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

Mental Workload of Common Voice-Based Vehicle Interactions across Six Different Vehicle Systems

October 2014

AAA Foundation for Traffic Safety

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

Title

Mental Workload of Common Voice-Based Vehicle Interactions across Six Different Vehicle Systems *(October 2014)*

Authors

Joel M. Cooper, Hailey Ingebretsen, and David L. Strayer

About the Sponsor

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

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

Executive Summary

In this report we present the results from a driving study that evaluated the mental demands of simple auditory-vocal vehicle commands using five 2013 and one 2012 model year OEM infotainment systems. Participants in this research completed a series of voice-based music functions and phone dialing tasks while driving an on-road course. Each participant drove six vehicles on a seven – nine minute loop through a residential neighborhood in which they were periodically instructed to dial a 10 digit number, call a contact, change the radio station, or play a CD. All interactions took place using "hands-free" voice systems which were activated with the touch of a button on the steering wheel. Evaluated systems included: A Ford equipped with MyFord Touch, a Chevrolet equipped with MyLink, a Chrysler equipped with Uconnect, a Toyota equipped with Entune, a Mercedes equipped with COMAND, and a Hyundai equipped with Blue Link. Mental workload was also assessed in a single-task baseline drive and during a demanding mental math task, which respectively formed the low and high workload baselines.

Across these eight conditions (6 OEM systems interactions + low and high workload baselines), measures of mental workload were derived from reaction time, subjective assessments, and heart rate. Reaction time measures were recorded using a standard stimulus response task. Results indicated that reaction time slowed reliably and in a consistent manner indicative of sensitivity to changes in task difficulty. Subjective mental workload was assessed using a common task load index. Similar to reaction time performance, results from the subjective workload ratings were highly sensitive to differences in the experimental conditions. Heart rate measures were also recorded. Results indicated that Heart Rate typically increased in a manner consistent with increases in task difficulty. However, Heart Rate was less sensitive to changes in workload than both the reaction time and subjective workload measures.

Following the process developed by Strayer et al. (2103), task performance measures were combined into a single workload metric. Results indicated that there were significant differences in the amount of cognitive attention required to complete voice interactions with each of the different vehicle systems. In the best case, we found that music functions and voice/contact dialing using Toyota's Entune system imposed modest additional demands over the single-task baseline, whereas those same activities using Chevy's MyLink imposed cognitive load that approached the demanding mental math task. Not surprisingly, the most critical element of mental workload appeared to be the duration of the interaction, of which the primary contributing factors were the number of steps required to complete the task as well as the number of comprehension errors that arose during the interaction.

A comparison of the results from the current study to the Strayer et al. study (2013) reveals that the average cognitive demand of voice interactions with actual vehicle systems is similar to that imposed by the speech-to-text mockup used by Strayer et al. (2013). That is, on the standardized rating scale, a mean demand score of approximately 3 out of 5 was observed. This indicates that common voice tasks are generally more demanding than natural conversations, listening to the radio, or listening to a book on tape. However, the data also indicate that, if designed well, these same basic commands can be completed with little error in very few steps, leading to little additional cognitive demand.

Background

Driving is a highly complex activity that requires a significant amount of visual and cognitive attention to be performed successfully. In order to allow drivers to maintain their eyes on the driving task, nearly every vehicle sold in the US and Europe can now be optionally equipped with a hands-free voice system. Using structured voice commands, drivers can access functions as varied as contact and number dialing, music selection, destination entry, and even climate adjustment. Hands-free, voice activated convenience features are a natural development in vehicle safety. Yet, a sizable body of literature cautions that even auditory-vocal tasks may lead to unexpected task demands (Delogu, Conte, & Sementina, 1998; Harbluk & Lalande, 2005; Paris et al., 1995; Recarte & Nunes, 2007). Research on cognitive distraction suggests that even if a driver's eyes remain on the forward roadway, his or her ability to detect and respond to targets within the visual field may be impaired if cognitive focus is not also on the forward roadway (Hyman, et al. 2010; Simons, 2000; Strayer & Drews, 2007). Furthermore, prior research suggests that the cognitive demands associated with speech-to-text system interactions may be higher than other common auditory-vocal tasks such as talking on a cell phone or listening to the radio (Strayer et al. 2013 & 2014). However, prior research has almost exclusively looked at voice based interactions with systems that differ in some way from those used in actual vehicles. Thus, it is not clear whether or how estimates of cognitive task load obtained using synthetic systems might apply to real world systems.

