complexity, and took one of the shortest time to complete (as measured by the time on task). By contrast, the Mazda 6 system rated the lowest on intuitiveness, highest on complexity, and had the second longest time to complete. This pattern is noteworthy because the intuitiveness, complexity, and time on task measures

were not included in the derivation of the cognitive workload scale. Nevertheless, they converge on the same interpretation of the driver's experience. A general principle that has emerged from this research is that robust, intuitive systems with lower levels of complexity and shorter task durations tend to have lower cognitive workload than more rigid, error-prone, time-consuming ones.

Our study evaluated the Chevy Malibu, an entry-level mid-sized sedan, and the Chevy Equinox, a compact sport utility vehicle. Both of these vehicles are manufactured by General Motors and were equipped with the MyLink system. We also evaluated the Buick LaCrosse, a luxury mid-size sedan, equipped with the Intellilink system, which is a rebranding of the MyLink system used by Buick. Panasonic manufactures both the MyLink and Intellilink systems and voice recognition software is produced by Nuance. Interestingly, the workload ratings for the Chevy Equinox (2.37) and the Buick LaCrosse (2.43) were virtually identical. However, the workload rating for the Chevy Malibu (3.39) was significantly higher than for the Chevy Equinox and Buick LaCrosse systems. An analysis of the ratings of intuitiveness and complexity for these three systems found that the Chevy Malibu system rated lower in intuitiveness than the Chevy Equinox and rated higher in complexity than the Chevy Equinox and Buick Lacrosse. During our testing, participants had more difficulty getting the Chevy Malibu voice-recognition system to understand their commands. This may stem, in part, from the ambient noise in the vehicle and the placement of the microphone.[5]

The analysis of workload using the on/off task DRT data found that "on-task" performance was associated with surprisingly high levels of workload (i.e., averaging 3.34 on our 5-point scale). The higher level of workload should serve as a caution that these voice-based interactions can be very mentally demanding and ought not to be used indiscriminately while operating a motor vehicle. Compared to our earlier research (Strayer et al., 2013), many of these IVIS interactions would appear to be significantly more demanding than typical cell phone conversations, which have cognitive workload levels around 2.3 on our 5-point scale. It is likely that the intuitiveness, complexity, and timing demands associated with the IVIS interactions are the reason for the increased level of cognitive workload.

Unexpected Costs

Interestingly, the off-task DRT performance provided evidence of persistent interference following the IVIS interactions. Despite the fact that the participants were not interacting with the system in any way, there were residual costs associated with the prior interaction. These residual costs are notable for their magnitude (in the seconds immediately following an interaction, the impairments are similar to that observed with OSPAN). These costs are also notable for their duration, lasting up to 27 seconds after an interaction had been completed. To put this in context, at 25 MPH a vehicle would have traveled 988 feet before the residual costs had completely dissipated. These findings have implications for self-regulatory strategies, such as choosing to dial or send a text message at a stoplight, because the costs of these interactions are likely to persist when the light turns green. The residual costs are likely related to the driver reestablishing situation awareness of the driving environment that was lost during the IVIS interaction (Fisher & Strayer, 2014; Strayer, in press).

[5] This information is based upon personal communication with representatives from Panasonic on November 20th 2015.

The voice-based interactions evaluated in the current study were designed to be completing using simple voice commands. However, like others (e.g., Reimer et al., 2014), we found that many participants routinely glanced at the displays during interactions. Additionally, we found that interactions with the voice-based systems changed the frequency of glances to the forward roadway and side and rear-view mirrors. Based on these findings, it is increasingly evident that natural visual scanning behavior is fundamentally coupled to cognitive processing demands. Quite simply, it is incorrect to assume that talking to your car is an "eyes-free" activity.

Caveats and Limitations

Cooper et al., (2014) also used the cognitive workload scale to benchmark the voice-based interactions of six vehicles. The workload ratings obtained in the current research are higher than those reported by Cooper et al., (2014). One reason for the difference in ratings stems from the way workload ratings were computed. In the current study, we use a refined approach to differentiate between "on-task" and "off-task" performance in the DRT measurements and excluded the "off-task" segments of the drive from the workload ratings. The Cooper et al., (2014) study did not have this ability and collapsed across the entire experimental segment for their workload estimates. Inclusion of the on- and off-task segments of the drive effectively collapses across momentary workload and time-on-task. By dissociating these factors, the current system provides a more fine-grained evaluation of workload. Overall-workload, be it cognitive, visual, or manual, is a function of momentary task demands *and* time-on-task (See Figures 14 and 17 respectively). Data from the current study suggest that these factors are sometimes, but not always, related. The independent measurement of these factors provides a more sophisticated method for evaluating driver workload.

Conclusion

The current research examined the impact of IVIS interactions on the cognitive workload experienced by drivers across the age range. The data supports six conclusions regarding the IVIS interactions while operating a motor vehicle.

