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

The Smartphone and the Driver's Cognitive Workload:

A Comparison of Apple, Google, and Microsoft's Intelligent Personal Assistants

October 2015

AAA Foundation for Traffic Safety

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

Title

The Smartphone and the Driver's Cognitive Workload: A Comparison of Apple, Google, and Microsoft's Intelligent Personal Assistants. *(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

The goal of this research was to examine the impact of voice-based interactions using three different intelligent personal assistants (Apple's Siri, Google's Google Now for Android phones, and Microsoft's Cortana) on the cognitive workload of the driver. In two experiments using an instrumented vehicle on suburban roadways, we measured the cognitive workload of drivers when they used the voice-based features of each smartphone to place a call, select music, or send text messages. Cognitive workload was derived from primary task performance through video analysis, secondary task performance using the Detection Response Task, and subjective mental workload. We found that workload was significantly higher than that measured in the single-task drive. There were also systematic differences between the smartphones: The Google system placed lower cognitive demands on the driver than the Apple and Microsoft systems, which did not differ in this regard. Video analysis revealed that the difference in mental workload between the smartphones was associated with the number of system errors, the time to complete an action, and the complexity and intuitiveness of the devices. Finally, surprisingly high levels of cognitive workload were observed when drivers were interacting with the devices - "on-task" workload measures did not systematically differ from that associated with a mentally demanding memory/math OSPAN task. The analysis also found residual costs associated with using each of the smartphones that took significant time to dissipate. The data suggest that caution is warranted in the use of smartphone voice-based technology in the vehicle because of the high levels of cognitive workload associated with these interactions.

Introduction

Driver distraction, operationalized here 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. The U.S. Department of Transportation estimated that in 2013, 3,154 people were killed and an additional 424,000 were injured in motor vehicle crashes involving driver distraction on U.S. roadways (Pickrell, 2015); however, the report acknowledged limitations to the way the that the data were collected and suggested that the actual number is likely much higher. In support of this, a recent report by the AAA Foundation for Traffic Safety found that 58% of all crashes among teenage drivers could be attributed to driver inattention (Carney et al., 2015).

The National Highway Safety Traffic Administration (NHTSA) is in the process of developing voluntary guidelines to minimize driver distraction created by electronic devices in the vehicle. 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 2015, 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 task, 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. Yet, a large and growing body of literature cautions that auditory/vocal tasks may have unintended consequences that adversely affect traffic safety.

In 2013, we reported on a methodology for assessing cognitive distraction in the vehicle (Strayer et al., 2013). Converging measures of mental workload from primary and secondary task performance, physiological recordings, and self-reports, were used to develop a rating system for 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.

In 2014, we reported on an extension of our earlier methods designed to measure cognitive workload in six 2013 vehicles equipped with voice-based technology that facilitates tuning the radio and placing outgoing calls. We found striking differences in the workload ratings associated with the different systems, with the Toyota system having a workload rating roughly equivalent to listening to a book on tape and the Chevy system having one of the highest workload ratings we have observed for any in-vehicle task. Clearly, the user interface had a large impact on driver workload, frequency of errors, and time to complete the various tasks.

One alternative to using a vehicle's embedded voice controls for many common tasks is the smartphone. The advantage of these systems is that they are already commonly available,

they are constantly being updated, they are familiar to drivers, and they offer nearly limitless capabilities. In this report, we present the findings of two on-road driving experiments designed to measure the cognitive workload associated with interactions using three different intelligent personal assistants (Apple's Siri, Google's Google Now for Android phones, and Microsoft's Cortana) on the cognitive workload of the driver[1].

The selected tasks and experimental structure were designed to extend our prior work using embedded vehicle systems (Cooper et al., 2014). In the first experiment, we evaluated the cognitive demand of common voice interactions while driving. In the second experiment, we evaluated the cognitive demands associated with sending voice-based text messages. How do the different smartphone systems compare with each other and what are the bases for any observed differences in the cognitive workload experienced by the driver? How do these smartphone systems compare with embedded systems found in the different OEM systems?

[1] In our discussions with representatives from Google, they indicated that: *"the Google voice system that you are planning to test has never been promoted for in-vehicle use by Google. And though we understand that some users may engage in this type of activity, Google does not encourage this behavior."*

Methods

Experiment 1

Participants

Following approval from the Institutional Review Board, participants were recruited by word of mouth, advertisements placed on online local classified websites, and flyers posted on the University of Utah campus. They were compensated $60 upon completion of the 2.5-hour study. Data were collected from February 27th 2015 through April 14th of 2015.

Thirty-one subjects participated in Experiment 1 (16 males, 15 females). The youngest participant was 21 and the oldest was 68 years old, with an average age of 42. The Division of Risk Management Department at the University of Utah ran a Motor Vehicles Record 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 of Utah policy, each prospective participant was required to complete a 20-minute online defensive driving course and pass the certification test. Participants were selectively recruited to balance gender across the eligible age range. Everyone who participated in this research owned a smartphone and 64% reported using their phone regularly while driving. Participants reported between 5 and 52 years of driving experience with the average being 26 years. Additionally, participants reported driving an average of 200 miles per week over 8.5 hours. All participants were recruited from the greater Salt Lake area and spoke with a western US English dialect.

Design

A 5 (condition) x 3 (age groups) mixed within and between subjects design was used. The 5 within-subject conditions were: Single-task, Apple's *Siri*, Google's *Google Now*, Microsoft's *Cortana* and the OSPAN task. The 3 between-subject age groups were: ages 21-34, ages 35-53, and ages 54-70. Each participant experienced each of the five experimental conditions in a counterbalanced order. During interactions with the intelligent personal assistants, participants completed 2 number dialing tasks, 2 contact calling tasks, and 4 music selection tasks presented in 2 blocks.

In addition, some of the dependent measures used in the study allowed the differentiation of on-task and off-task performance during the three intelligent personal assistant conditions. For these analyses, an 8(condition) x 3(age group) design was used.

Materials and Equipment

Access to intelligent personal assistants engineered by Apple, Google, and Microsoft, was provided using an Apple iPhone 6 with iOS 8.2 (Build 12D508), a Google Nexus phone running Android 5.0.1 (Build LRby22C), and a Nokia Lumia 635 running Windows 8.1 (O.S. Version 8.10.12400.899), respectively. Identical music and contacts libraries were loaded on to each of the phones, providing the basis for the task evaluations.

An Apple "EarPods with Remote and Mic" was attached to each of the phones. The right speaker lead was inserted into participants' right ear and the left speaker lead was taped to the microphone input of the video collection system. A small button, attached to the cord of the headphones controlled the activation / deactivation of each of the three intelligent

personal assistants. This setup was selected because, at the time of testing, the single-ear system was legal in all 50 states. By using identical headphones we could ensure that any potential differences between the phones were related to characteristics of the verbal interface, and not potential differences in audio quality, microphone sensitivity, or other aspects of the physical interface.

Cellular phone service for all three systems was provided by T-Mobile. Excellent cell coverage (4-5 bars) was available during the entire drive on all phones. Phones were secured to the center console, just to the right of the steering wheel, using a universal suction mount that securely held each of the phones during interactions.

The vehicles used in the experiment were a 2015 Chevy Malibu with an automatic transmission and a 2015 Chrysler 200c with an automatic transmission.[2] Participants were familiarized with the vehicle and allowed to adjust the seat and mirrors before the study commenced. Participants drove the vehicle for approximately 20 minutes before the experiment began.

Two Sony Action Cams were used to collect video and audio feeds during experimentation. One was mounted to the front windscreen, just under the rear-view mirror, and faced the driver. The other was mounted between the two front seats via a rigid pole attached to the passenger seat headrest; it captured a view of the vehicle interior, including the screen of each phone, as well as the forward roadway. The two video feeds were synchronized for later video analysis.

During all phases of testing, participants wore a head-mounted Detection Response Task (DRT) device that was manufactured by Precision Driving Research. The DRT protocol 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-finger switch attached to the participant's left hand, 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 configuration used in this research adhered to the ISO standard 17488 with red LED stimuli configured to flash every 3-5 seconds. Timing was controlled on 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 the recall probe). In order to standardize presentation for

[2]A preliminary analysis found that the data collected in the Chevy Malibu and Chrysler 200c vehicles did not differ.

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, 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?*

A study facilitator was assigned to ride with each participant for the duration of the study. Facilitators were trained to precisely administer the research procedure and adhered to a scripted evaluation protocol. Additionally, facilitators were to ensure the safety of the driver, provide in-car training, and deliver task cues to participants. All facilitators had current driver's licenses and were over the age of 21.

Procedure

Upon arrival, participants filled out an IRB approved consent form and a brief intake questionnaire to assess basic characteristics of phone and driving usage. Once completed, drivers were familiarized with the controls of the instrumented vehicle, adjusted the mirrors and seat, and were informed of the tasks that would be completed while driving. Next, participants drove one circuit of the 2.7-mile 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. After the practice drive, participants began the experimental portions of the research. Data collection occurred between the hours of 9 a.m. and 8 p.m., Monday through Saturday. Driving conditions during data collection were sunny and clear with low traffic density.

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 task until they felt comfortable. In most cases, this took a couple of minutes.

Two baseline conditions and three experimental conditions were evaluated during the course of the research. The first condition was the single-task baseline. During the single-task condition, participants simply drove around the predefined driving course and responded to the lights generated by the DRT task. The second condition was a high workload condition in which participants drove while concurrently performing the OSPAN math and memory task. In each of the other 3 conditions, participants completed a series of common secondary tasks using either Apple's *Siri*, Android's *Google Now*, or Microsoft's *Cortana*.

Six distinct tasks were given to participants during each of the conditions involving interactions with the intelligent personal assistants. The tasks were initiated once

participants reached pre-specified locations on the course. Participants were not told where on the course the new tasks would be given but the task onset location remained constant for interactions with each of the voice assistants. Once tasks were completed, participants were allowed to return their undivided attention to the driving task until instructions were given for the subsequent task. Data were not collected during the turns. All tasks began when participants pressed the micro button located on the Apple EarPods to initialize the voice command systems. Once initiated, each of the tasks was completed through auditory + vocal system interactions. The tasks were presented to participants in a fixed order, progressing from Task 1 through Task 6 as participants circumnavigated the course. System interactions were as follows:

Task 1: "Phone: Joel Cooper"
Task 2: "Music: Fleetwood Mac" once completed… "Music: The Beatles"
Task 3: "Phone: Own Number"
Task 4: "Music: Stevie Wonder" once completed… "Music: Frank Sinatra"
Task 5: "Phone: Amy Smith at work"
Task 6: "Phone: Own number"

Prior to evaluations of the three intelligent personal assistants, each of the systems underwent a standard reset and voice model training procedure. This protocol was developed in conjunction with feedback from engineers working at Apple and Google. For the Apple iPhone, Siri and Siri dictation were reset for each participant prior to each run. In order to reset Siri, the following switch was toggled with each new participant: Settings -> General -> Siri -> Siri Off/On. In addition, dictation was reset for each participant by toggling the following: Settings -> Keyboard -> Enable Dictation -> Off/On. For the Android phone, the Google Now digital assistant was retrained prior to each drive through a simple voice training provided by Google, accessible through the following menu: Settings -> Language & Input -> Voice input -> Enhanced Google Services -> "Okay Google" Detection -> Retrain Voice Model. There was no voice training protocol for the Microsoft windows phone.

