Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale
2017, Vol. 71, No. 2, 93–110

1196-1961/17/$12.00 http://dx.doi.org/10.1037/cep0000104

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

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

The goal of this research was to examine the impact of voice-based interactions using 3 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 2 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 (DRT), 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. 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 Operation Span (OSPAN) task. The analysis also found residual costs associated using each of the smartphones that took a 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.

*Keywords:* cognitive distraction, cognitive workload, divided attention, driving, multitasking

Driver distraction, operationalized here as "the diversion of attention away from activities critical for safe driving toward a competing activity" (Regan, Hallett, & Gordon, 2011; see also Engström et al., 2005; Regan & Strayer, 2014), is increasingly recognized as a significant source of injuries and fatalities on the roadway. In fact, a recent report by the AAA Foundation for Traffic Safety found that 58% of all crashes among teenaged drivers could be attributed to driver inattention (Carney et al., 2015). Impairments to driving arise from visual/manual interference, for example when a driver takes his or her eyes off the road to look at or manually interact with a device (this is often referred to as "structural interference," that is, your eyes cannot focus on two disparate locations at the same time). Impairments can also stem from cognitive sources of distraction when attention is withdrawn from the processing of information necessary for the safe operation of a motor vehicle. For example, talking on a cell phone requires mental resources to perform the conversation task. Performing the cell phone conversation while driving diverts attention from the driving task (Strayer & Johnston, 2001). Given the capacity limitations of human attention (e.g., Kahneman, 1973; Heathcote et al., 2015), the mental resources available for driving are inversely related to the cognitive workload of the concurrent secondary task.

A motorist's situational awareness of the driving environment is a mental state that is contingent upon several cognitive processes that are dependent on attentional resources (Endsley, 1995, 2015). These include visual *Scanning* of the driving environment for indications of threats, *Predicting* and anticipating where potential threats might materialize if they are not visible, *Identifying* threats and objects in the driving environment when they are in the field of view, *Deciding* whether an action is necessary and what action is necessary, and *Executing* appropriate *Responses* (*SPIDER* for short; for a review, see Strayer & Fisher, 2016). When drivers engage in secondary-task activities that are unrelated to the safe operation of the vehicle, attention is often diverted from driving, impairing performance on one or more of these SPIDER-related processes (Regan & Strayer, 2014). Consequently, activities that divert attention from the task of driving degrade the driver's situational awareness and compromise the ability of the driver to safely operate their vehicle.

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 phases to the NHTSA (2012) guidelines. The Phase 1 guidelines, entered into the Federal Register on March 15, 2012, address visual-manual interfaces for devices installed by manufactures. The Phase 2 guidelines, scheduled for release sometime in

David L. Strayer, Joel M. Cooper, Jonna Turrill, James R. Coleman, and Rachel J. Hopman, Department of Psychology, University of Utah.

This research was supported by a grant from the AAA Foundation for Traffic Safety.

Correspondence concerning this article should be addressed to David L. Strayer, Department of Psychology, University of Utah, 380 South 1530 East, RM 502, Salt Lake City, UT 84112. E-mail: David.Strayer@utah.edu

2017, 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.

To allow drivers to maintain their eyes on the driving environment, nearly every vehicle sold in the U.S. 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 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 (Strayer, 2015). In particular, although voice-based technology may allow the driver to keep their eyes on the road, they may actually increase the level of cognitive workload associated with interactions with technology in the vehicle (Strayer et al., 2015).

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 nondistracted 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. This method was also used to assess the cognitive workload in six 2013 vehicles equipped with voice-based technology (Cooper et al., 2014). We found striking differences in the workload ratings associated with the different systems, with the Toyota Entune system having a workload rating roughly equivalent to listening to a book on tape and the Chevy Mylink 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.

An 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]

### Research Objectives and Experimental Overview

Following the protocol used by Strayer et al., (2013), 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. Unlike our prior testing where the secondary-tasks were continuous, the smartphone interactions reported in this article were intermittent. Six different voice-based interactions were initiated when participants reached prespecified locations on the 2.7-mile course (see below for details). This provided a repeating on-task/off-task pattern with the on-task interval approximately 30 seconds in duration and the off-task interval approximately 40 seconds in duration. This method affords a fine-grained exploration of the structure and time-course of mental workload across the driving interval. In particular, we sought to assess the workload in the on-task interval relative to the single-task and OSPAN benchmarks. Is the cognitive distraction from these interactions so severe that it is clearly incompatible with safe driving or is it sufficiently benign that it is nearly indistinguishable from activities such as listening to the radio? Additionally, the on/off task structure provides an opportunity to learn how quickly any impairment from the dual-task interaction abates. That is, how quickly does performance return to single-task baseline levels? The task switching literature (e.g., Rogers & Monsell, 1995) suggests that the abatement is unlikely to be instantaneous, but is the interval a matter of milliseconds or seconds? The longer the recovery interval, the greater negative impact on traffic safety.

Our prior research studied younger drivers (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 the older population and it is unclear how demanding older drivers will find these dual-task interactions. Studies have documented greater costs of multitasking for older adults in the laboratory (e.g., Hartley & Little, 1999; Kramer & Larish, 1996; McDowd & Shaw, 2000); however, the level of cognitive workload experienced by older drivers using smartphone systems is unknown. The current research recruited drivers between the ages of 21 and 70 to learn if older adults exhibit greater dual-task costs when operating a motor vehicle than younger drivers.

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 (e.g., dialing, music selection, etc.) while driving. In the second experiment, we evaluated the cognitive demands associated with sending voice-based text messages. We anticipated that the voice-based interactions would be more demanding in Experiment 2, given the more complex nature of the interactions. The objective of this research was to determine how the different smartphone systems compare with each other and to identify the bases for any observed differences in the cognitive workload experienced by the driver. The standardized testing protocol also facilitated a comparison of the smartphone systems with the embedded systems found in the different OEM systems?[2]

## Experiment 1

### Method

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

[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."

[2] Prior to the study, we solicited feedback on the research design from representatives from Google, Apple, and Microsoft, and this resulted in a number of refinements to the research protocol.

Thirty-one participants were recruited for Experiment 1 (16 males, 15 females).[3] Participants ranged in age from 21 to 70 years ($\bar{x}$ = 42 years). 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., valid drivers' license, no at-fault accidents in the past year). In addition, following University of Utah policy, each prospective participant was required to complete a 20-min online defensive driving course and pass a 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 ($\bar{x}$ = 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 U.S. English dialect.

**Design.** A 5 (condition) × 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-subjects 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) × 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 onto 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 states in the United 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 centre 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.[4] 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 min 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. The DRT protocol followed the specifications outlined in ISO DIS 17488 (2015). The device consisted of an LED light mounted to a flexible arm that was connected to a headband, a microswitch attached to the participant's left thumb, and a dedicated microprocessor to handle all stimulus timing and response data. The light was positioned in the periphery of participants' left eye (approximately 15° to the left and 7.5° above participants' 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 and responses were collected 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 × 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). 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, 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 as follows:

1. How mentally demanding was the task?
2. How physically demanding was the task?
3. How hurried or rushed was the pace of the task?

[3] A preliminary analysis found that the main effect of Gender was not significant, nor did Gender interact with any of the other factors (all *ps* > .50), hence we collapsed across this variable for all the analyses reported in this article.

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

4. How successful were you in accomplishing what you were asked to do?

5. How hard did you have to work to accomplish your level of performance?

6. 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.

**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, 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 during daylight hours 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 using a fully within-subjects design with the order of conditions counterbalanced across participants. One of the conditions 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. Another of the conditions 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 prespecified 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. 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. 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 $26 \times 26$?

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-min condition, the driver completed the DRT. 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 100 msec or greater than 2500 msec 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. The portions of the roadway used in the data analyses were straight sections with a speed limit of 25 MPH that were identical for the five conditions in the experiment. During testing, task engagement was flagged by the experimenter through keyboard input that facilitated comparison of performance in the secondary-task smartphone conditions when the participant was actively engaged in an activity (on-task) or had finished that 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 RTs to the DRT task divided by the number of valid RTs.
- 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, and 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 2.4 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 user's 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.