- The momentary cognitive workload ratings associated with IVIS interaction averaged 3.34 on our 5-point scale and ranged from 2.37 to 4.57. These findings reflect a moderate to a high level of cognitive workload. The workload ratings were associated with the intuitiveness and complexity of the IVIS and the time it took participants to complete the interaction.
- The momentary cognitive workload experienced by older drivers performing the IVIS interactions was significantly greater than that experienced by younger drivers. In fact, the age-related differences that were observed in the single-task condition doubled when participants interacted with the IVIS.
- Practice does not eliminate the interference caused by IVIS interactions. IVIS interactions that were easy on the first day were also easy after five days of practice and those interactions that were difficult on the first day were still relatively difficult to perform after five days of practice.
- There were differences in the cognitive workload of the different IVIS systems over and above any differences associated with simply driving the vehicles. We found that robust, intuitive systems with lower levels of complexity and shorter task durations tend to have lower cognitive workload than more rigid, error-prone, time-consuming ones.
- There were long-lasting residual costs after IVIS interactions had terminated. These residual costs were notable for their magnitude and duration – in fact, it took 27 seconds to return to single-task baseline levels of performance. At 25 MPH, drivers would have traveled more than 3 football fields in this interval.

References

Bergen, B., Medeiros-Ward, N., Wheeler, K., Drews, F., & Strayer, D. L. (2013). The crosstalk hypothesis: Language interferes with driving because of modality-specific mental simulation. *Journal of Experimental Psychology: General, 142*, 119-130.

Carney, C., McGehee, D., Harland, K., Weiss, M., & Raby, M. (2015). Using naturalistic driving data to access the prevalence of environmental factors and driver behaviors in teen driver crashes. *AAA Foundation for Traffic Safety*.

Cerella, J. (1985). Information processing rates in the elderly. *Psychological Bulletin, 98*, 67-83.

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

Cooper, J. M., Ingebretsen, H., & Strayer, D. L. (2014). Measuring Cognitive Distraction in the Automobile IIa: Mental Demands of Voice-Based Vehicle Interactions with OEM Systems. *AAA Foundation for Traffic Safety*.

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.

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.

Hartley, A. A., & Little, D. M. (1999). Age-related differences and similarities in dual task interference. *Journal of Experimental Psychology: General, 128,* 416–449.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7*, 185-207.

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., & Larish, J. (1996). Aging and dual-task performance. In W. Rogers, A. D. Fisk, & N. Walker (Eds.), *Aging and skilled performance* (pp. 83–112). Hillsdale, NJ: Erlbaum.

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.

McDowd, J. M., & Shaw, R. J. (2000). Attention and aging: A functional perspective. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (2nd ed., pp. 221–292). Mahwah, NJ: Erlbaum.

Newell, A. & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55).

Hillsdale, NJ: Erlbaum.

NHTSA (2012). Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices. Department of Transportation. Docket No. NHTSA-2010-0053.

ISO DIS 17488 (2015). Road Vehicles -Transport information and control systems - Detection Response Task (DRT) for assessing selective attention in driving. Draft International Standard, ISO TC 22/SC39/WG8.

Pickrell, T. M. (2015, April). Driver electronic device use in 2013. (Traffic Safety Facts Research Note. Report No. DOT HS 812 114). Washington, DC: National Highway Traffic Safety Administration.

Regan, M. A., Hallett, C. & Gordon, C. P. (2011). Driver distraction and driver inattention: Definition, relationship and taxonomy. *Accident Analysis and Prevention*, *43*, 1771-1781.

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., Mehler, B., Dobres, J., McAnulty, H., Mehler, A., Munger, D., & Rumpold, A. (2014). Effects of an 'Expert Mode' Voice Command System on Task Performance, Glance Behavior & Driver Pysiology. *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicle Applications (AutoUI 2014),* Seattle, WA.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127-190.

Sivak, M. (2013). Marketing implications of the changing age composition of vehicle buyers in the U.S. Online publication downloaded on August 3, 2015 from at http://deepblue.lib.umich.edu/bitstream/handle/2027.42/97760/102946.pdf?sequence=1&isAllowed=y.

Strayer, D. L. (In Press). Attention and Driving. In J. Fawcett, E. F. Risko, & A. Kingstone (Eds.) *The Handbook of Attention*, pp. xxx-xxx, MIT Press.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., & Hopman, R. J. (In Press). The smartphone and the driver's cognitive workload: A comparison of Apple, Google, and Microsoft's intelligent personal assistants. *AAA Foundation for Traffic Safety.*

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

Strayer, D. L., Turrill, J., Coleman, J., Ortiz, E., & Cooper, J. M. (2014). Measuring Cognitive Distraction in the Automobile: II. Assessing In-vehicle Voice-based Interactive Technologies. *AAA Foundation for Traffic Safety.*

Watson, J. M., & Strayer, D. L. (2010). Supertaskers: Profiles in extraordinary multi-tasking ability. *Psychonomic Bulletin and Review*, *17*, 479-485.