After each phone was ready for use, participants were allowed to explore the various functionalities of the voice assistant and were required to successfully retrieve the answer to 8 of the following 10 questions.

1. *What is the time in Sydney, Australia?*
2. *What is the tallest mountain in the world?*
3. *Who is the Speaker of the House in the United States?*
4. *What is the weather outside?*
5. *Where is the closest gas station?*
6. *When did we land on the moon?*
7. *What is 26 x 26?*
8. *What area code is 801?*
9. *What is 1 + 2 + 3 + 4?*
10. *What are the first 4 digits of pi?*

Once completed, participants were given a brief training on number dialing, contact calling, and music selection. Before each run, participants were then asked to complete a series of contact calling, number dialing, and music selection tasks until they reached proficiency.

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 restriction. Throughout each 10-minute condition, the driver completed the DRT. Safety directions were reiterated before each driving condition. At the conclusion of the study, participants returned to the University parking lot and they were compensated for their time and debriefed.

Dependent Measures

Cognitive workload was determined by collecting several dependent 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 this standard, all responses briefer than 100ms or greater than 2500ms were rejected for calculations of Reaction Time (RT). Responses that occurred later than 2.5 seconds from the stimulus onset were coded as misses. Any DRT data collected around turns was removed from the analysis. During testing, task engagement was flagged by the experimenter through a keyboard that facilitated comparison of performance in the secondary-task smart phone conditions when the participant was actively engaged in an activity (on-task) or had finished that an activity and was operating the vehicle without secondary-task interaction (off-task).

- DRT –Reaction Time (both on-task and off-task). Defined as the sum of all valid reaction times to the DRT task divided by the number of valid reaction times.
- DRT – Hit Rate (both on-task and off-task). Defined as the number of valid responses divided by the total number of stimuli presented during each condition.

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 task, the final 2 were questions added to assess the intuitiveness and complexity of the tasks.

- 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 2 questions on task intuitiveness and complexity.

Three critical performance metrics were distilled from coding the video recorded during testing. These were time to complete the task, error count, and average driving speed. The 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. Task completion time reflects the average task duration across the 6 tasks

- Video Analysis – Vehicle Speed. Average driving speed was derived from the time required to traverse the northern and southern legs of the route. During video coding, the time that corresponded to the start and end sections of roadway was recorded. The total distance of these two roadway sections was 2.39 miles.
- Video Analysis – Error count. Defined as the total number of system errors that arose during the 6 tasks. System behaviors classified as errors were: Instances when the

system was unresponsive to the users intention (e.g., not carrying out any action at all or indicating that the user should try again); instances where the system understood what the participant said but carried out an action that was inconsistent with the participants expectations (e.g., searching the internet for Stevie Wonder rather than playing a music selection by Stevie Wonder); instances where the system failed to correctly understand the words spoken by the user (e.g., "calling *Jane* Doe" instead of "calling *John* Doe"); and instances where the system entered an error state due to a pacing error by the participant (e.g., speaking prior to the tonal listening cue).

- Video Analysis – Task Completion Time. Defined as the time from the moment the voice activation button on the headphones was pressed to initiate a task to the time the button was pressed to terminate a task.

To assess the overall performance of each of the three intelligent personal assistants (Siri, Google Now, and Cortana), the three classes of voice tasks completed during this experiment (number dialing, contact calling, and music selection) were aggregated. Thus, workload measures presented in this report are a general reflection of overall system performance and are not specifically indicative of performance on any one of the tasks.

Experiment 2

In Experiment 1, we tested a variety of voice-based interactions that are common in many OEM vehicles (e.g., Cooper et al., 2014). However, smartphones have additional voice-based capabilities that go beyond dialing and music selection. In Experiment 2 we tested the voice-texting features of these phones to determine how these seemingly more complex interactions would affect the driver's performance while operating a motor vehicle. We kept the testing protocol identical to that used Experiment 1, with the exception that the dialing and music selection tasks were replaced with sending short text messages.

Participants

Thirty-four subjects participated in Experiment 2 (19 males, 15 females). Participants were recruited using the same methods as Experiment 1. All data were collected from March 26th 2015 through April 19th of 2015. The youngest participant was 22 and the oldest was 68 years old, with an average age of 42.5. All eligibility requirements were identical to those used in Experiment 1. Participants reported between 4 and 52 years of driving experience with the average being 26.8 years. Additionally, participants reported driving an average of 207 miles per week over 9.3 hours. All participants were recruited from the greater Salt Lake area and spoke with a western US English dialect.

Materials and Equipment

The equipment used in Experiment 2 was identical to that used in Experiment 1.

Procedure

The procedure for Experiment 2 was identical to that used in Experiment 1 with the exception that participants dictated unique text messages at each of the 6 task locations throughout the driving course. In order to adequately train participants on the text message functionality of each of the phones, they were required to send 6 practice text messages prior to using the phone's digital assistant that was to be used in the forthcoming

condition. Voice training and resetting for each of the phones was identical to that used in Experiment 1.

Once trained, participants were reminded of the upcoming task and asked if they had any questions. Text messaging prompts were given in the same location as the task prompts in Experiment 1 and were:

1. *"Tell Amy Smith that you saw her flight is early, but you're on your way now."*
2. *"Tell John Doe you're running late in traffic, & ask him to start the meeting without you."*
3. *"Tell Anna Pearl your car is in the shop, & can she come pick you up."*
4. *"Ask Chris Hunter if he wants to eat out & what movie he wants to watch tonight."*
5. *"Tell Amy Smith you're running late. Ask her to start dinner."*
6. *"Tell John Doe you picked up lunch & you're on your way to the meeting."*

In all cases, every effort was made to keep the experimental procedure between Experiment 1 and Experiment 2 as identical as possible.

Results

Experiment 1

DRT

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

Reaction Time

The reaction time data from the DRT when participants were on-task were analyzed using a MANOVA with a 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 8 (Condition: Single-task, Apple-0, Google-0, Microsoft-0, Apple-1, Google-1, Microsoft-1, and OSPAN) split-plot factorial design. RT increased with Condition, $F(7, 196) = 29.83$, $p < .001$, $\eta^2 = .516$, but neither Age, $F(2, 28) = 1.76$, $p = .190$, $\eta^2 = .112$, nor the Age by Condition interaction, $F(14, 196) = 1.08$, $p = .375$, $\eta^2 = .072$, were significant. Planned comparisons indicated that the single-task condition was significantly faster than the other secondary-task conditions ($p < .001$), that the Google-0 condition was faster than both the Apple-0 condition ($p = .023$) and Micorsoft-0 condition (p.021), that the Apple-0 and Microsoft-0 conditions did not significantly differ ($p = .630$), and that each of these conditions differed from their respective on-task performance (Apple-0 vs. Apple-1, $p < .001$; Google-0 vs. Google-1, $p < .001$; Microsoft-0 vs. Microsoft-1, $p < .001$). Importantly, neither the Apple-1, nor the Microsoft-1 conditions significantly differed from the OSPAN condition ($p = .061$ and $p = .130$), whereas the Google-1 condition was significantly faster than OSPAN ($p = .003$). Finally, the on-task performance for the three conditions did not differ from each other, (Apple-1 vs. Google-1, $p = .527$; Apple-1 vs. Microsoft-1, $p = .426$; Google-1 vs. Microsoft-1, $p = .153$).

Hit Rate

The Hit Rate data from the DRT task were analyzed using a MANOVA with a 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 8 (Condition: Single-task, Apple-0, Google-0, Microsoft-0, Apple-1, Google-1, Microsoft-1, and OSPAN) split-plot factorial design. Hit Rate decreased with Condition, $F(7, 196) = 11.30$, $p < .001$, $\eta^2 = .287$, but neither Age, $F(2, 28) = 0.11$ $p = .895$, $\eta^2 = .008$, nor the Age by Condition interaction, $F(14, 196) = 1.27$, $p = .227$, $\eta^2 = .083$, were significant. Planned comparisons indicated that Hit Rate was significantly higher in the single-task condition than the other secondary-task conditions ($p < .001$) with the exception of the single-task vs. Google-0 comparison, which did not significantly differ ($p = .599$). Hit Rate was higher in the Google-0 condition than the Apple-0 ($p = .006$) and Micorsoft-0 (p.017) conditions, and the Apple-0 and Microsoft-0 conditions did not significantly differ ($p = .815$). The off-task performance differed from on-

task performance for Google-0 vs. Google-1, $p < .006$, and Microsoft-0 vs. Microsoft-1, $p < .013$, but not for Apple-0 vs. Apple-1 ($p = .057$). Hit Rate was higher for each of the on-task secondary-task conditions than OSPAN ($p = .032$, $p = .002$, and $p = .021$ for Apple-1, Google-1, and Microsoft-1, respectively). Finally, the on-task performance for the three secondary-task conditions did not differ from each other (Apple-1 vs. Google-1, $p = .051$; Apple-1 vs. Microsoft-1, $p = .851$; Google-1 vs. Microsoft-1, $p = .058$).

NASA TLX

The 6 scales of the NASA TLX, presented in Figure 3, were analyzed using a MANOVA with a 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 5 (Condition: Single-task, Apple, Google, Microsoft and OSPAN) split-plot factorial design. The MANOVA revealed a main effect of Condition, $F(24, 440) = 5.56$, $p < .001$, $\eta^2 = 233$, but neither Age, $F(12,48) = 0.82$, $p = ..627$, $\eta^2 = .171$, nor the interaction were significant, $F(48, 672) = 0.84$, $p = .767$, $\eta^2 = .057$.

Univariate tests were also performed on the 6 NASA TLX subscales. The *mental* sub-scale increased as a function of Condition, $F(4, 112) = 50.58$, $p < .001$, $\eta^2 = .644$, and Age, $F(2, 28) = 4.11$, $p = .027$, $\eta^2 = .227$, but the interaction was not significant, $F(8, 112) = 0.65$, $p = .734$, $\eta^2 = .044$. The *physical* sub-scale increased as a function of Condition, $F(4, 112) = 9.80$, $p < .001$, $\eta^2 = .259$, but neither the Age, $F(2, 28) = 0.33$, $p = .719$, $\eta^2 = .023$, nor the interaction were significant, $F(8, 112) = 0.67$, $p = .713$, $\eta^2 = .046$. The *temporal* sub-scale increased as a function of Condition, $F(4, 112) = 33.99$, $p < .001$, $\eta^2 = .548$, but neither the Age, $F(2, 28) = 2.36$, $p = .113$, $\eta^2 = .114$, nor the interaction were significant, $F(8, 112) = 0.63$, $p = .747$, $\eta^2 = .043$. The *performance* sub-scale increased as a function of Condition, $F(4, 112) = 5.55$, $p < .001$, $\eta^2 = .165$, but neither the Age, $F(2, 28) = 0.43$, $p = .657$, $\eta^2 = .030$, nor the interaction were significant, $F(8, 112) = 0.81$, $p = .598$, $\eta^2 = .054$. The *effort* sub-scale increased as a function of Condition, $F(4, 112) = 29.79$, $p < .001$, $\eta^2 = .516$, but neither the Age, $F(2, 28) = 2.06$, $p = .146$, $\eta^2 = .129$, nor the interaction were significant, $F(8, 112) = 0.81$, $p = .597$, $\eta^2 = .055$. Finally, the *frustration* sub-scale increased as a function of Condition, $F(4, 112) = 21.02$, $p < .001$, $\eta^2 = .429$ but neither the Age, $F(2, 28) = 1.31$, $p = .285$, $\eta^2 = .014$, nor the interaction were significant, $F(8, 112) = 0.45$, $p = .889$, $\eta^2 = .031$.