## Results

**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 1a and 2a, 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," nonresponses 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 "-On," or when participants were operating the vehicle without concurrent secondary interaction (e.g., off-task), denoted by a suffix of "-Off."

***Reaction time.*** The RT data from the DRT when participants were on-task were analyzed using a ANOVA with a 3 (Age Group: ages 21–34, 35–53, and 54–70) × 8 (Condition: Single-task, Apple-Off, Google-Off, Microsoft-Off, Apple-On, Google-On, Microsoft-On, 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 with the DRT 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-Off condition was faster than both the Apple-Off condition ($p = .023$) and Microsoft-Off condition ($p = .021$), that the Apple-Off and Microsoft-Off conditions did not significantly differ ($p = .630$), and that each of these conditions differed from their respective on-task performance (Apple-Off vs. Apple-On, $p < .001$; Google-Off vs. Google-On, $p < .001$; Microsoft-Off vs. Microsoft-On, $p <$

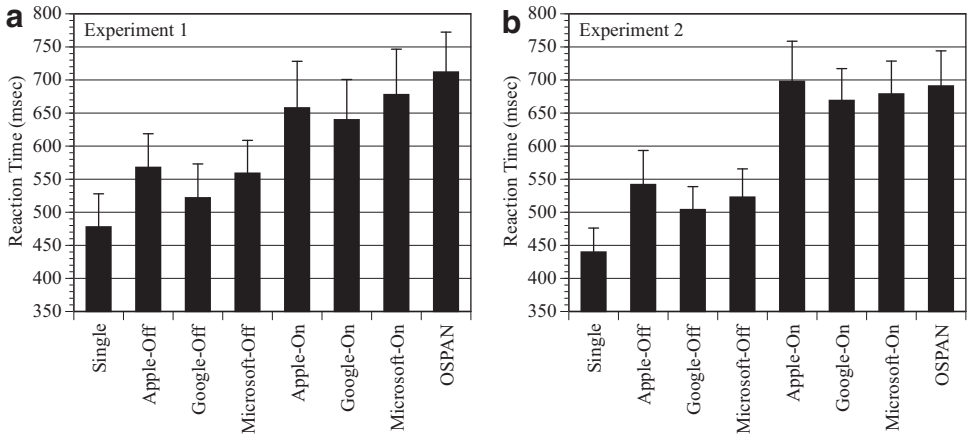


*Figure 1.* Mean DRT RT (in msec) for the single-task, OSPAN, and off-task (e.g., Google-Off) and on-task (e.g., Google-On) performance for the Apple, Google, and Microsoft secondary-tasks in Experiment 1 (a; left panel) and Experiment 2 (b; right panel). Error bars reflect 95% confidence intervals around the point estimate. OSPAN = Operation Span.

.001).[5] Importantly, neither the Apple-On, nor the Microsoft-On conditions significantly differed from the OSPAN condition ($p = .061$ and $p = .130$), whereas the Google-On condition was significantly faster than OSPAN ($p = .003$). Finally, the on-task performance for the three smartphone conditions did not differ from each other, (Apple-On vs. Google-On, $p = .527$; Apple-On vs. Microsoft-On, $p = .426$; Google-On vs. Microsoft-On, $p = .153$).

***Hit rate.*** The Hit Rate data from the DRT task were analyzed using a ANOVA with a 3 (Age Group) × 8 (Condition) 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 × 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 versus Google-Off comparison, which did not significantly differ ($p = .599$). Hit Rate was higher in the Google-Off condition than the Apple-Off ($p = .006$) and Microsoft-Off ($p = .017$) conditions, and the Apple-Off and Microsoft-Off conditions did not significantly differ ($p = .815$).[6] The off-task performance differed from on-task performance for Google-Off versus Google-On ($p < .006$), and Microsoft-Off versus Microsoft-On ($p < .013$), but not for Apple-Off versus Apple-On ($p = .057$). Hit Rate was higher for each of the on-task secondary-task conditions than OSPAN ($p = .032$, $p = .002$, and $p = .021$ for Apple-On, Google-On, and Microsoft-On, respectively). Finally, the on-task performance for the three secondary-task conditions did not differ from each other (Apple-On vs. Google-On, $p = .051$; Apple-On vs. Microsoft-On, $p = .851$; Google-On vs. Microsoft-On, $p = .058$).

**NASA TLX.** The 6 scales of the NASA TLX, presented in Figure 3a, were analyzed using a MANOVA with a 3 (Age Group) × 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 Age × Condition 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* subscale 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* subscale 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 Age × Condition interaction were significant, $F(8, 112) = 0.67$, $p = .713$, $\eta^2 = .046$. The *temporal* subscale 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 Age × Condition interaction were significant, $F(8, 112) = 0.63$, $p = .747$, $\eta^2 = .043$. The *performance* subscale 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 Age × Condition interaction were significant, $F(8, 112) = 0.81$, $p = .598$, $\eta^2 = .054$. The *effort* subscale 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 Age × Condition interaction were significant, $F(8, 112) = 0.81$, $p = .597$, $\eta^2 = .055$. Finally, the *frustration* subscale 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 Age × Condition interaction were significant, $F(8, 112) = 0.45$, $p = .889$, $\eta^2 = .031$.[7]

**Intuitiveness and complexity.** Participants were also asked to rate how intuitive, usable, and easy it was to use the different smartphones' intelligent personal assistants. They also rated how complex, difficult, and confusing it was to use the different smartphones' intelligent personal assistants. Figure 4a presents the in-

[5] Effect size estimates for pairwise differences in RT are presented in Table 1.

[6] Effect size estimates for pairwise differences in Hit Rate are presented in Table 2.

[7] Effect size estimates for the pairwise differences in NASA TLX ratings are presented in Table 3.

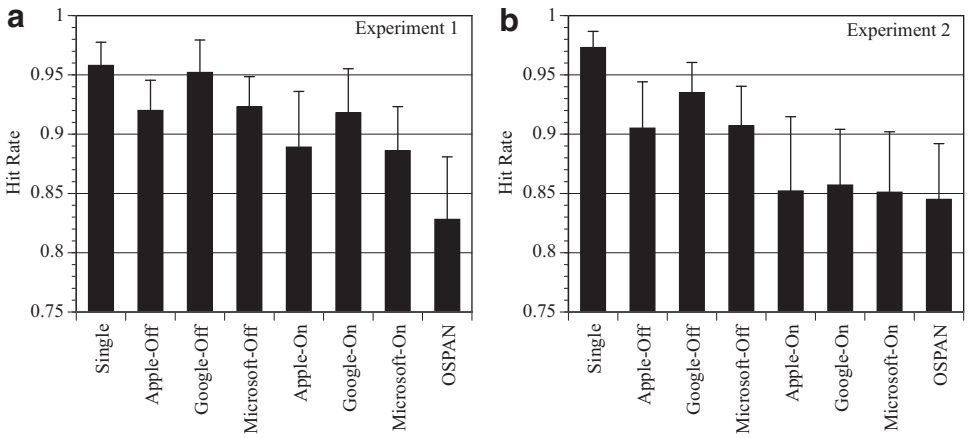


*Figure 2.* Mean DRT Hit Rate (an accuracy measure computed by determining the number of valid responses divided by the total number of responses for the single-task, OSPAN, and off-task (e.g., Google-Off) and on-task (e.g., Google-On) performance for the Apple, Google, and Microsoft secondary-tasks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

tuitiveness and complexity ratings on a 21-point scale where 1 reflected *not at all* and 21 reflected *very much*.

***Intuitiveness.*** A 3 (Age Group) × 3 (Condition: Apple, Google, Microsoft) split-plot ANOVA 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 × 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, $p = .036$).[8]

***Complexity.*** A 3 (Age Group) × 3 (Condition: Apple, Google, Microsoft) split-plot ANOVA 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 × 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.** A video analysis of participants' interactions was performed to determine the vehicle speed (see Figure 5), the number of observed errors (see Figure 6), and the time to complete the task (see Figure 7). The relative frequency of the four error categories for each of the smartphones is provided in Figure 8a.