Unlike visual attention, which can be directly observed, shifts in cognitive attention may be nearly impossible to notice, especially if they are not accompanied by some secondary task manifestation, such as speech production. Due to the self-evident nature of visual attention, many reliable measures of visual demand have been developed and refined, and these have in turn led to clear guidelines for reducing visual distractions in vehicles (e.g., NHTSA, 2012). On the flipside, the elusiveness of cognitive attention has made reliable measures more challenging to develop and interpret. Indeed, measures of driving behavior under cognitive load can be very subtle and at times even counterintuitive. This nuanced complexity has made the development of cognitive distraction metrics significantly more challenging.

One ongoing effort to understand cognitive distraction in vehicles is being led by Strayer and colleagues (see Strayer et al., 2013, and Strayer et al., 2014). In their research, Strayer et al. (2013) investigated a comprehensive set of common cognitive tasks across various data collection environments, using a complementary set of primary, secondary, physiological, and subjective measures. Through the use of a consistent protocol, Strayer at al. (2013) completed a laboratory, simulator, and on road assessment of common auditory-vocal tasks performed by drivers. This allowed a variety of everyday secondary cognitive driving tasks to be directly compared. Results indicated that the tasks could be roughly clustered into three distinct groups (See Figure 1). This clustering is based on both statistical and practical task differentiation. Listening to the radio or a book on tape led to low levels of cognitive demand, similar to baseline driving. Conversation, whether with a passenger or through a hand-held or hands-free cellular phone led to slightly elevated levels of mental workload, comprising the second group. Finally, a synthetic speech-to-text email interaction system led to still more elevated levels of mental workload, forming a separate workload category that was greater than the other two.

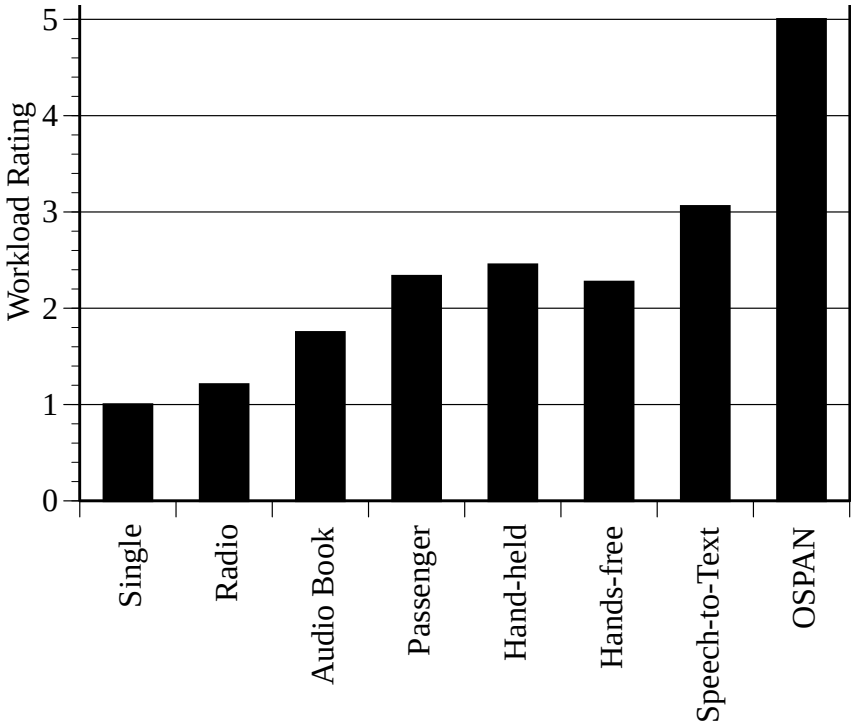

Figure 1. The Workload Rating Scale developed by Strayer et al. (2013).