Intuitiveness and Complexity

Participants were also asked to rate how intuitive, usable, and easy it was to use the different smartphones. They also rated how complex, difficult, and confusing it was to use the different smartphones. Figure 4 presents the intuitiveness and complexity ratings on a 21-point scale where 1 reflected "not at all" and 21 reflected "very much."

Intuitiveness

A 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 3 (Condition: Apple, Google, Microsoft) split-plot MANOVA found that intuitiveness varied as a function of Condition, $F(2, 56) = 5.66$, $p = .006$, $\eta^2 = .168$, but not Age, $F(2, 28) = 1.64$, $p = .212$, $\eta^2 = .105$; however, the Age by Condition was significant, $F(4, 56) = 2.98$, $p = .026$, $\eta^2 = .176$. Planned comparisons revealed that the intuitiveness of the Apple and Google systems did not differ ($p = .244$), and both were rated as more intuitive than the Microsoft system (Apple vs. Microsoft, $p = .009$; Google vs. Microsoft, $p = .036$).

Complexity

A 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 3 (Condition: Apple, Google, Microsoft) split-plot MANOVA found that complexity varied as a function of Condition, $F(2, 56) = 9.83$, $p = .006$, $\eta^2 = .168$, but neither the Age, $F(2, 28) = 1.06$, $p = .360$, $\eta^2 = .070$, nor the Age by Condition interaction were significant, $F(4, 56) = 0.61$, $p = .660$, $\eta^2 = .042$. Planned comparisons revealed that the complexity of the Apple and Google systems did not differ ($p = .772$), and both were rated as less complex than the Microsoft system (Apple vs. Microsoft, $p = .002$; Google vs. Microsoft, $p = .001$).

Video Analysis of Interactions

An analysis of the video of the participant's interactions was performed to determine the vehicle speed, presented in Figure 5, the number of observed errors, presented in Figure 6 and the time to complete the task, presented in Figure 7. The relative frequency of the four error categories for each of the smartphones is provided in Figure 8.

Vehicle Speed

A 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 3 (Condition: Apple, Google, Microsoft) split-plot MANOVA found that vehicle speed varied as a function of Condition, $F(4, 112) = 4.87$, $p < .001$, $\eta^2 = .148$, but not Age, $F(2, 28) = 1.43$, $p = .256$, $\eta^2 = .093$. The Age by Condition interaction was also significant, $F(8, 112) = 2.89$ $p = .006$, $\eta^2 = .171$. Planned comparisons revealed that the driving speed was higher in the single-task condition than in all other conditions ($p = .006$, $p < .001$, $p < .001$, and $p = .013$, respectively) and that speed did not differ from OSPAN for the Apple ($p = .222$), and Google ($p = .508$) conditions, but the Microsoft condition was significantly faster than OSPAN ($p = .041$). Vehicle speed did not significantly differ between the smartphone conditions (Apple vs. Google; $p = .737$; Apple vs. Microsoft, $p = .379$; and Google vs. Microsoft, $p = .508$).

Error Count

A 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 3 (Condition: Apple, Google, Microsoft) split-plot MANOVA found that the number of errors differed as a function of Condition, $F(2, 56) = 3.94$, $p = .025$, $\eta^2 = .123$, Age, $F(2, 28) = 4.56$, $p = .020$, $\eta^2 = .245$, but the Age by Condition interaction was not significant, $F(4, 56) = 0.84$ $p = .504$, $\eta^2 = .057$. Planned comparisons revealed that the number of errors did not differ between the Apple and Google ($p = .508$) or Apple and Microsoft ($p = .101$), but the difference between the Google and Microsoft was significant ($p = .041$).

Task Completion Time

A 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 3 (Condition: Apple, Google, Microsoft) split-plot MANOVA found that the time to complete the task did not differ as a function of Condition, $F(2, 56) = 1.80$, $p = .174$, $\eta^2 = .060$, Age, $F(2, 28) = 1.46$, $p = .249$, $\eta^2 = .095$, and the Age by Condition interaction was also not significant, $F(4, 56) = 1.69$ $p = .166$, $\eta^2 = .108$. None of the pair-wise planned comparisons was significant (Apple vs. Google, $p = .508$; or Apple vs. Microsoft, $p = .101$; and Google vs. Microsoft, $p = .576$).

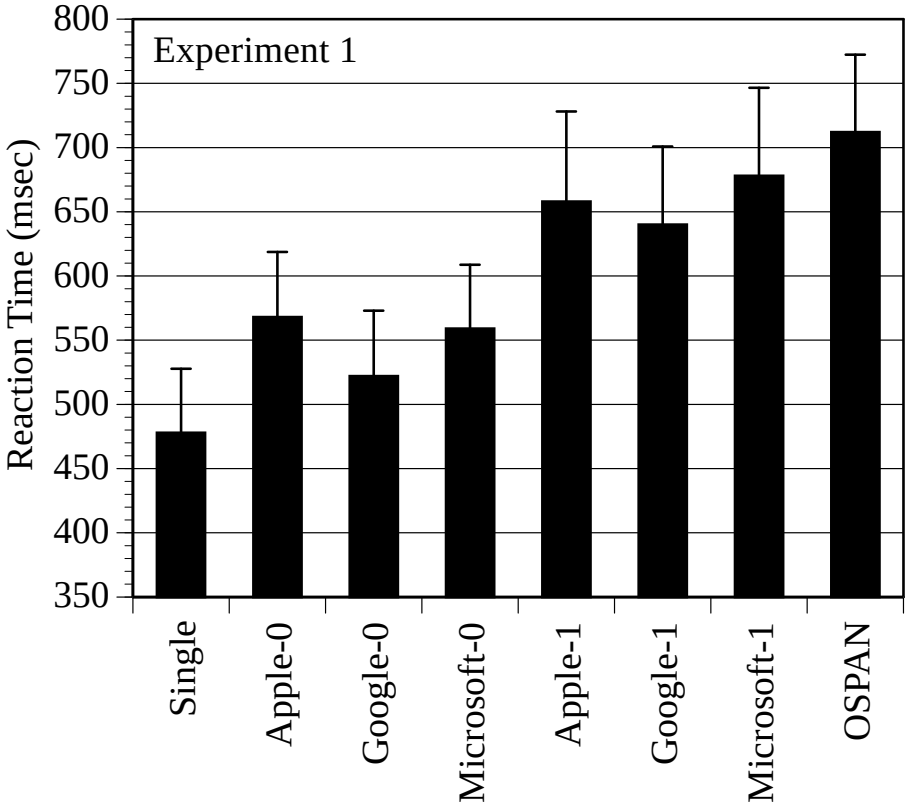

Figure 1. Mean DRT reaction time (in msec) for the single-task, OSPAN, and off-task (e.g., Google-0) and on-task (e.g., Google-1) performance for the Apple, Google, and Microsoft secondary tasks in Experiment 1. Error bars reflect 95% confidence intervals around the point estimate.

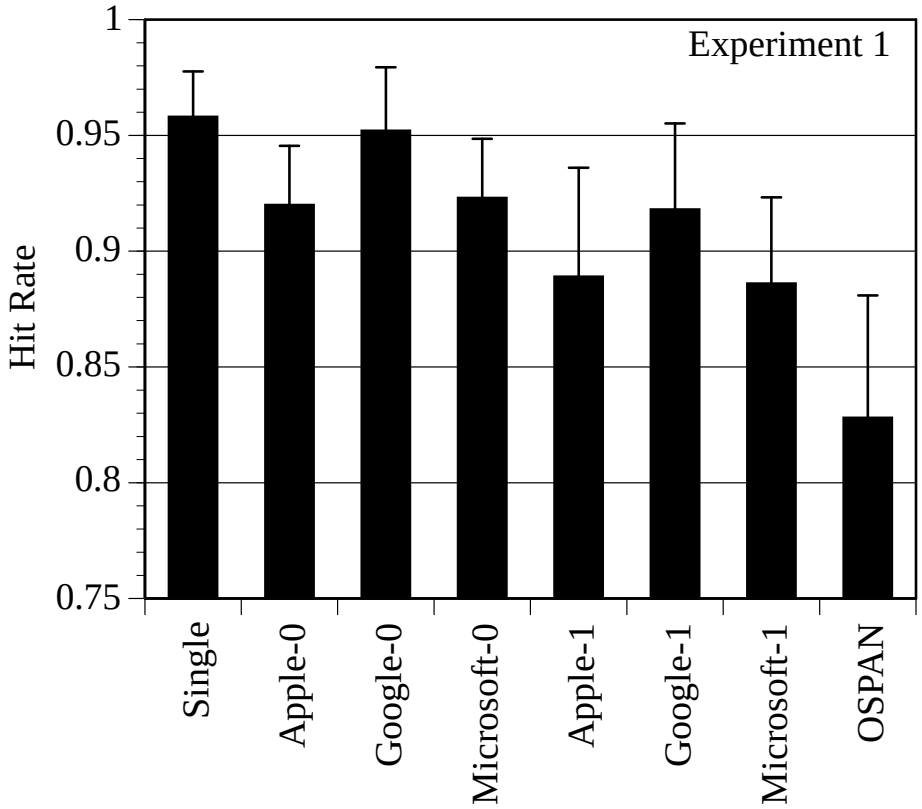

Figure 2. Mean DRT Hit Rate (an accuracy measure computed by determining the number of valid responses divided by the total number of responses for the single-task, OSPAN, and off-task (e.g., Google-0) and on-task (e.g., Google-1) performance for the Apple, Google, and Microsoft secondary tasks in Experiment 1. Error bars reflect 95% confidence intervals around the point estimate.

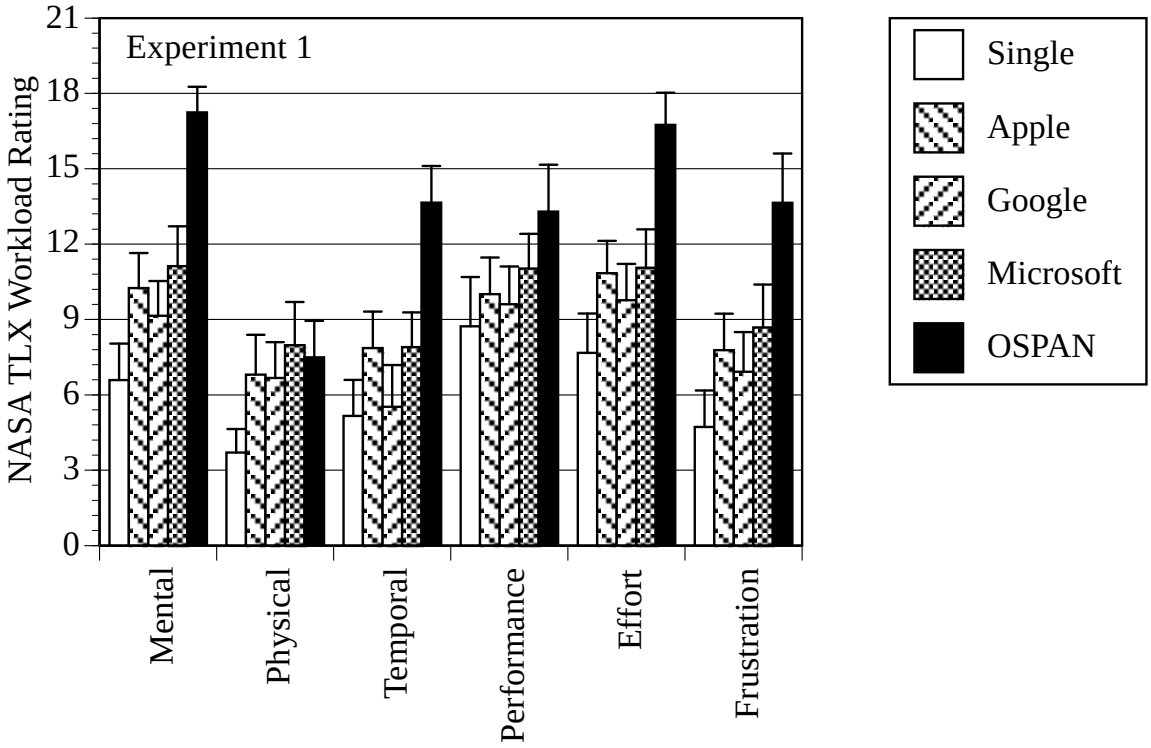

Figure 3. Mean NASA TLX ratings for the six sub-scales in the 5 conditions of Experiment 1. Error bars reflect 95% confidence intervals around the point estimate.