***Vehicle speed.*** A 3 (Age Group) × 3 (Condition: Apple, Google, Microsoft) split-plot ANOVA 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 × 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, $p = .508$).[9]

***Error count.*** A 3 (Age Group) × 3 (Condition: Apple, Google, Microsoft) split-plot ANOVA 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 ($p = .041$).[10]

***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 ANOVA found that the time to complete the tasks 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 × Condition interaction was also not significant, $F(4, 56) = 1.69$ $p = .166$, $\eta^2 = .108$. None of the pairwise planned comparisons was significant (Apple vs. Google, $p = .508$; or Apple vs. Microsoft, $p = .101$; and Google vs. Microsoft, $p = .576$).

## Discussion

Experiment 1 examined the impact of intelligent personal assistant interactions using three different smartphone systems (Apple's

[8] Effect size estimates for the pairwise differences in Intuitiveness and Complexity are presented in Table 4.

[9] Effect size estimates for the pairwise differences in Vehicle Speed are presented in Table 5.

[10] Effect size estimates for the pairwise differences in Error Count and Task Completion Time are presented in Table 6.

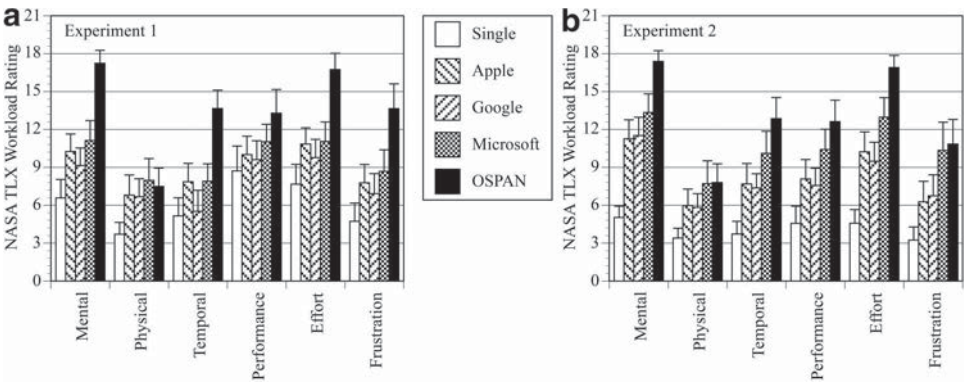


*Figure 3.* Mean NASA TLX ratings for the six subscales in the 5 conditions of Experiment 1 (a; left panel) and Experiment 2 (b; right panel). Error bars reflect 95% confidence intervals around the point estimate.

*Siri,* Google's *Google Now* for Android phones, and Microsoft's *Cortana*). Each of the smartphone conditions impaired performance when compared with 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.

## 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.

### Method

**Participants.** Thirty-four participants were recruited for Experiment 2 (19 males, 15 females) using the same methods as Experiment 1. All data were collected from March 26th through April 19th of 2015. Participants ranged in age from 22 to 70 years old ($\bar{x}$ = 42.5). All eligibility requirements were identical to those used in Experiment 1. Participants reported between 4 and 52 years of driving experience ($\bar{x}$ = 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 U.S. 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. To adequately train participants on the text message functionality of each of the phones, they were required to send 6 practice text messages using the phone's digital assistant that was to be used in the forthcoming condition. Voice

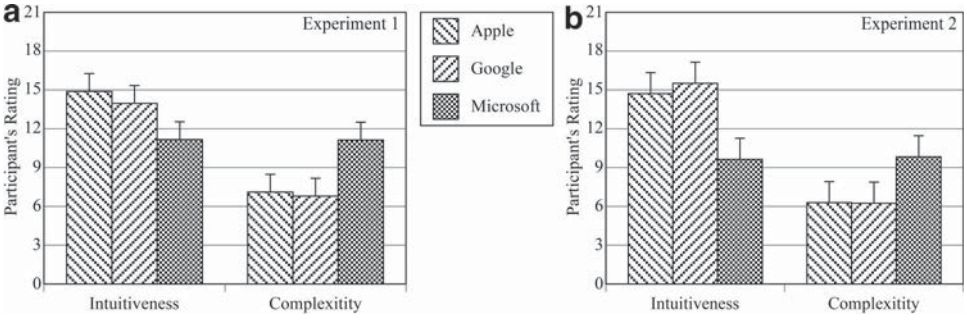


*Figure 4.* Mean ratings of intuitiveness and complexity for the Apple, Google, and Microsoft systems in Experiment 1 (a; left panel) and Experiment 2 (b; right panel). Error bars reflect 95% confidence intervals around the point estimate.

Table 1
*Cohen's d Values for the Pair-Wise DRT Differences in Experiments 1 and 2*

| Condition | Experiment | Apple-Off | Google-Off | MSFT-Off | Apple-On | Google-On | MSFT-On | OSPAN |
|---|---|---|---|---|---|---|---|---|
| Single-Task | Experiment1 | .87 | .66 | 1.16 | 1.42 | 1.45 | 1.77 | 1.98 |
| | Experiment2 | .84 | 1.02 | .84 | 1.60 | 1.75 | 1.90 | 1.93 |
| Apple-Off | Experiment1 | | .41 | .13 | .89 | .57 | | 1.23 |
| | Experiment2 | | .35 | .19 | 1.22 | .67 | .88 | .82 |
| Google-Off | Experiment1 | | | .35 | 1.09 | 1.13 | 1.22 | 1.69 |
| | Experiment2 | | | .18 | 1.39 | 1.38 | 1.83 | 1.63 |
| MSFT-Off | Experiment1 | | | | .89 | .69 | 1.24 | 1.37 |
| | Experiment2 | | | | 1.21 | 1.14 | 1.72 | 1.36 |
| Apple-On | Experiment1 | | | | | .17 | .05 | .31 |
| | Experiment2 | | | | | .16 | .11 | .03 |
| Google-On | Experiment1 | | | | | | .25 | .64 |
| | Experiment2 | | | | | | .12 | .20 |
| MSFT-On | Experiment1 | | | | | | | .34 |
| | Experiment2 | | | | | | | .08 |

*Note*. The first row in each cell is for RT differences in Experiment 1 and the second row in each cell is for RT differences in Experiment 2. Condition refers to the Single-Task, OSPAN, and Off-Task (e.g., Google-Off) and On-task (e.g., Google-On) performance for the Apple, Google, and Microsoft (MSFT) secondary-tasks. OSPAN = Operation Span.

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 whether they had any questions. Text messaging prompts were given in the same location as the task prompts in Experiment 1 and were as follows:

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, and ask him to start the meeting without you."
3. "Tell Anna Pearl your car is in the shop, and can she come pick you up."
4. "Ask Chris Hunter if he wants to eat out and 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 and 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

**DRT.** The RT and Hit Rate data for the DRT task are plotted as a function of secondary-task condition in Figures 1b and 2b, respectively. Like Experiment 1, these are denoted by a "-Off" for off task performance (e.g., not interacting with the digital voice assistant) and a "-On" for on-task performance (e.g., interacting with the digital voice assistant).