Overall, the Strayer et al. (2013) study accomplished three objectives. First, it developed a method for the standardized measurement and analysis of cognitive demand. Second, it established that cognitive demands vary between commonly performed mental tasks. Third, it highlighted the need to better understand auditory-vocal vehicle interactions, especially as they relate to real-world systems that are currently available to consumers.

Building on the research reported by Strayer et al. (2013), Strayer et al. (2014) investigated the relative cognitive demands of speech production vs. speech comprehension, synthetic vs. natural speech, and error free vs. error prone systems. Results showed that speech production is significantly more demanding than speech comprehension and that system errors increase cognitive workload (See Figure 2). However, natural speech does not appear to directly confer a benefit over synthetic speech. In order to create the most broadly applicable results, the speech-to-text tasks evaluated by Strayer et al. (2013) and Strayer et al. (2014) were carefully scripted and controlled using functional mock-ups rather than actual systems. Thus, it is unknown how the cognitive demands reported in their research might compare to similar tasks using actual vehicle systems.

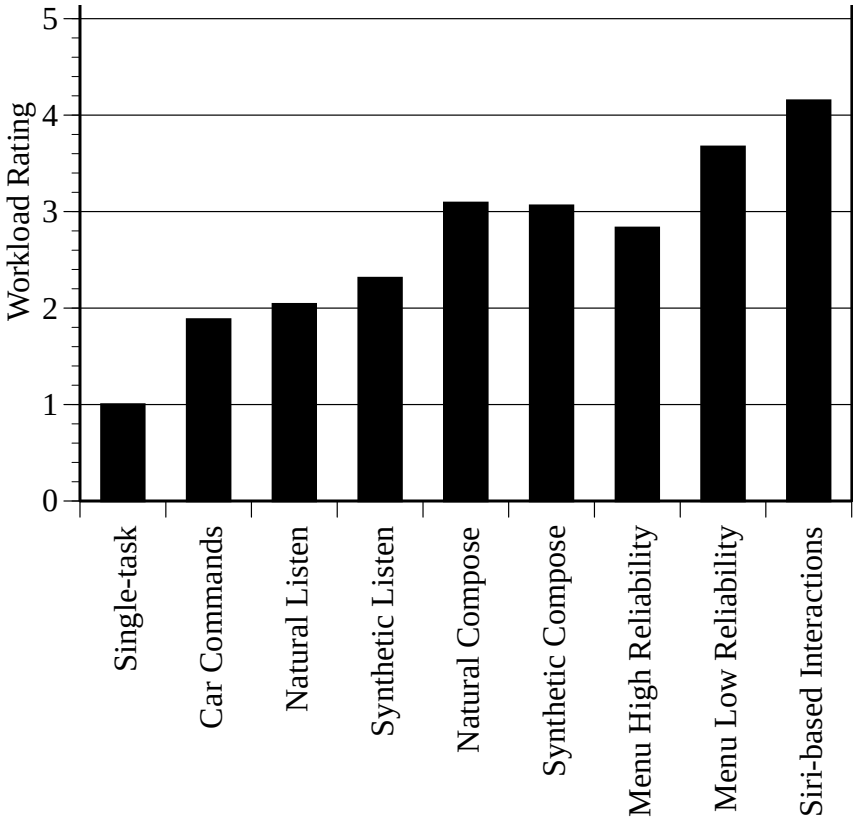

Figure 2. The Workload Rating Scale of tasks evaluated by Strayer et al. (2014).

The purpose of the current study is to assess cognitive workload associated with selected speech-to-text functionality of actual OEM systems in order to extend results from Strayer et al. (2013) and Strayer et al. (2014) to interactions with commonly available vehicle systems. This research addresses the following questions: How cognitively demanding are auditory-vocal vehicle interactions with actual OEM systems? How similar/dissimilar is the cognitive demand associated with different OEM systems? How do the cognitive demands of OEM speech interactions compare to the cognitive demands of the various tasks evaluated in Strayer et al. (2013, 2014)?

Method

Participants

Following Institutional Review Board approval, participants were recruited through ads placed on online local classifieds websites, flyers posted on the University of Utah campus, and by word of mouth. They were compensated $100 upon completion of the three-hour study. All data were collected from December 5th through December 10th of 2013.