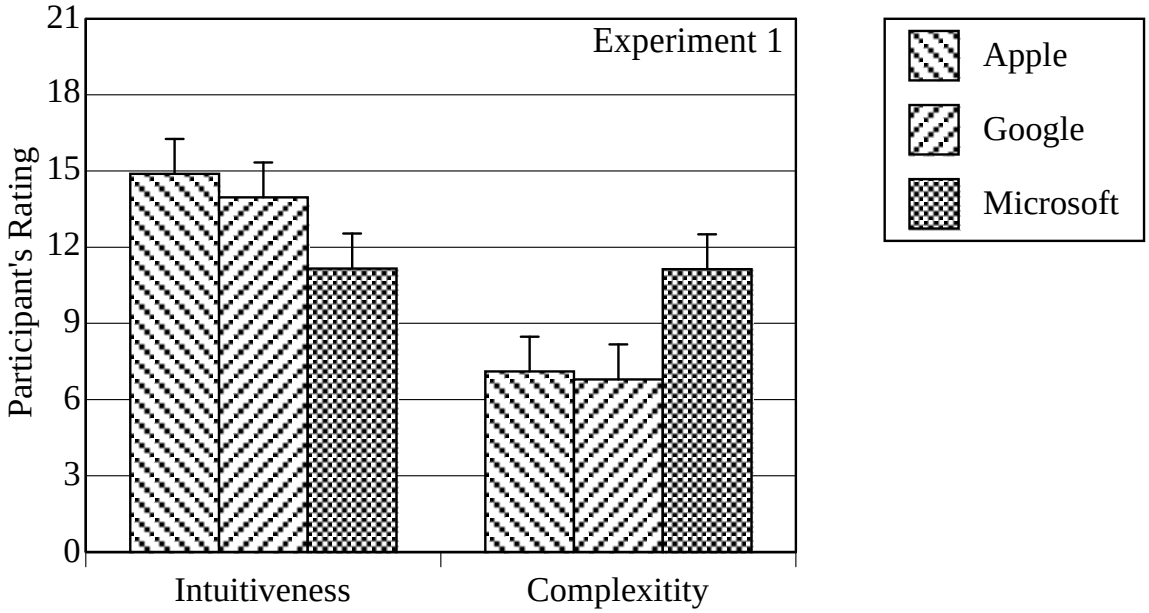

Figure 4. Mean ratings of intuitiveness and complexity for the Apple, Google, and Microsoft systems in Experiment 1. Error bars reflect 95% confidence intervals around the point estimate.

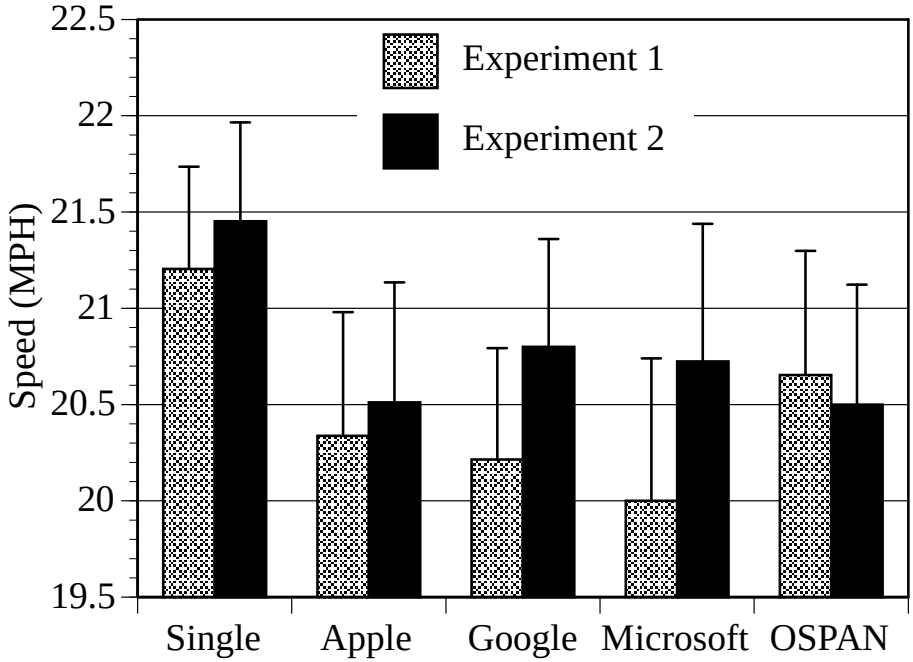

Figure 5. Average driving speed (in MPH) for the 5 conditions in Experiment 1. Error bars reflect 95% confidence intervals around the point estimate.

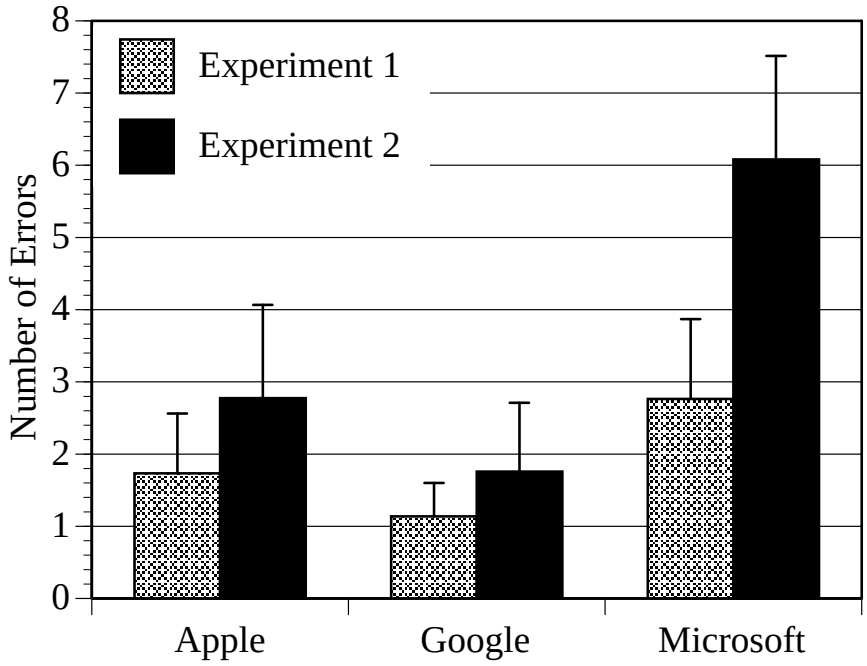

Figure 6. Average number of errors experienced by participants for the Apple, Google, and Microsoft systems in Experiments 1 and 2. Error bars reflect 95% confidence intervals around the point estimate.

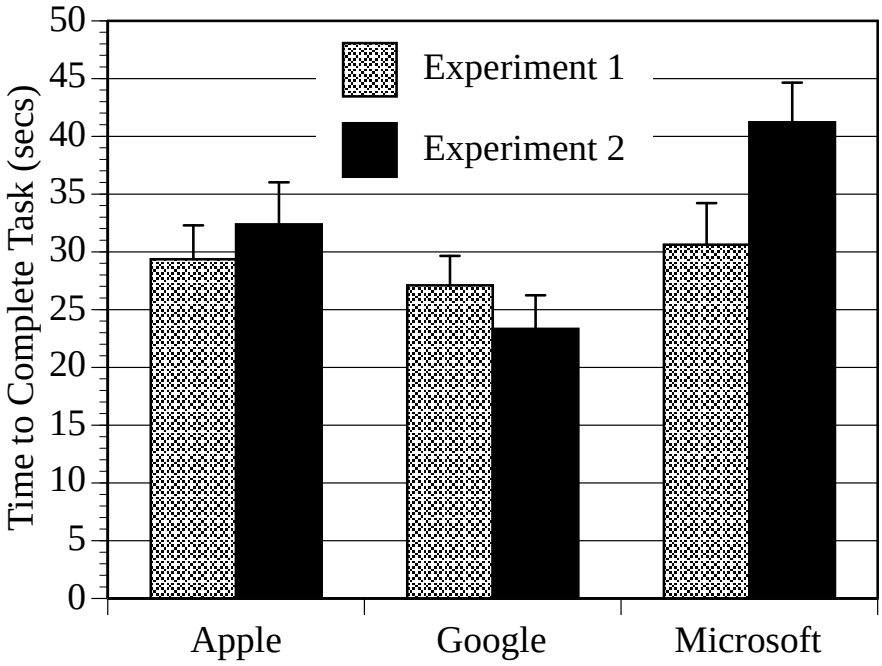

Figure 7. Average time to complete the secondary tasks for the Apple, Google, and Microsoft systems in Experiments 1 and 2. Error bars reflect 95% confidence intervals around the point estimate.