Table 2
*Cohen's d Values for the Pair-Wise DRT Differences in Experiments 1 and 2*

| Condition | Experiment | Apple-Off | Google-Off | MSFT-Off | Apple-On | Google-On | MSFT-On | OSPAN |
|---|---|---|---|---|---|---|---|---|
| Single-Task | Experiment 1 | .83 | .17 | .64 | .74 | .53 | .82 | 1.02 |
| | Experiment 2 | .57 | .50 | .59 | .60 | .68 | .66 | 1.01 |
| Apple-Off | Experiment 1 | | .50 | .10 | .42 | .01 | .37 | .77 |
| | Experiment 2 | | .33 | .01 | .35 | .34 | .47 | .47 |
| Google-Off | Experiment 1 | | | .37 | .84 | .50 | .96 | .97 |
| | Experiment 2 | | | .35 | .45 | .50 | .52 | .85 |
| MSFT-Off | Experiment 1 | | | | .38 | .08 | .54 | .83 |
| | Experiment 2 | | | | .35 | .53 | .40 | .64 |
| Apple-On | Experiment 1 | | | | | .44 | .01 | .43 |
| | Experiment 2 | | | | | .12 | .03 | .07 |
| Google-On | Experiment 1 | | | | | | .39 | .70 |
| | Experiment 2 | | | | | | .15 | .18 |
| MSFT-On | Experiment 1 | | | | | | | .56 |
| | Experiment 2 | | | | | | | .10 |

*Note*. The first row is for hit rate differences in Experiment 1 and the second row is for hit rate differences in Experiment 2. Condition refers to the Single-Task, OSPAN, and Off-Task (e.g., Google-Off) and On-task (e.g., Google-On) performance for the Apple, Google, and Microsoft (MSFT) secondary-tasks. OSPAN = Operation Span.

Table 3
*Cohen's d Values for the Pair-Wise NASA TLX Ratings in Experiments 1 and 2*

| Condition | Apple | Google | Microsoft | OSPAN |
|---|---|---|---|---|
| | | Experiment 1 | | |
| Single-Task | | | | |
| Mental | 1.08 | .62 | .59 | 2.40 |
| Physical | .73 | .76 | 1.02 | 1.05 |
| Temporal | .78 | .59 | .68 | 1.72 |
| Performance | .24 | .18 | .38 | .69 |
| Effort | .69 | .40 | .57 | 1.65 |
| Frustration | .74 | .48 | .79 | 1.40 |
| Apple | | | | |
| Mental | | .43 | .17 | 1.67 |
| Physical | | .05 | .23 | .12 |
| Temporal | | .10 | .00 | 1.30 |
| Performance | | .10 | .18 | .62 |
| Effort | | .31 | .02 | 1.28 |
| Frustration | | .25 | .25 | .88 |
| Google | | | | |
| Mental | | | .54 | 2.09 |
| Physical | | | .37 | .15 |
| Temporal | | | .12 | 1.51 |
| Performance | | | .28 | .66 |
| Effort | | | .36 | 1.51 |
| Frustration | | | .45 | 1.04 |
| Microsoft | | | | |
| Mental | | | | 1.25 |
| Physical | | | | .16 |
| Temporal | | | | 1.67 |
| Performance | | | | .44 |
| Effort | | | | 1.46 |
| Frustration | | | | .75 |
| | | Experiment 2 | | |
| Single-Task | | | | |
| Mental | 1.47 | .47 | .73 | 3.63 |
| Physical | .81 | .77 | .97 | 1.16 |
| Temporal | .90 | .84 | 1.30 | 1.84 |
| Performance | .75 | .75 | 1.04 | 1.40 |
| Effort | 1.21 | 1.08 | 1.68 | 3.19 |
| Frustration | .74 | .78 | 1.18 | 1.35 |
| Apple | | | | |
| Mental | | .09 | .38 | 1.78 |
| Physical | | .07 | .47 | .55 |
| Temporal | | .08 | .47 | .91 |
| Performance | | .11 | .50 | .88 |
| Effort | | .18 | .40 | 1.32 |
| Frustration | | .04 | .63 | .71 |
| Google | | | | |
| Mental | | | .34 | 1.60 |
| Physical | | | .52 | .54 |
| Temporal | | | .51 | 1.16 |
| Performance | | | .54 | 1.12 |
| Effort | | | .60 | 2.06 |
| Frustration | | | .60 | .68 |
| Microsoft | | | | |
| Mental | | | | 1.00 |
| Physical | | | | .04 |
| Temporal | | | | .53 |
| Performance | | | | .47 |
| Effort | | | | .87 |
| Frustration | | | | .08 |

*Note*. OSPAN = Operation Span.

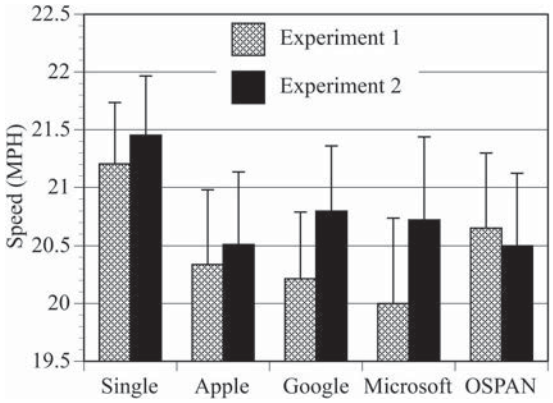


*Figure 5.* Average driving speed (in MPH) for the 5 conditions in Experiment 1. Error bars reflect 95% confidence intervals around the point estimate.

***Reaction time.*** The RT data from the DRT task were analyzed using a ANOVA with a 3 (Age Group: ages 21–34, ages 35–53, and ages 54–70) × 8 (Condition: Single-task, Apple-Off, Google-Off, Microsoft-Off, Apple-On, Google-On, Microsoft-On, 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 × 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-Off vs. Google-Off, $p = .070$; Apple-Off vs. Microsoft-Off, $p = .392$; Google-Off vs. Microsoft-Off, $p = .189$). Each of these off-task conditions differed from their respective on-task performance (Apple-Off vs. Apple-On, $p < .001$; Google-Off vs. Google-On, $p < .001$; Microsoft-Off vs. Microsoft-On, $p < .001$). Importantly, none of the on-task secondary tasks differed significantly from the OSPAN condition ($p = .805$, $p = .297$, and $p = .569$ for Apple-On, Google-On, and Microsoft-On, respectively). Finally, the on-task performance for the three conditions did not

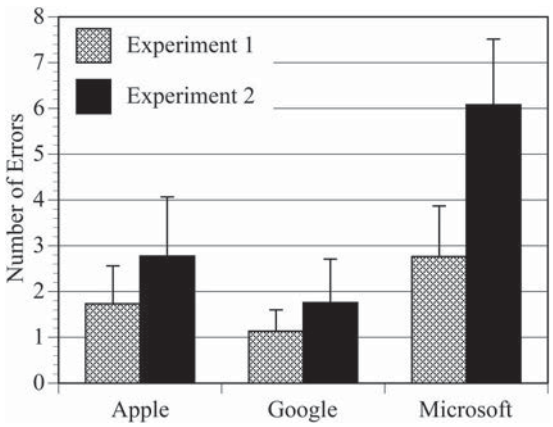


*Figure 6.* Average number of errors experienced by participants for the Apple, Google, and Microsoft systems in Experiments 1 and 2. Error bars reflect 95% confidence intervals around the point estimate.

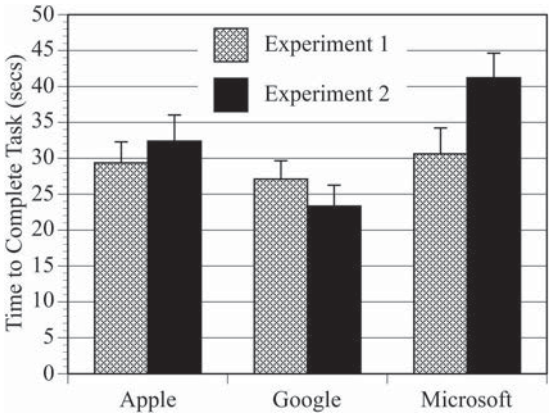


*Figure 7.* Average time to complete the secondary-tasks for the Apple, Google, and Microsoft systems in Experiments 1 and 2. Error bars reflect 95% confidence intervals around the point estimate.

differ from each other (Apple-On vs. Google-On, $p = .365$; Apple-On vs. Microsoft-On, $p = .411$; Google-On vs. Microsoft-On, $p = .612$).