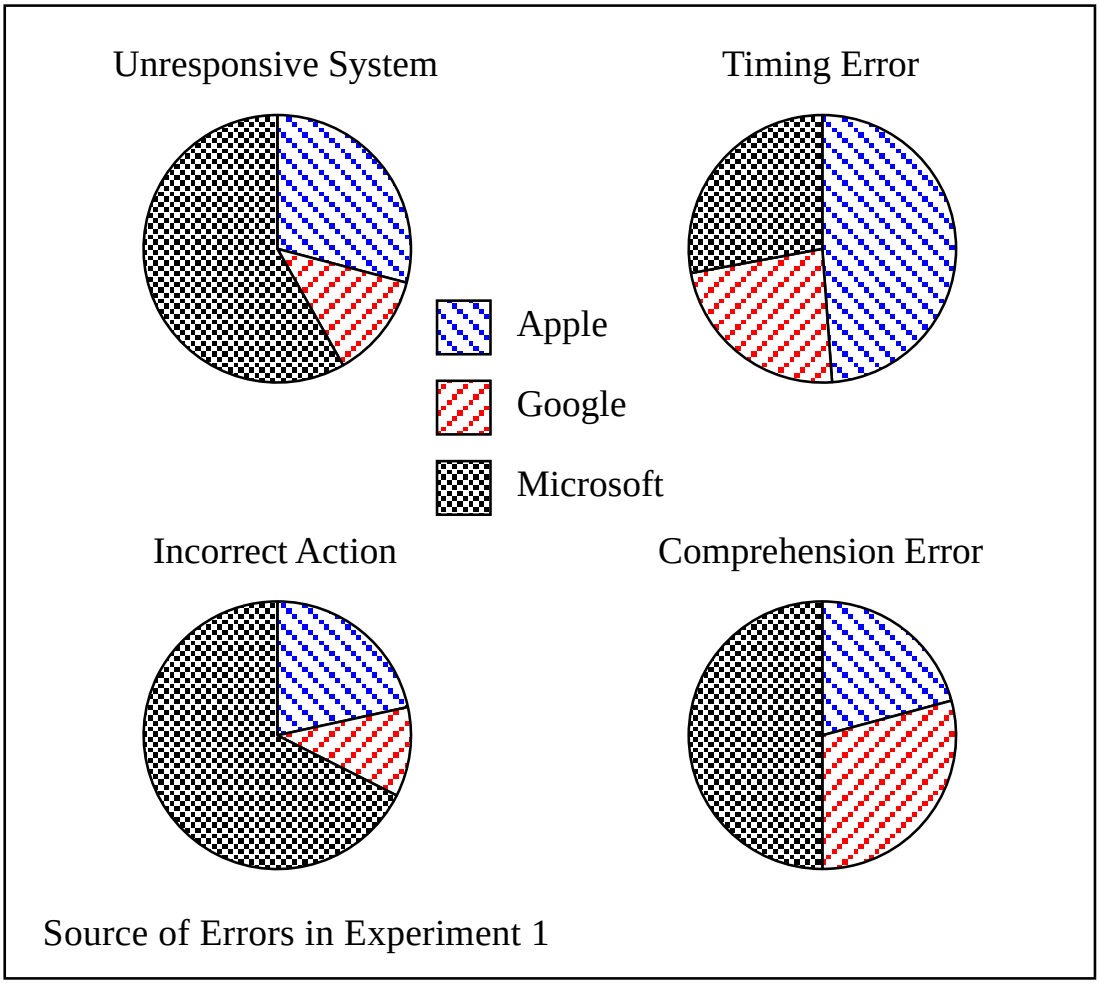

Figure 8. Relative proportion of errors by category for the Apple, Google, and Microsoft systems in Experiment 1.

Experiment 2

DRT

The RT and Hit Rate data for the DRT task are plotted as a function of secondary-task condition in Figures 9 and 10, respectively. Like Experiment 1, these are denoted by a "-0" for off task performance (e.g., not interacting with the digital voice assistant) and a "-1" for on-task performance (e.g., interacting with the digital voice assistant).

Reaction Time

The reaction time data from the DRT task were analyzed using a MANOVA with a 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 8 (Condition: Single-task, Apple-0, Google-0, Microsoft-0, Apple-1, Google-1, Microsoft-1, and OSPAN) split-plot factorial design. RT increased with Condition, $F(7, 217) = 38.87$, $p < .001$, $\eta^2 = .556$, and Age, $F(2, 31) = 5.00$, $p = .013$, $\eta^2 = .244$, but the Age by Condition interaction was not significant, $F(14, 217) = 1.01$, $p = .447$, $\eta^2 = .061$. Planned comparisons indicated that the single-task condition was significantly faster than the other secondary-task conditions ($p < .001$) and that the off-task secondary-tasks did not differ from each other (Apple-0 vs. Google-0, $p = .070$; Apple-0 vs. Microsoft-0, $p = .392$; Google-0 vs. Microsoft-0, $p = .189$). Each of these off-task conditions differed from their respective on-task performance (Apple-0 vs. Apple-1, $p < .001$; Google-0 vs. Google-1, $p < .001$; Microsoft-0 vs. Microsoft-1, $p < .001$). Importantly, none of the on-task secondary tasks differed significantly from the OSPAN condition *(*$p = .805$, $p = .297$, and $p = .569$ for Apple-1, Google-1, and Microsoft-1, respectively). Finally, the on-task performance for the three conditions did not differ from each other, (Apple-1 vs. Google-1, $p = .365$; Apple-1 vs. Microsoft-1, $p = .411$; Google-1 vs. Microsoft-1, $p = .612$).

Hit Rate

The Hit Rate data from the DRT task were analyzed using a MANOVA with a 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 8 (Condition: Single-task, Apple-0, Google-0, Microsoft-0, Apple-1, Google-1, Microsoft-1, and OSPAN) split-plot factorial design. Hit Rate decreased with Condition, $F(7, 217) = 9.33$, $p < .001$, $\eta^2 = .231$, and Age, $F(2, 31) = 4.00$ $p = .029$, $\eta^2 = .205$, and the Age by Condition interaction was also significant, $F(14, 217) = 1.81$, $p = .039$, $\eta^2 = .104$. Planned comparisons indicated that Hit Rate was significantly higher in the single-task condition than the other secondary-task conditions ($p < .001$) and that the off-task secondary-task conditions did not differ from each other (Apple-0 vs. Google-0, $p = .055$; Apple-0 vs. Microsoft-0, $p = .913$; Google-0 vs. Microsoft-0, $p = .052$). The off-task performance differed from on-task performance for Google-0 vs. Google-1, $p = .050$, and Microsoft-0 vs. Microsoft-1, $p = .002$, but not for Apple-0 vs. Apple-1 *(*$p = .100$). Importantly, none of the on-task secondary task conditions differed significantly from the OSPAN condition *(*$p = .821$, $p = .595$, and $p = .817$ for Apple-1, Google-1, and Microsoft-1, respectively). Finally, the on-task performance for the three conditions did not differ from each other, (Apple-1 vs. Google-1, $p = .741$; Apple-1 vs. Microsoft-1, $p = .946$; Google-1 vs. Microsoft-1, $p = .635$).

NASA TLX

The 6 scales of the NASA TLX, presented in Figure 11, were analyzed using a MANOVA with a 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 5 (Condition: Single-task, Apple, Google, Microsoft and OSPAN) split-plot factorial design. The MANOVA revealed a main effect of Condition, $F(24,488) = 6.40$, $p < .001$, $\eta^2 = .239$, but neither the Age, $F(12,54) = 1.15$, $p = .342$, $\eta^2 = .204$, nor the interaction were significant, $F(48,744) = 1.31$, $p = .076$, $\eta^2 = .078$.

Univariate tests were also performed on the 6 NASA TLX subscales. The *mental* sub-scale increased as a function of Condition, $F(4, 124) = 76.43$, $p < .001$, $\eta^2 = .771$, Age, $F(2, 31) = 3.59$, $p = .039$, $\eta^2 = .188$, and these two factors interacted, $F(8, 124) = 2.19$, $p = .032$, $\eta^2 = .124$. The *physical* sub-scale increased as a function of Condition, $F(4, 124) = 18.65$, $p < .001$, $\eta^2 = .376$, but neither the Age, $F(2, 31) = 2.74$, $p = .080$, $\eta^2 = .150$, nor the interaction were significant, $F(8, 124) = 1.32$, $p = .241$, $\eta^2 = .078$. The *temporal* sub-scale increased as a function of Condition, $F(4, 124) = 33.09$, $p < .001$, $\eta^2 = .516$, but neither the Age, $F(2, 31) = 2.73$, $p = .081$, $\eta^2 = .150$, nor the interaction were significant, $F(8, 124) = 1.98$, $p = .054$, $\eta^2 = .113$. The *performance* sub-scale increased as a function of Condition, $F(4, 124) = 24.16$, $p < .001$, $\eta^2 = .438$, but neither the Age, $F(2, 31) = 0.72$, $p = .495$, $\eta^2 = .044$, nor the interaction were significant, $F(8, 124) = 1.07$, $p = .391$, $\eta^2 = .064$. The *effort* sub-scale increased as a function of Condition, $F(4, 124) = 59.64$, $p < .001$, $\eta^2 = .658$, but neither the Age, $F(2, 31) = 2.30$, $p = .117$, $\eta^2 = .129$, nor the interaction were significant, $F(8, 124) = 1.40$, $p = .203$, $\eta^2 = .083$. Finally, the *frustration* sub-scale increased as a function of Condition, $F(4, 124) = 21.40$, $p < .001$, $\eta^2 = .408$ but neither the Age, $F(2, 31) = 0.22$, $p = .806$, $\eta^2 = .014$, nor the interaction were significant, $F(8, 124) = 1.01$, $p = .432$, $\eta^2 = .061$.

Intuitiveness and Complexity

Participants were also asked to rate how intuitive, usable, and easy it was to use the different smartphones. They also rated how complex, difficult, and confusing it was to use the different smartphones. Figure 12 presents the intuitiveness and complexity ratings on a 21-point scale where 1 reflected "not at all" and 21 reflected "very much."

Intuitiveness

A 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 3 (Condition: Apple, Google, Microsoft) split-plot MANOVA found that intuitiveness varied as a function of Condition, $F(2, 62) = 18.25$, $p < .001$, $\eta^2 = .371$ but neither Age, $F(2, 31) = 0.55$, $p = .581$, $\eta^2 = .034$, nor the Age by Condition were significant, $F(4, 62) = 0.87$, $p = .486$, $\eta^2 = .053$. Planned comparisons revealed that the intuitiveness of the Apple and Google systems did not differ ($p = .278$), and both were rated as more intuitive than the Microsoft system (Apple vs. Microsoft, $p < .001$; Google vs. Microsoft, $p < .031$).

Complexity

A 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 3 (Condition: Apple, Google, Microsoft) split-plot MANOVA found that complexity varied as a function of Condition, $F(2, 62) = 9.00$, $p < .001$, $\eta^2 = .225$ but neither the Age, $F(2, 31) = 1.10$, $p = .364$, $\eta^2 = .066$, nor the Age by Condition interaction were significant, $F(4, 62) = 0.22$, $p = .928$, $\eta^2 = .014$. Planned comparisons revealed that the complexity of the Apple and Google systems did not differ ($p = .949$), and both were rated as less complex than the Microsoft system (Apple vs. Microsoft, $p = .003$; Google vs. Microsoft, $p < .001$).

Video Analysis of Interactions

An analysis of the video of the participant's interactions was performed to determine the vehicle speed, presented in Figure 5, the number of observed errors, presented Figure 6 and the time to complete the task, presented in Figure 7. The relative frequency of the four

error categories for each of the smartphones is provided in Figure 13.

Vehicle Speed

A 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 3 (Condition: Apple, Google, Microsoft) split-plot MANOVA found that vehicle speed varied as a function of Condition, $F(4, 124) = 4.93$, $p < .001$, $\eta^2 = .137$, but neither Age, $F(2, 31) = 0.31$, $p = .736$, $\eta^2 = .020$ nor the Age by Condition interaction were significant, $F(8, 124) = 1.73$ $p = .098$, $\eta^2 = .100$. Planned comparisons revealed that the driving speed was higher in the single-task condition than in all other conditions ($p < .001$, $p = .012$, $p = .035$, and $p < .001$, respectively) and that speed did not differ from OSPAN for the Apple ($p = .963$), Google ($p = .162$), or Microsoft ($p = .420$) conditions Vehicle speed also did not significantly differ between the smartphone conditions (Apple vs. Google; $p = .179$; Apple vs. Microsoft, $p = .619$; and Google vs. Microsoft, $p = .480$).

Error Count

A 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 3 (Condition: Apple, Google, Microsoft) split-plot MANOVA found that the number of errors differed as a function of Condition, $F(2, 62) = 13.95$, $p < .001$, $\eta^2 = .310$, but neither the Age, $F(2, 31) = 0.88$, $p = .916$, $\eta^2 = .006$, nor the Age by Condition interaction were not significant, $F(4, 62) = 0.35$ $p = .840$, $\eta^2 = .022$. Planned comparisons revealed that the number of errors did not differ between the Apple and Google ($p = .177$), but the differences between Apple and Microsoft ($p < .001$) and the Google and Microsoft were significant ($p < .001$).

Task Completion Time

A 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 3 (Condition: Apple, Google, Microsoft) split-plot MANOVA found that the time to complete the task differed as a function of Condition, $F(2, 62) = 30.98$, $p < .001$, $\eta^2 = .500$, but neither the Age, $F(2, 31) = 0.95$, $p = .397$, $\eta^2 = .058$, nor the Age by Condition interaction were significant, $F(4, 62) = 0.28$ $p = .891$, $\eta^2 = .018$. All of the pair-wise planned comparisons were significant (Apple vs. Google, $p < .001$; or Apple vs. Microsoft, $p < .001$; and Google vs. Microsoft, $p < .001$).

A Comparison across Experiments

A number of analyses were performed to determine if the pattern obtained in the two experiments differed in any substantive way. For the analysis of the DRT data, a 2 (Experiment) by 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 8 (Condition: Single-task, Apple-0, Google-0, Microsoft-0, Apple-1, Google-1, Microsoft-1, and OSPAN) split-plot MANOVA was conducted to determine if the pattern differed across experiments. For RT, neither the main effect of Experiment, $F(1, 59) = 0.07$, $p = .792$, $\eta^2 = .001$, nor the Experiment by Age interaction, $F(2, 59) = 1.23$, $p = .301$, $\eta^2 = .040$, nor the Experiment by Condition interaction, $F(7, 413) = 1.74$, $p = .098$, $\eta^2 = .029$, nor the Experiment by Age by Condition interaction, $F(14, 413) = .728$, $p = .746$, $\eta^2 = .024$ were significant. For Hit Rate, neither the main effect of Experiment, $F(1,59) = 0.75$, $p = .390$, $\eta^2 = .013$, nor the

Experiment by Age interaction, $F(2, 59) = 2.86$, $p = .066$, $\eta^2 = .088$, nor the Experiment by Condition interaction, $F(7, 413) = 1.71$, $p = .104$, $\eta^2 = .028$, nor the Experiment by Age by Condition interaction, $F(14, 413)$ 1.65, $p = .065$, $\eta^2 = .053$ were significant. Like the course-grained analysis reported above, the overall pattern obtained in the fine-grained analysis of the two experiments was virtually identical.

A MANOVA compared the six NASA TLX measures using a 2 (Experiment) by 3 (Age Group: ages 21-34, ages 35-53, and ages 54-70) by 5 (Condition: Single-task, Apple, Google, Microsoft and OSPAN) split-plot factorial design. The MANOVA revealed that neither the main effect of Experiment, $F(6, 54) = 2.16$, $p = .062$, $\eta^2 = .193$, nor the Experiment by Age interaction, $F(12, 110) = 1.67$, $p = .083$, $\eta^2 = .154$, nor the Experiment by Age by Condition interaction, $F(48,1416)$ 1.09, $p = .315$, $\eta^2 = .036$ were significant. However, the Experiment by Condition interaction was significant, $F(24,936) = 2.39$, $p < .001$, $\eta^2 = .058$. On the whole, the pattern obtained in the two experiments was virtually identical.

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 are residual costs that persist after the smartphone interaction had terminated. To evaluate this residual cost in more detail, DRT performance in the off-task segments of the drive were 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 a smartphone interaction would be sorted into the second bin (which reflects the average of events between 3 and 6 seconds). Figure 14 presents the switch cost function collapsed over the two experiments and the different smartphone conditions within each experiment, as they did not produce different patterns in the data. In the figure, "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:

$$f(x) = a * (x^{-.1837641}), \text{ where } a = \exp(6.697153), \text{ with } R^2 = .97.$$

The residual switch costs show that it takes a surprisingly long time to dissipate. In fact, the data indicate that off-task performance (cf. Figures 1, 2, 9, and 10) reflects a mixture of "single-task" performance and the lingering costs associated with the voice-based interactions in the preceding on-task period. This is notable effect given the actual time to complete the tasks, approximately 30 seconds, (cf. Figure 7) was just over twice as long as the time it took for the residual costs to subside. While residual switch costs of much smaller magnitude have been observed in standard cognitive experiments (e.g., Rogers & Monsell, 1995), they often involve switching between two active tasks (Task A and Task B). The switch costs depicted in Figure 14 are notable because of their magnitude, their duration, and the fact that they are obtained even when there is no active switch to Task B. They appear to reflect the lingering act of disengaging from the cognitive processing associated with the smartphone task. From a practical perspective, the data indicate than just because a driver terminates a call or text message doesn't mean that they are no longer impaired.

The Cognitive Distraction Scale

The primary objective of the current research was to compare the cognitive workload associated with using 3 different intelligent personal assistants to complete common voice tasks while driving (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 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) 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.
>
> Step 3: The standardized condition averages across dependent measures were computed with an equal weighting for primary, secondary, and subjective 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 Figure 15. 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 Apple was 3.7, for Google was 3.3, and for Microsoft was 4.1.[3] The error bars represent 95% confidence intervals and document that the Google system

[3] If the workload ratings were based solely upon Experiment 1, they would be 3.4 for Apple Siri, 3.0 for Google Now and 3.8 for Microsoft Cortana.

was associated with a lower workload rating that the Apple and Microsoft systems, which did not significantly differ.

Figure 16 helps to put these workload ratings into perspective. Our prior research (Strayer et al., 2013) found that listening to the radio (1.2) or an audio book (1.7) were associated with a small increase in cognitive distraction, the conversation activities of conversing with a friend on a hand-held (2.4) or hands-free cell phone (2.3) were associated with a moderate increase in cognitive distraction, and interacting with a highly reliable speech-to-text condition (3.1) had a large cognitive distraction rating. Cooper et al., (2014) also used the cognitive workload scale to benchmark six 2013 voice-based systems. The ratings were Toyota (1.7), Hyundai (2.2), Chrysler (2.7), Ford (3.0), Mercedes (3.1) and Chevy (3.7).

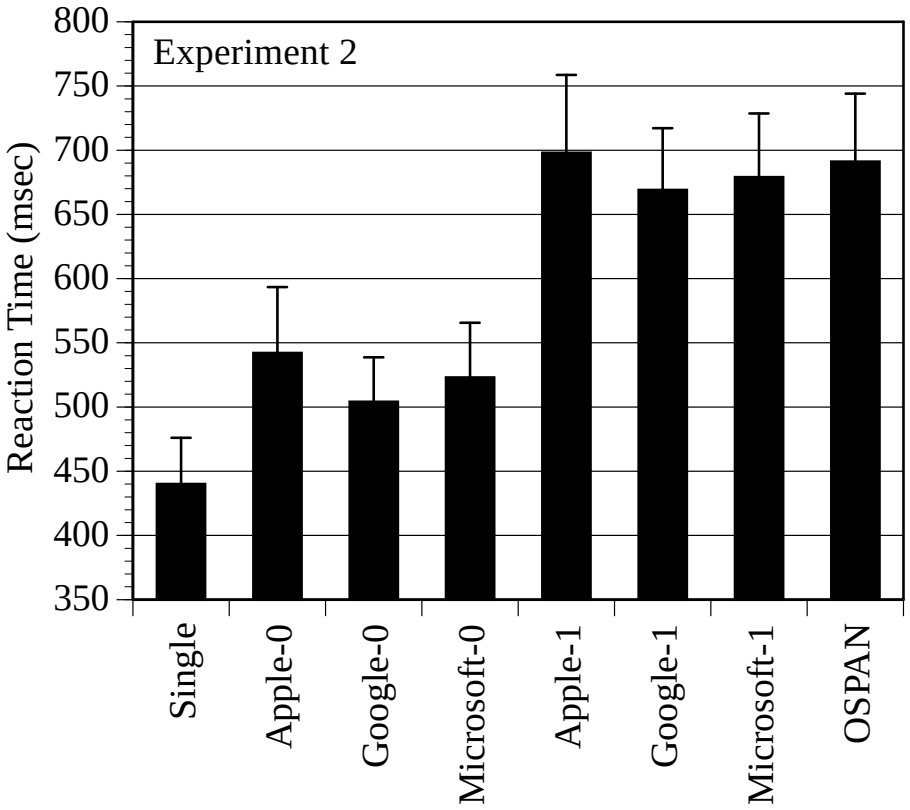

Figure 9. Mean DRT reaction time (in msec) for the single-task, OSPAN, and off-task (e.g., Google-0) and on-task (e.g., Google-1) performance for the Apple, Google, and Microsoft secondary tasks in Experiment 2. Error bars reflect 95% confidence intervals around the point estimate.

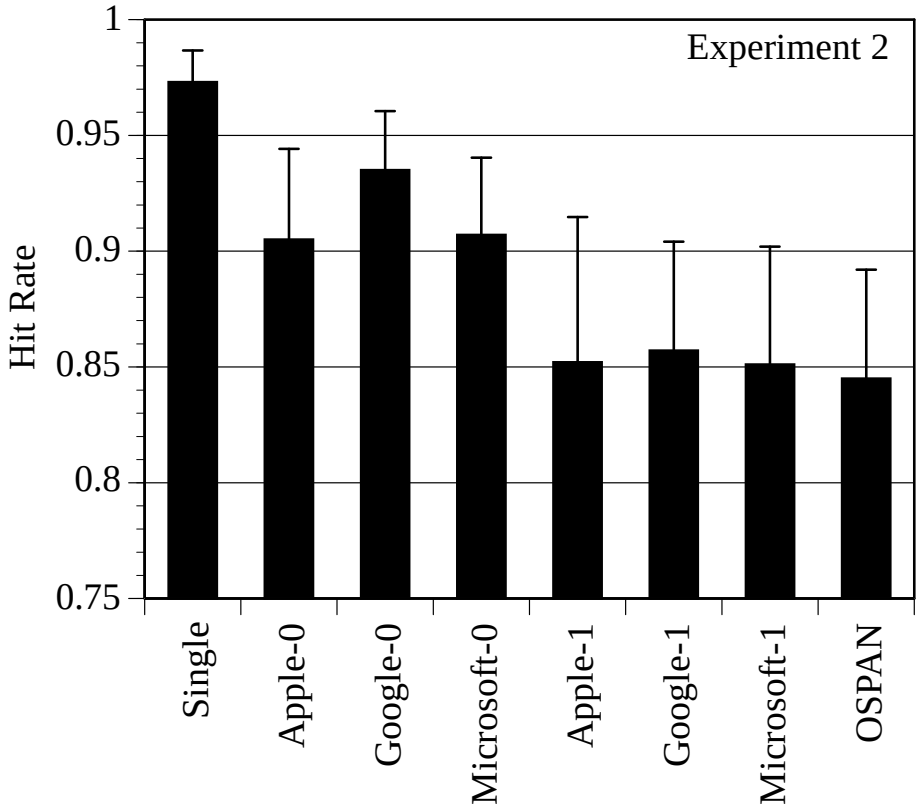

Figure 10. Mean DRT Hit Rate (an accuracy measure computed by determining the number of valid responses divided by the total number of responses) for the single-task, OSPAN, and off-task (e.g., Google-0) and on-task (e.g., Google-1) performance for the Apple, Google, and Microsoft secondary tasks in Experiment 2. Error bars reflect 95% confidence intervals around the point estimate.