***Hit rate.*** The Hit Rate data from the DRT task were analyzed using a ANOVA with a 3 (Age Group) × 8 (Condition) 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 × 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-Off vs. Google-Off, $p = .055$; Apple-Off vs. Microsoft-Off, $p = .913$; Google-Off vs. Microsoft-Off, $p = .052$). The off-task conditions differed from on-task performance for Google-Off versus Google-On ($p = .050$), and Microsoft-Off versus Microsoft-On ($p = .002$), but not for Apple-Off versus Apple-On ($p = .100$). Importantly, none of the on-task secondary-task conditions differed significantly from the OSPAN condition ($p = .821$, $p = .595$, and $p = .817$ for Apple-On, Google-On, and Microsoft-On, respectively). Finally, the on-task performance for the three smartphone conditions did not differ from each other (Apple-On vs. Google-On, $p = .741$; Apple-On vs. Microsoft-On, $p = .946$; Google-On vs. Microsoft-On, $p = .635$).

**NASA TLX.** The 6 scales of the NASA TLX, presented in Figure 3b, were analyzed using a MANOVA with a 3 (Age Group) × 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 Age × Condition 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* subscale 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* subscale 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 Age × Condition interaction were significant, $F(8, 124) = 1.32$, $p = .241$, $\eta^2 = .078$. The *temporal* subscale 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 Age × Condition interaction were significant, $F(8, 124) = 1.98$, $p = .054$, $\eta^2 = .113$. The *performance* subscale increased as a

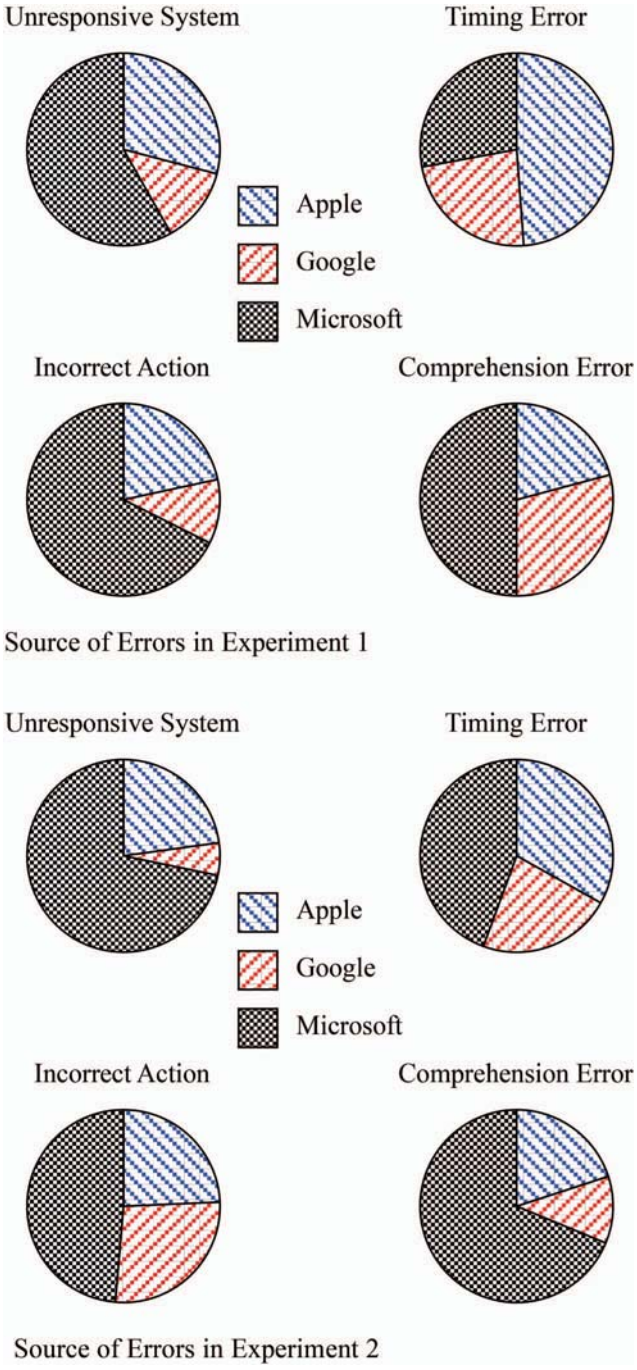


*Figure 8.* Relative proportion of errors by category for the Apple, Google, and Microsoft systems in Experiment 1 (upper panel) and Experiment 2 (lower panel). See the online article for the color version of this figure.

Table 4
*Cohen's d Values for the Pair-Wise Differences in Intuitiveness and Complexity Obtained in Experiments 1 and 2*

| | | Intuitiveness | | Complexity | |
|---|---|---|---|---|---|
| Condition | Experiment | Google | Microsoft | Google | Microsoft |
| Apple | Experiment 1 | .22 | .47 | .08 | .57 |
| | Experiment 2 | .17 | .79 | .02 | .61 |
| | Experiment 1 | | .33 | | .65 |
| Google | Experiment 2 | | .88 | | .69 |

*Note*. The first row in each cell is for Experiment 1, and the second row in each cell is for Experiment 2.

Table 6
*Cohen's d Values for the Pair-Wise Differences in Error Count and Task Completion Time Obtained in Experiments 1 and 2*

| | | Error count | | Task completion time | |
|---|---|---|---|---|---|
| Condition | Experiment | Google | Microsoft | Google | Microsoft |
| Apple | Experiment 1 | .28 | .25 | .29 | .08 |
| | Experiment 2 | .29 | .58 | .79 | .64 |
| Google | Experiment 1 | | .49 | | .36 |
| | Experiment 2 | | .88 | | 1.46 |

*Note*. The first row in each cell is for Experiment 1, and the second row is for Experiment 2.

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 Age × Condition interaction were significant, $F(8, 124) = 1.07$, $p = .391$, $\eta^2 = .064$. The *effort* subscale 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 Age × Condition interaction were significant, $F(8, 124) = 1.40$, $p = .203$, $\eta^2 = .083$. Finally, the *frustration* subscale 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 Age × Condition interaction were significant, $F(8, 124) = 1.01$, $p = .432$, $\eta^2 = .061$.[11]

**Intuitiveness and complexity.** Participants were also asked to rate how intuitive, usable, and easy it was to use the different smartphones' intelligent personal assistants. They also rated how complex, difficult, and confusing it was to use the different smartphones' intelligent personal assistants. Figure 4b 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) × 3 (Condition: Apple, Google, Microsoft) split-plot ANOVA 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 × 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, $p < .031$).

Table 5
*Cohen's d Values for the Pair-Wise Vehicle Speed in Experiments 1 and 2*

| Condition | Experiment | Apple | Google | Microsoft | OSPAN |
|---|---|---|---|---|---|
| Single-Task | Experiment 1 | .51 | .57 | .59 | .52 |
| | Experiment 2 | .84 | .45 | .41 | .76 |
| Apple | Experiment 1 | | .05 | .14 | .14 |
| | Experiment 2 | | .24 | .08 | .00 |
| Google- | Experiment 1 | | | .09 | .22 |
| | Experiment 2 | | | .15 | .24 |
| Microsoft | Experiment 1 | | | | .26 |
| | Experiment 2 | | | | .08 |

*Note*. The first row is for speed differences in Experiment 1, and the second row is for speed differences in Experiment 2.

***Complexity.*** A 3 (Age Group) × 3 (Condition: Apple, Google, Microsoft) split-plot ANOVA 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.** A video analysis of participants' interactions was performed to determine the vehicle speed (see Figure 5), the number of observed errors (see Figure 6), and the time to complete the task (see Figure 7). The relative frequency of the four error categories for each of the smartphones is provided in Figure 8b.

***Vehicle speed.*** A 3 (Age Group) × 3 (Condition: Apple, Google, Microsoft) split-plot ANOVA 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 × 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) × 3 (Condition: Apple, Google, Microsoft) split-plot ANOVA 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 × Condition interaction were 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) × 3 (Condition: Apple, Google, Microsoft) split-plot ANOVA found that the time to complete the tasks differed as a function of Condition, $F(2,$

[11] Effect size estimates for the pairwise differences in NASA TLX ratings are presented in Table 3.

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 × Condition interaction were significant, $F(4, 62) = 0.28$ $p = .891$, $\eta^2 = .018$. All of the pairwise 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 whether the pattern obtained in the two experiments differed in any substantive way. For the analysis of the DRT data, a 2 (Experiment) × 3 (Age Group: ages 21–34, ages 35–53, and ages 54–70) × 8 (Condition: Single-task, Apple-Off, Google-Off, Microsoft-Off, Apple-On, Google-On, Microsoft-On, and OSPAN) split-plot ANOVA 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 × Age interaction, $F(2, 59) = 1.23$, $p = .301$, $\eta^2 = .040$, nor the Experiment × Condition interaction, $F(7, 413) = 1.74$, $p = .098$, $\eta^2 = .029$, nor the Experiment × Age × 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 × Age interaction, $F(2, 59) = 2.86$, $p = .066$, $\eta^2 = .088$, nor the Experiment × Condition interaction, $F(7, 413) = 1.71$, $p = .104$, $\eta^2 = .028$, nor the Experiment × Age × Condition interaction, $F(14, 413) = 1.65$, $p = .065$, $\eta^2 = .053$, were significant. The overall pattern obtained with the DRT in the two experiments was virtually identical.

A MANOVA compared the six NASA TLX measures using a 2 (Experiment) × 3 (Age Group) by 5 (Condition: Single-task, Apple, Google, Microsoft and OSPAN) split-plot design. The MANOVA revealed that neither the main effect of Experiment, $F(6, 54) = 2.16$, $p = .062$, $\eta^2 = .193$, nor the Experiment × Age interaction, $F(12, 110) = 1.67$, $p = .083$, $\eta^2 = .154$, nor the Experiment × Age × Condition interaction, $F(48, 1416) = 1.09$, $p = .315$, $\eta^2 = .036$ were significant. However, the Experiment × Condition interaction was significant, $F(24, 936) = 2.39$, $p < .001$, $\eta^2 = .058$ and reflected the slightly higher workload ratings for the IVIS condition in Experiment 2 (i.e., the average TLX rating for the smartphone conditions in Experiment 2 was of 0.73 higher than that obtained on Experiment 1 on a 21-point scale). On the whole, the pattern obtained in the subjective workload obtained in the two experiments was very similar with the caveat that, as anticipated, voice-texting was more demanding than placing a call or selecting music.

A MANOVA compared the intuitiveness and complexity measures using a 2 (Experiment) × 3 (Age Group) by 3 (Condition: Apple, Google, and Microsoft) split-plot design. The MANOVA revealed that neither the main effect of Experiment, $F(2, 58) = 0.71$, $p = .496$, $\eta^2 = .024$, nor the Experiment × Age interaction, $F(4, 118) = 0.25$, $p = .909$, $\eta^2 = .008$, nor the Experiment × Condition interaction $F(4, 236) = 1.95$, $p = .107$, $\eta^2 = .032$, nor the Experiment × Age × Condition interaction, $F(8, 236) = 0.82$, $p = .590$, $\eta^2 = .027$ were significant. In terms of the subjective measures of intuitiveness and complexity, 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.[12] To evaluate this residual cost in more detail, DRT performance in the off-task segments of the drive were sorted into 3-s 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 9 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 Figure 9, "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:

$$f(x) = a \times (x^{-.1837}),\ \textit{Where } a = exp(6.697),\ \textit{With } R^2 = .97$$

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 and 2) 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 a notable effect given that the actual time to complete the tasks, approximately 32 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 9 are striking 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 that just because a driver terminates a call or text message does not mean that they are no longer impaired.

## 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 head-

[12] The on-task interval averaged 32 seconds whereas the off-task interval averaged 37 seconds.

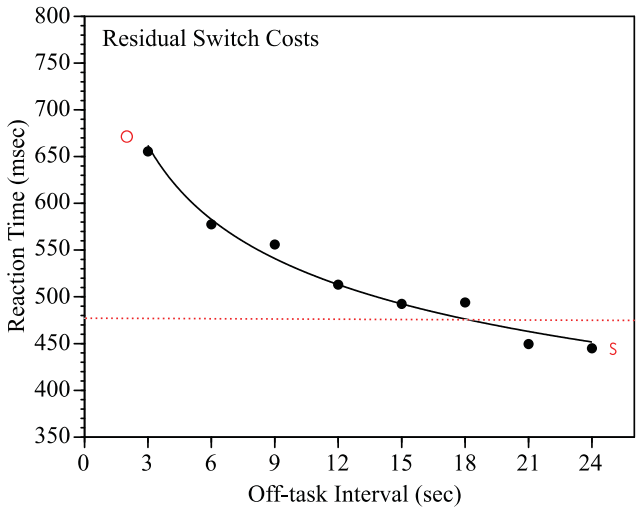


*Figure 9.* Residual switch costs in transitioning from on-task to off-task performance. The red "O" indicates average OSPAN RT from the DRT task; the red "S" indicates the average single-task RT from the DRT task. Off-task performance is distributed into 3-s intervals (relative to when the on-task activity terminated). The solid line represents the best fitting power function relating transition from on-task to single-task levels of performance. The dotted 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. See the online article for the color version of this figure.

phone (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 training 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.

## 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. The standardized score 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 nondistracted 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 with 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 participants 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 participants.

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 nondistracted single-task condition had a rating of 1.0 and the OSPAN task had a rating of 5.0

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

The cognitive workload scale for the different conditions is presented in Figure 10. 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, Google was 3.3, and Microsoft was 4.1.

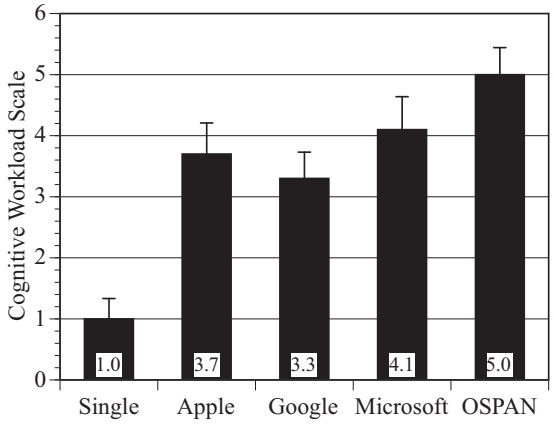


*Figure 10.* The cognitive workload scale for the Apple, Google, and Microsoft systems compared with single-task (category 1) and OSPAN (category 5). Error bars reflect 95% confidence intervals around the point estimate. OSPAN = Operation Span.

The error bars represent 95% confidence intervals and document that the Google system was associated with a lower workload rating that the Apple and Microsoft systems, which did not significantly differ. The data indicate that voice-based systems can have unintended consequences that result in high levels of cognitive workload.

Figure 11 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).

## Why Do Voice-Based Smartphones Interactions Increase motorists' Cognitive Workload?

The level of cognitive workload that we observed may be surprising given the automotive industry's push to add voice-based capabilities to their new vehicles (including the capability to pair a smartphone via bluetooth to the vehicle infotainment system). Indeed, our market survey has found that virtually every major automaker supports some form of voice-based interaction in the vehicle (Cooper & Strayer, in press). The current research found that some of the workload associated with smartphone use was linked to the complexity and intuitiveness of the voice-based interactions; systems that were rated higher in complexity and lower in intuitiveness were associated with higher levels of cognitive workload. Even so, the workload associated with smartphone interactions was significantly higher than talking to a person on the smartphone. In the following paragraphs, we explore some of the bases for these differences.