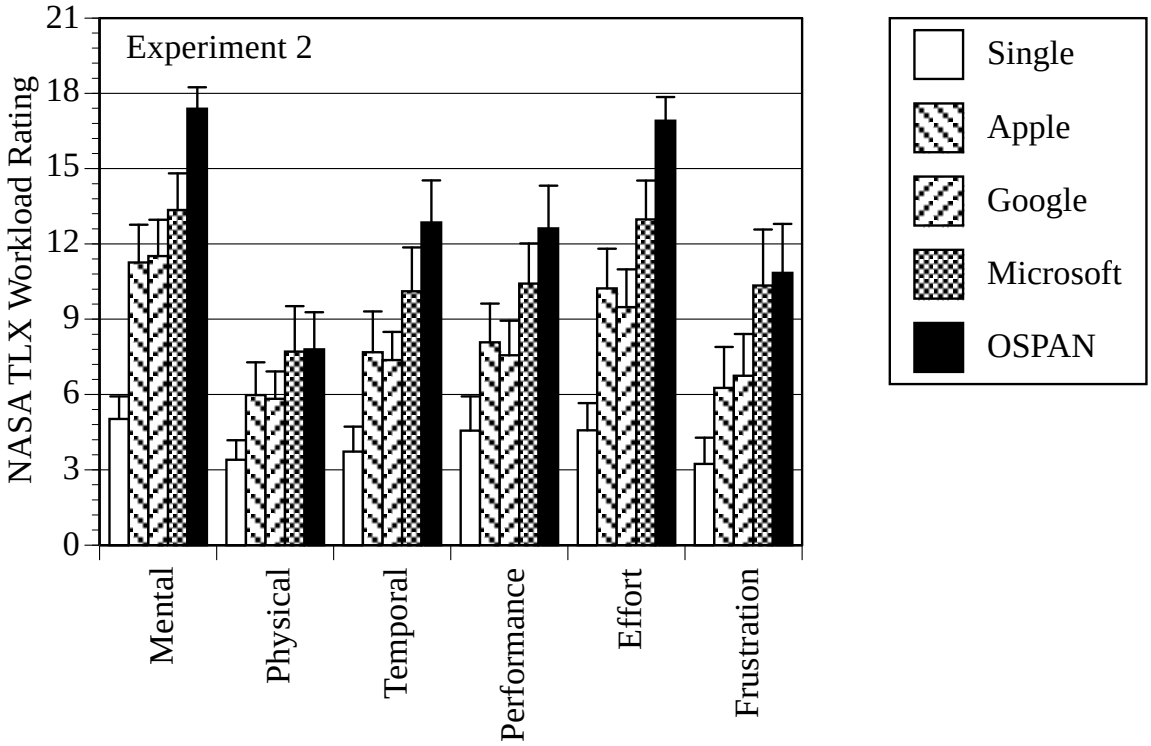

Figure 11. Mean NASA TLX ratings for the six sub-scales in the 5 conditions of Experiment 2. Error bars reflect 95% confidence intervals around the point estimate.

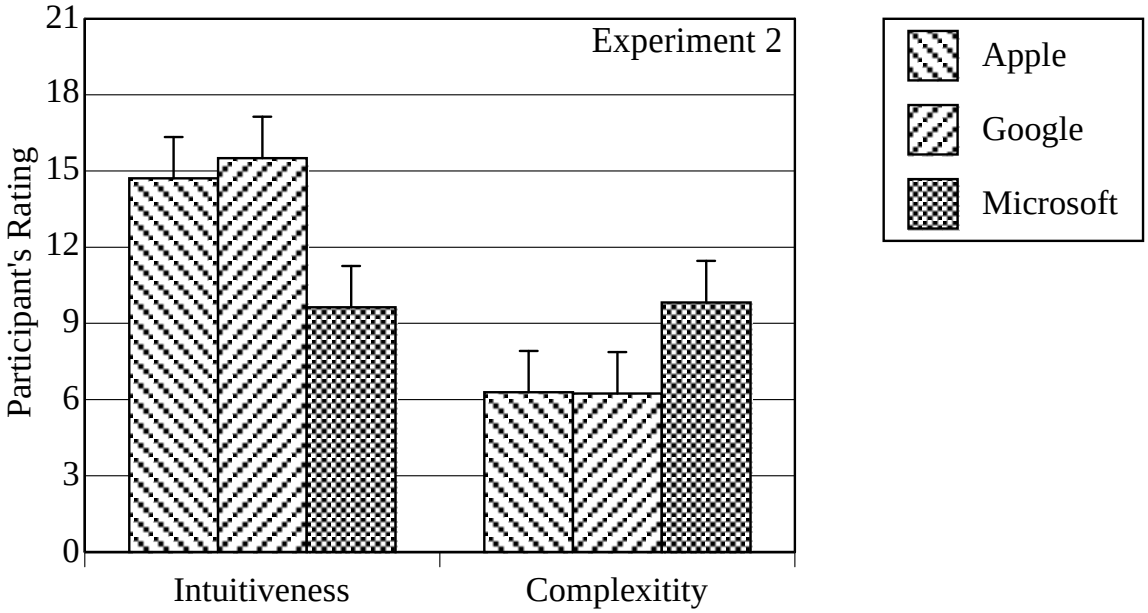

Figure 12. Mean ratings of intuitiveness and complexity for the Apple, Google, and Microsoft systems in Experiment 2. Error bars reflect 95% confidence intervals around the point estimate.

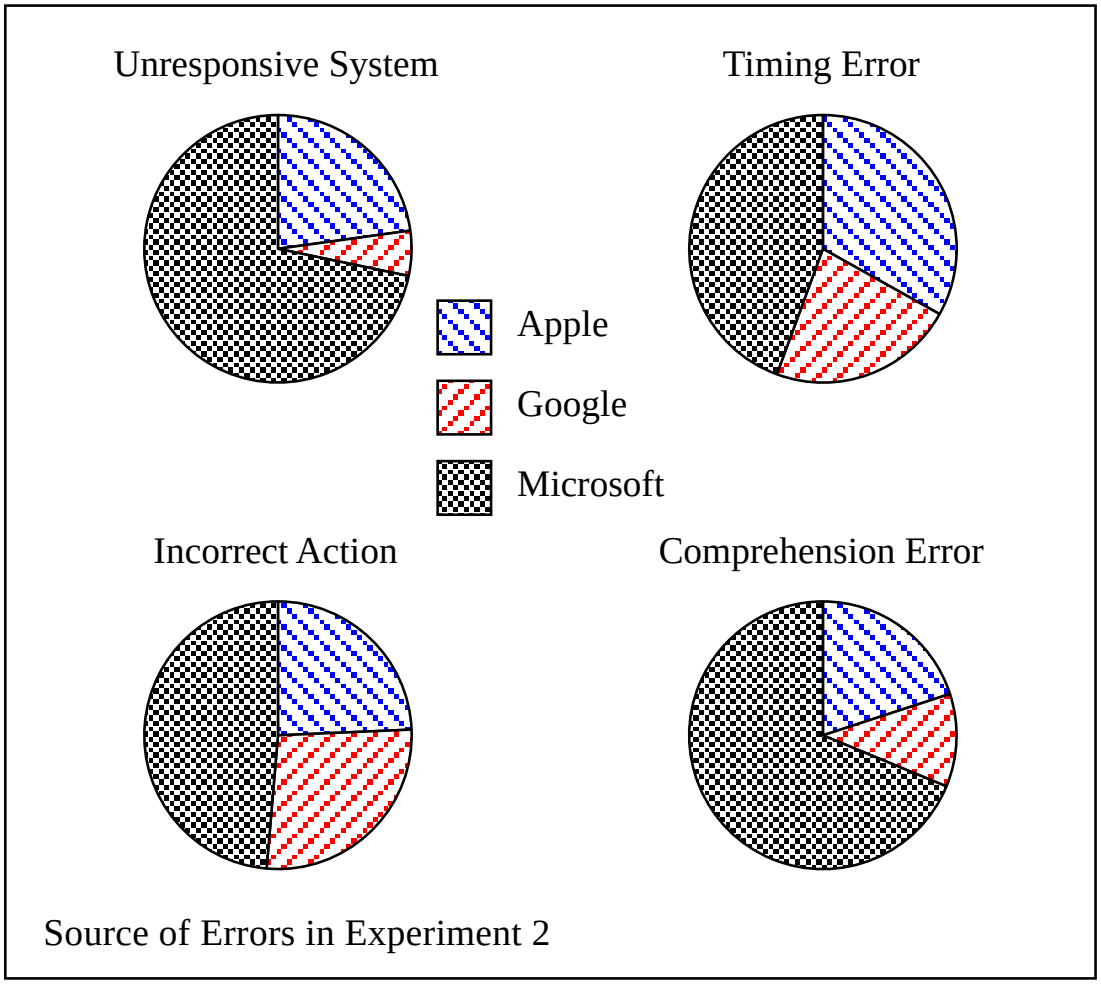

Figure 13. Relative proportion of errors by category for the Apple, Google, and Microsoft systems in Experiment 2.

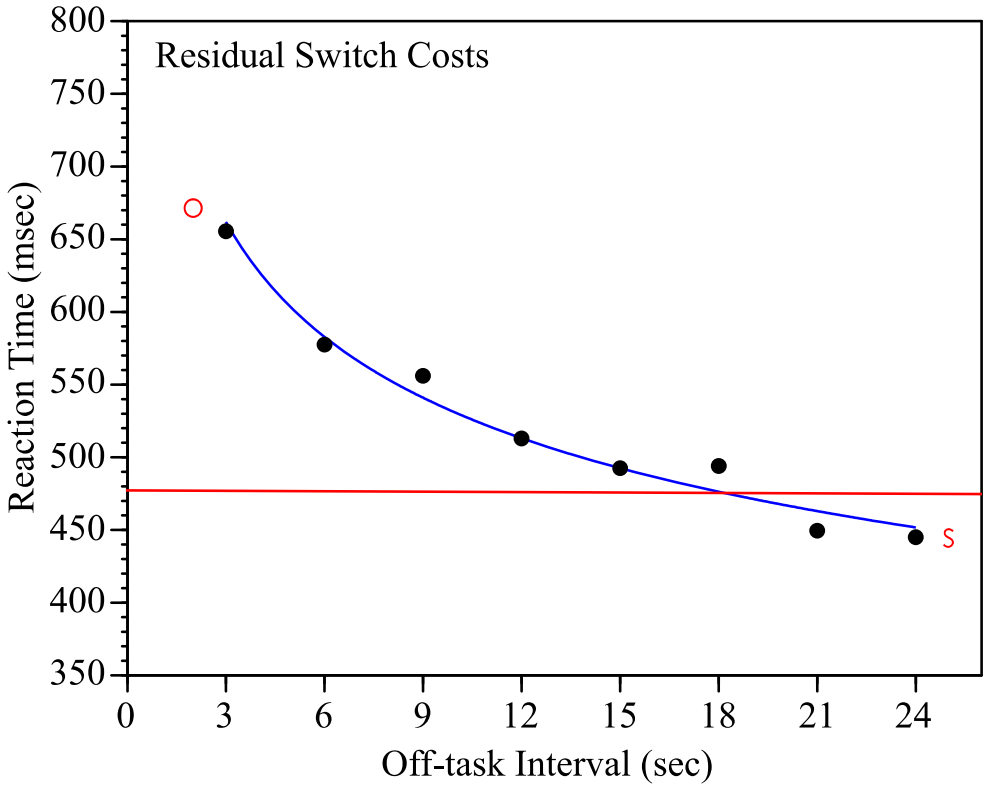

Figure 14. Residual switch costs in transitioning from on-task to off-task performance. The red "O" indicates average OSPAN RT from the DRT task; the red "S" indicates the average single-task RT from the DRT task. Off-task performance is distributed into 3-second intervals (relative to when the on-task activity terminated). The blue line represents the best fitting power function relating transition from on-task to single-task levels of performance. The solid red line represents the critical t-value for significant differences from the single-task condition. From the figure, residual switch costs are significantly different from the single-task baseline up to 18 seconds after the on-task interval had terminated.