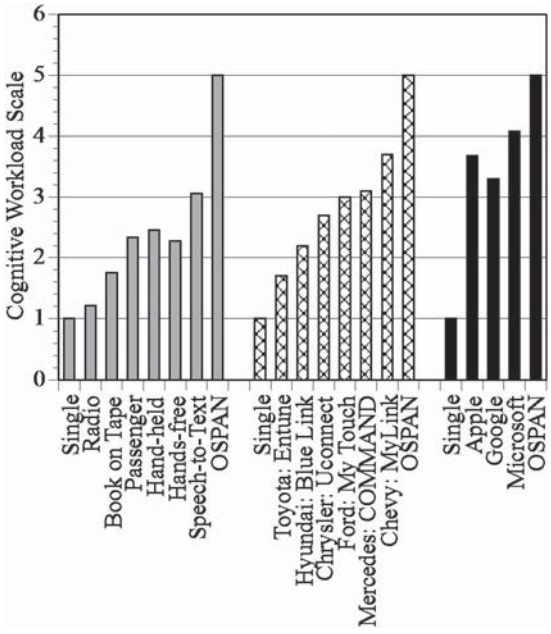


*Figure 11.* The workload scale for Strayer et al. (2013, the gray bars), Cooper et al. (2014, the hatched bars), and the current research (black bars). OSPAN = Operation Span.

One difference between interactions using a smartphone and interactions with a human is that the former involves computerized speech recognition and computerized speech generation, whereas the latter involves interactions with a person (i.e., the difference between talking to a computer vs. talking to a person). Our prior evaluation of state-of-the-art speech generation systems found that the mental workload associated with comprehending *computerized speech* was equivalent to the mental workload associated with comprehending *human speech* (assuming the same content, see Strayer et al., 2014); however, the comprehension aspects of speech were associated with lower levels of workload than the production aspects of speech. Using a computer-based voice-recognition system with 100% reliability, we found that the mental workload associated with speech comprehension was about half of that observed with speech production (Strayer et al., 2014).

Moreover, if the voice-based interactions are error-prone, as they often are with smartphone interactions (cf. Figure 8), the workload experienced by the driver significantly increases. Unlike human language where pauses in speech are part of the natural ebb and flow of a conversation (e.g., Drews, Pasupathi, & Strayer, 2004, 2008), computerized speech recognition systems have difficulty parsing a sentence with pauses. Several of the smartphone system errors we observed involved situations where the driver paused their dictation in midsentence to handle a tricky aspect of the driving task. In these circumstances, the smartphone interaction failed whereas adult interlocutors more often adapt to the pauses in the conversation. The added time pressure induced by the smartphone likely increases the attentional burden especially if drivers attempt to load and hold their sentence in working memory until they can output the speech in one continuous utterance.

If this "load and hold" strategy for interacting with smartphones is correct, the effect may be similar to the loading effects observed in the prospective memory literature (e.g., Heathcote, Loft, & Remington, 2015). The prospective memory task requires participants to perform one block of primary task trials (e.g., lexical decision) and another block of trials in which they are asked to perform that same primary task but also to perform a deferred action if a target is detected in the future (e.g., report if a particular sequence of letters was presented in the letter string). The typical pattern in this task is that participants respond slower on the primary task with the additional prospective memory requirements. This delay is thought to be attributable to the additional burden placed on working memory that changes the information processing dynamics of the primary task (e.g., Heathcote, Loft, & Remington, 2015). If smartphone interactions place an additional burden on working memory over and above voice-based interactions with another person, future research should be able to model the data and show that the same parameters of information processing are at work.[13]

[13] This modeling effort requires at least an order of magnitude more data than were collected in the current study for the modeling parameters to become stable.

The finding that smartphone interactions interfere with the operation of a motor vehicle also has implications for our theoretical understanding of dual-task interactions. Driving is predominately a visual/spatial/manual task whereas hands-free smartphone interaction is predominately an auditory/verbal/vocal task. Consequently, multiple-resource models (e.g., Wickens, 1980, 1984) would seem to predict little or no interference when the two tasks are performed concurrently, because there is little overlap in the attentional resources used to perform them. However, Bergen et al. (2014) suggests that language interferes with driving because of modality-specific mental simulation (i.e., cross-talk). Interference is observed when the mental representations of two tasks overlap. Importantly, Bergen et al. (2014) found that if the language task did not lend itself to mental simulation that the interference on the driving task diminished.

It is newsworthy that the workload measured with the DRT for on-task voice texting was essentially equivalent to that obtained in the memory-demanding OSPAN task. The well-intentioned attempt to allow motorists to send and receive text messages while keeping their eyes on the road and their hands on the wheel appears to have the unintended consequence of driving cognitive workload to excessively high levels. The apparent cross talk between the concurrent tasks, suggests that they are in competition for the same limited capacity resource. We hypothesize that driving-related activities and other nondriving tasks place competing demands on the limited capacity working memory system.

## Threaded Cognition and Smartphone Interactions

Using ACT-R, Salvucci (2006; see also Salvucci & Taatgen, 2008) developed a threaded cognition of model to predict driving behavior highlighting the fact that ACT-R has built in perceptual and motor modules that works in parallel resembling human behavior. In addition, a cognitive processor that receives information from the perceptual module and is in charge of all that goes to the motor module. Although these input/output modules operate in parallel, the cognitive processor operates sequentially, thereby servicing only one quantized thread at any time. In this model, the cognitive processor operates on a first-in, first-out basis. Salvucci (2006) argued that when drivers engage in secondary tasks, the cognitive processor must switch between the secondary tasks and driving, which results in suboptimal driving performance.

The pattern of data reported in this article is consistent with a refinement of threaded cognition wherein the disparate task threads have different processing priority. Notably, we have observed a general pattern where driving-related activities have a higher priority than nondriving threads. Indeed, drivers tend to give short shrift to the DRT task when they were making turns; hence the ISO DIS 17488 (2015) stipulation to not include segments of the drive with turns in the DRT analyses. Drivers also pause their conversation with passengers during difficult sections of the drive (e.g., Drews, Pasupathi, & Strayer, 2004, 2008). Consequently, when more than one thread requires processing, threads unrelated to driving tend to be delayed by driving-related threads (i.e., driving is prioritized over other secondary tasks). However, processing of the driving-related threads is delayed when the serial cognitive processor is occupied with a nondriving operation, thereby providing situations where driving is impaired by secondary-task interactions. In this circumstance, drivers may miss or react slowly to imperative events in the driving environment (e.g., drivers engaged in a secondary task may fail to detect a changing traffic light or other hazard on the roadway). Moreover, responses to the DRT tend to be delayed by processing of *both* driving-related activities and smartphone interactions, suggesting an even lower processing priority for the simple DRT interactions. That is, the DRT is processed after the driving and smartphone interactions have been serviced and in many instances the workload of these two activities is such that the driver fails to respond to the DRT, resulting in the pattern of miss rates observed in the current research. Note, however, that the DRT can impose a small cost on driving if the cognitive processor is servicing the DRT thread when an imperative driving-related event occurs (i.e., during the serial lockout of the cognitive processor, processing of driving-related information is slightly delayed; a pattern that was reported by Strayer et al., 2013).

The workload associated with sending a text message, as indexed by the on-task DRT data, was essentially the same as that observed in the continuous OSPAN task, suggesting that these voice-based interactions place a surprisingly high demand on the cognitive processor, effectively locking out the processing of other activities. We also observed a significant residual cost following the on-task interaction (see Figure 9). Using the threaded cognition model, we interpret the residual costs in the context of reacquiring situation awareness that was lost during the smartphone interaction (e.g., Strayer & Fisher, 2016). That is, the smartphone interactions lock out the processing of the driving-related threads, thereby diminishing the driver's situation awareness. Under this interpretation, the residual costs reflect the servicing these driving-related threads once the smartphone interaction has completed. This predicts a significant increase over baseline in driving-related activities associated with good situation awareness (e.g., an increased scanning of mirrors and other peripheral locations), a pattern observed elsewhere in the context of other supervisory control tasks (e.g., Gartenberg, Breslow, McCurry, & Trafton, 2014).

## 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 with 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 nearly identical for placing calls, selecting music, and the seemingly more demanding activity of sending of text messages (i.e., the only differences were observed in slightly higher levels of subjective workload with voice-texting). 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. In particular, the Google system outperformed the Apple and Microsoft systems. Our analysis found that this difference was directly related to the number of system errors and the intuitive-

ness/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 with 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 persistent 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 remarkable 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 also be related to the driver reestablishing situation awareness of the driving environment that was lost during the smartphone interaction (Fisher & Strayer, 2014; Strayer & Fisher, 2016; Strayer, in press).

## Summary and Conclusions

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.

## Résumé

Le but de cette recherche consistait à examiner, au moyen de trois différents assistants personnels (*Siri* de Apple, *Google Now* de Google pour téléphone Androïde et *Cortana* de Microsoft), l'impact d'interactions vocales sur la charge de travail cognitive du conducteur. À l'aide de deux expériences employant un véhicule instrumenté sur des routes de banlieue, nous avons mesuré la charge de travail cognitive de conducteurs alors qu'ils utilisaient les fonctionnalités vocales de chacun des téléphones intelligents pour effectuer un appel, sélectionner de la musique ou envoyer un message texte. La charge de travail cognitive a pu être déterminée après évaluation de la performance de la tâche principale par analyse-vidéo, de la performance de la tâche secondaire par tâche de détection-réponse (DRT) puis, de la charge de travail mentale subjective. Nous avons constaté que la charge de travail y était nettement plus élevée que celle associée à la tâche simple de conduire. Il y avait aussi des différences systématiques entre les téléphones intelligents. Le système Google était moins demandant cognitivement sur le conducteur que les systèmes Apple et Microsoft, lesquels avaient le même effet. L'analyse vidéo a montré que la différence au niveau de la charge de travail mentale entre téléphones intelligents était associée au nombre d'erreurs de système, à la durée de temps requise pour mener à bien une action et à la complexité et à l'intuitivité des appareils. Finalement, des niveaux étonnamment élevés de charge de travail cognitive ont été observés lorsque les conducteurs étaient en interaction avec leurs appareils : Les mesures de la charge de travail associée à la concentration sur une tâche ne différaient pas systématiquement de celles associées à une tâche (OSPAN) exigeante sur le plan mental. L'analyse a aussi révélé la présence de coûts résiduels associée à l'utilisation de chacun des téléphones intelligents, lesquels ont pris un temps considérable pour se dissiper. Les données suggèrent que la prudence est de mise en ce qui a trait à l'utilisation de technologie vocale sur téléphone intelligent dans un véhicule étant donné les niveaux élevés de la charge de travail cognitive associée à ces interactions.

*Mots-clés* : distraction cognitive, charge de travail cognitive, attention divisée, conduite, multitâches

## References

Bergen, B., Medeiros-Ward, N., Wheeler, K., Drews, F., & Strayer, D. L. (2014). The crosstalk hypothesis: Why language interferes with driving. *Journal of Experimental Psychology: General, 142,* 119–On30. http://dx.doi.org/10.1037/a0028428

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*. Washington, DC: 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*. Washington, DC: AAA Foundation for Traffic Safety.

Cooper, J. M., & Strayer, D. L. (in press). An inventory of voice-based technology in new automobiles. Manuscript submitted for publication.

Drews, F. A., Pasupathi, M., & Strayer, D. L. (2004). Passenger and cell-phone conversations in simulated driving. *Proceedings of the 48nd Annual Meeting of the Human Factors and Ergonomics Society* (pp. 2210–2212).

Drews, F. A., Pasupathi, M., & Strayer, D. L. (2008). Passenger and cell phone conversations in simulated driving. *Journal of Experimental Psychology: Applied, 14,* 392–400. http://dx.doi.org/10.1037/a0013119

Endsley, M. R. (1995). Towards a theory of situation awareness in dynamic systems. *Human Factors, 37,* 32–64. http://dx.doi.org/10.1518/001872095779049543

Endsley, M. R. (2015). Situation awareness misconceptions and misunderstandings. *Journal of Cognitive Engineering and Decision Making, 9,* 4–32. http://dx.doi.org/10.1177/1555343415572631

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour, 8,* 97–120. http://dx.doi.org/10.1016/j.trf.2005.04.012

Fisher, D. L., & Strayer, D. L. (2014). Modeling situation awareness and crash risk. *Annals of Advances in Automotive Medicine, 58,* 33–39.

Gartenberg, D., Breslow, L., McCurry, J. M., & Trafton, J. G. (2014). Situation awareness recovery. *Human Factors, 56,* 710–727. http://dx.doi.org/10.1177/0018720813506223

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload. Amsterdam,* the Netherlands: North-Holland Press. http://dx.doi.org/10.1016/S0166-4115(08)62386-9

Hartley, A. A., & Little, D. M. (1999). Age-related differences and similarities in dual-task interference. *Journal of Experimental Psychology: General, 128,* 416–449. http://dx.doi.org/10.1037/0096-3445.128.4.416

Heathcote, A., Coleman, J. R., Eidels, A., Watson, J. M., Houpt, J., & Strayer, D. L. (2015). Working memory's workload capacity. *Memory & Cognition, 43,* 973–989. http://dx.doi.org/10.3758/s13421-015-0526-2

Heathcote, A., Loft, S., & Remington, R. W. (2015). Slow down and remember to remember! A delay theory of prospective memory costs. *Psychological Review, 122,* 376–410. http://dx.doi.org/10.1037/a0038952

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.

Kahnemann, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

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.

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.

NHTSA. (2012). Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices. Department of Transportation. Docket No. NHTSA-2010-Off053.

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. http://dx.doi.org/10.1016/j.aap.2011.04.008

Regan, M. A., & Strayer, D. L. (2014). Towards an understanding of driver inattention: Taxonomy and theory. *Annals of Advances in Automotive Medicine, 58,* 5–14.

Rogers, R. D., & Monsell, S. (1995). The cost of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124,* 207–231. http://dx.doi.org/10.1037/0096-3445.124.2.207

Salvucci, D. D. (2006). Modeling driver behavior in a cognitive architecture. *Human Factors, 48,* 362–380. http://dx.doi.org/10.1518/001872006777724417

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review, 115,* 101–130. http://dx.doi.org/10.1037/0033-295X.115.1.101

Strayer, D. L. (2015). Is the technology in your car driving you to distraction. *Policy Insights from Behavioral and Brain Sciences, 2,* 156–165. http://dx.doi.org/10.1177/2372732215600885

Strayer, D. L. (In Press). Attention and driving. In J. Fawcett, E. F. Risko, & A. Kingstone (Eds.), *The handbook of attention*. Cambridge, MA: 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*. Washington, DC: AAA Foundation for Traffic Safety.

Strayer, D. L., & Fisher, D. L. (2016). SPIDER: A framework for understanding driver distraction. *Human Factors, 58,* 5–12. http://dx.doi.org/10.1177/0018720815619074

Strayer, D. L., & Johnston, W. A. (2001). Driven to distraction: Dual-Task studies of simulated driving and conversing on a cellular telephone. *Psychological Science, 12,* 462–466. http://dx.doi.org/10.1111/1467-9280.00386

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*. Washington, DC: AAA Foundation for Traffic Safety.

Strayer, D. L., Turrill, J., Cooper, J. M., Coleman, J., Medeiros-Ward, N., & Biondi, F. (2015). Assessing cognitive distraction in the automobile. *Human Factors, 53,* 1300–1324.

Watson, J. M., & Strayer, D. L. (2010). Supertaskers: Profiles in extraordinary multitasking ability. *Psychonomic Bulletin & Review, 17,* 479–485. http://dx.doi.org/10.3758/PBR.17.4.479

Wickens, C. D. (1980). The structure of attentional resources. In R. S. Nickerson (Ed.), *Attention and performance VIII* (pp. 239–257). Hillsdale, NJ: Erlbaum.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of attention* (pp. 63–101). New York, NY: Academic Press.

Received May 3, 2016
Accepted September 10, 2016 ■