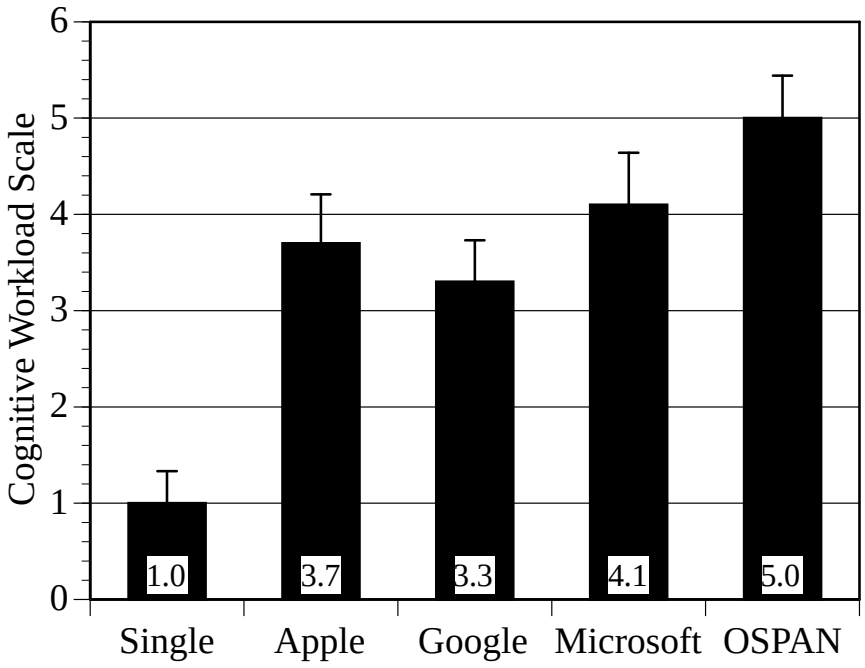

Figure 15. The cognitive workload scale for the Apple, Google, and Microsoft systems compared to single-task (category 1) and OSPAN (category 5). Error bars reflect 95% confidence intervals around the point estimate.

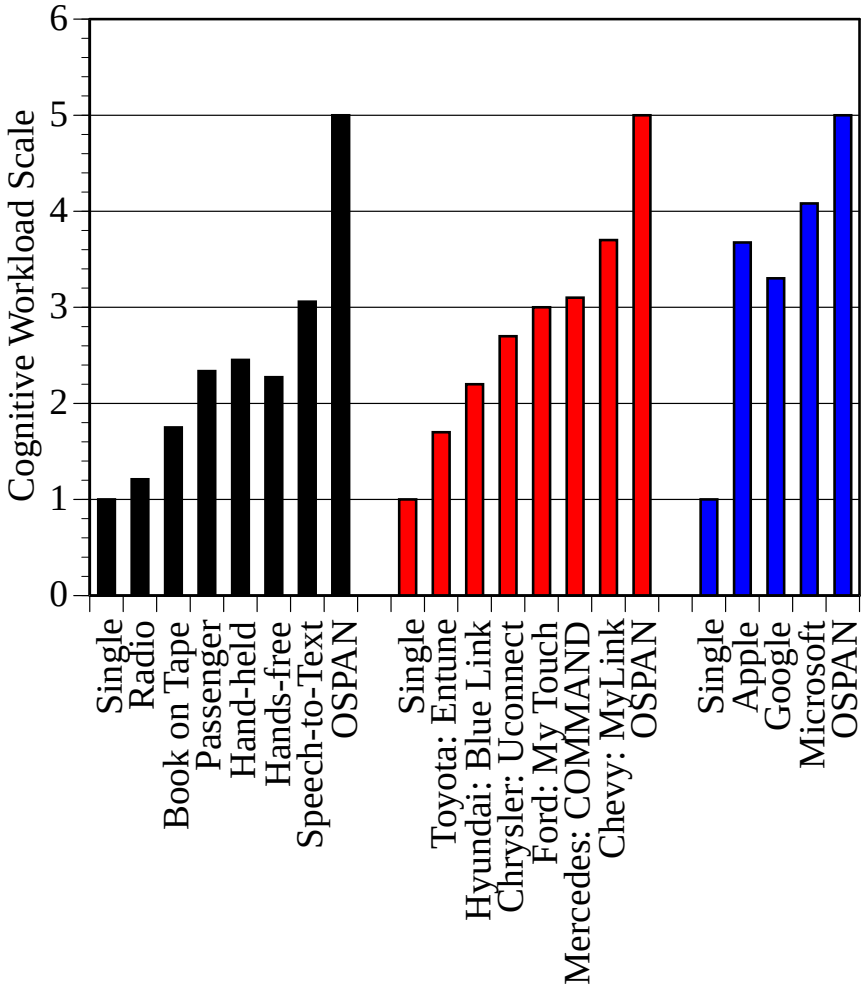

Figure 16. The workload scale for Strayer et al., (2013, the black bars), Cooper et al., (2014, the red bars), and the current research (blue bars).

Discussion

Experiment 1

Experiment 1 examined the impact of intelligent personal assistant interactions using three different smartphone systems (Apple's *Siri,* Google's *Google Now* for Android phones, and Microsoft's *Cortana*). Each of the smartphone conditions impaired performance when compared to the single-task baseline. There were also systematic differences between the smartphone systems, such that interactions using the Google system had lower levels of workload than the Apple and Microsoft systems. Our analysis revealed that these differences were associated with the number of system errors and the complexity and intuitiveness of the systems. Surprisingly large delays in RT were observed in the DRT data when drivers were interacting with the devices – in each case, on-task DRT performance was similar to that of the demanding OSPAN task. Importantly, the analysis of DRT performance found that off-task performance was impaired relative to the single-task baseline. This pattern suggests that there are residual costs associated using each of the devices that take a significant time to dissipate.

General Discussion

The objective of the current research was to examine the impact of voice-based interactions using three different smartphone systems (Apple's *Siri,* Google's *Google Now*, and Microsoft's *Cortana*) on the workload experienced by the driver. We selected tasks (voice dialing, contact calling, music selection, and voice-texting) that could be performed with no visual component, and only a minimal button press to initiate the interaction. As such, the interactions were primarily cognitive in nature (i.e., aside from the initial button push on the remote headphone there was no requirement for visual or manual interaction with the device). The experiments were structured such that the car, driving environment, wireless provider (T-Mobile with 4-5 bars of service), and headphone (with ear-bud, microphone, and remote button) were identical and the order in which the conditions were performed was counterbalanced across participants. Moreover, before each test began, participants practiced with each system to ensure that they were familiar with the device and its functions. Note that in some cases this also involved resetting the smartphone so that it could learn the user's voice patterns. Thus, the only difference between the conditions was the smartphone functionality provided by the Apple, Google, and Microsoft systems.

In both studies, the cognitive workload when using the smartphones was significantly higher than that of the single-task baseline. There were also systematic differences between the smartphone systems, such that interactions using the Google system had significantly lower levels of workload than the Apple and Microsoft systems. Video analysis revealed that these differences were associated with the number of system errors, the time to complete an action, and the complexity and intuitiveness of the systems. Finally, high levels of workload were observed in the analysis of the DRT data when drivers were interacting with the devices – on-task DRT performance did not significantly differ from that of the demanding OSPAN task.

Takeaways from the current research

There are four key takeaways from the current research. First, using the voice-based intelligent personal assistants to complete common in-vehicle tasks such as: calling a contact, dialing a phone number, selecting music, or sending a text messages was associated with a significant increase in the workload of the driver, compared to single-task driving conditions. In our testing, the overall workload ratings associated with using the smartphone ranged from 3.3 to 4.1, reflecting a moderate to high level of cognitive workload. Moreover, the workload of the driver was virtually identical for placing calls, selecting music and the seemingly more demanding activity of sending of text messages. These levels of workload are similar those reported by Cooper et al., (2014) in their evaluation of voice-based interactions in 2013 vehicles.

Second, there were significant differences in the cognitive workload experienced by the driver when they used the different smartphones to perform the same tasks in the same driving conditions. Notably, the Google system outperformed the Apple and Microsoft systems. Our analysis found that this was directly related to the number of system errors and the intuitiveness/complexity of the different systems. It is noteworthy that this same factor differentiated the levels of workload in the evaluation by Cooper et al. (2014). Indeed, a general principle to emerge from the research is that robust, error-free systems tend to have lower workload than rigid error-prone ones. Thus, enhanced usability testing and an iterative design process to minimize system errors in the user interface have the potential to make these systems less cognitively demanding on the driver.

Third, the analysis of workload using the on/off task DRT data found that "on-task" performance was associated with surprisingly high levels of workload. In fact, in many instances the on-task levels of workload experienced by the driver did not differ from the mentally demanding OSPAN task (a category-5 level of workload). This high level of workload should serve as a caution that these "hands-free" 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), these voice-based smartphone interactions would appear to be significantly more demanding than typical cell phone conversations, which had cognitive workload levels around 2.3. It is possible that the timing and wording demands associated with the smartphone interactions may be a source of the increased level of cognitive workload.

Fourth, the off-task DRT data provided evidence of persistence interference following voice-based interactions on the smartphones. Despite the fact that the participants were not interacting with the smartphone in any way, there were residual costs associated with the prior interaction that were evident in both experiments and for all three smartphones. These residual switch 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 18 seconds after an interaction had been completed. These findings have implications for self-regulatory strategies, such as choosing to dial or send a text at a stoplight, because the costs of these interactions are likely to persist when the light turns green. The residual switch costs may be related to the driver reestablishing situation awareness of the driving environment that was lost during the smartphone interaction (Fisher & Strayer, 2014; Strayer, in press).

Conclusion

The goal of the current research was to examine the impact of voice-based interactions using three different smartphone systems (Apple's *Siri,* Google's *Google Now* for Android phones, and Microsoft's *Cortana*) on the cognitive workload of the driver. We found systematic differences between the systems and video analysis revealed that the differences were associated with the number of system errors, the time to complete an action, and the complexity and intuitiveness of the devices. The data suggest caution in introducing voice-based interactions in the vehicle because of the surprisingly high levels of workload associated with some of these interactions.

References

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.*

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.

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.

Parasuraman, R., & Davies, D. R. (1984). Varieties of Attention. Academic Press.

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.

Rogers R. D., Monsell S. (1995). The cost of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*, 207–231.

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., Medeiros-Ward, N., & Biondi, F. (2013). Measuring cognitive distraction in the automobile. *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.