A total of 36 participants completed this research (18 male, 18 female). Participants ranged in age from 22 to 36 years (M = 28.1, SD = 3.89). All participants were required to have a valid driver's license and have fewer than two accidents in the past two years. Additionally, participants were selectively recruited to balance gender. Thirty-four participants reported having no accidents in the past two years, with two reporting having one accident in the

past two years. Participants' years of driving experience ranged from 3 to 20 years, with an average of 11 years. None of the participants were familiar with any of the voice systems implemented across the six different vehicles. However, approximately half of the participants indicated familiarity with smartphone-based voice interactions.

Equipment

Systems from six different vehicle manufacturers were investigated. All were chosen because of their popularity and hands-free voice controlled functionality. These cars included a 2013 Ford Explorer Limited featuring SYNC with MyFord Touch, a 2013 Chevy Cruz Eco featuring Chevrolet MyLink, a 2013 Chrysler 300 with the Uconnect System, a 2012 Toyota Prius V Three with Entune, a 2013 Mercedes E350 featuring the COMAND® system, and a 2013 Hyundai Sonata SE with a Blue Link Telematics System. These six systems had many features in common, including steering wheel-mounted controls, Bluetooth phone pairing, voice-activated music functions, voice-activated CD playing, voice-controlled satellite radio, hands-free calling, and access to calling features (i.e. phonebook, call log, etc.).

An Alcatel One-Touch Fierce phone was paired via Bluetooth to each of the voice-controlled systems. Once paired, the phones allowed drivers to make hands-free voice calls using a standardized contact list or through voice controlled number dialing. The phones also provided the respective vehicle systems with a wireless internet connection. For this research, drivers never had to manually interact with the phone but could access many phone features through the vehicle infotainment system. Phones were placed in an out of the way location and were never directly viewed or manipulated by participants during the study.

Two Sony Compact POV Action Cams were placed in each vehicle. One camera was attached to the windshield just below the rearview mirror, pointing at the driver's face. The other was positioned between the passenger and driver seats facing forward, with the center stack and outside environment both visible. These cameras were chosen because of their compact size, high definition picture quality (1080p), Wi-Fi connectivity, built-in microphones, and GPS tracking abilities.

During all phases of testing, participants wore a head-mounted reaction time assessment device. These Detection Response Task (DRT) devices were assembled for the purpose of this study and follow the specifications outlined in ISO WD 17488 rev 10.1 (ISO, 2012). The devices consisted of an LED light mounted to a flexible arm that was connected to a headband.

Figure 3. DRT headband placement

The light was positioned in the periphery of the participants' left eye so that it could be seen while looking forward at the road but did not obstruct their view. The devices featured

a simple user interface which was optimized to assess mental workload. The precise configuration used in this research differed from the draft standard in two ways. First, the stimulus lights were configured to flash red *or green* every three-five seconds. Each time a light turned on, there was a 60 percent chance it would be red and 40 percent chance it would be green. Second, participants were given a response button and instructed to respond only to *green* lights as quickly as possible by clicking the button against the steering wheel. Timing was controlled on Asus Transformer Book T100s with quad-core Intel® Atom™ processors running at 1.33GHz.

Participants were outfitted with a Zephyr BioHarness 3 heart rate monitor. These professional quality heart rate monitors, and their accompanying software algorithms, have been tested to be within +/- 1 beat per minute of accuracy. The BioHarness 3 collects and stores comprehensive physiological data about the person wearing the monitor, including heart rate, heart-rate variability, breathing rate, posture, and activity level. The monitors attach around the chest with a flexible strap. For the purposes of this study, only Heart Rate, operationalized as the beats per minute, was used. Prior research suggests that of the many potential cardiovascular measures, Heart Rate is the most sensitive to mental workload (Mehler, Reimer, & Wang, 2011).

Three outputs were used to derive average Heart Rate for each subject and condition. These were the internal clock, the algorithm confidence, and the Heart Rate. The internal clock of each heart rate monitor was used to identify the segment of heart data which corresponded to each condition. Internal clocks for all six of the monitors were recalibrated each night to UTC. Internal clocks were found to drift < 2 seconds per day. Once activated, heart rate monitors began collecting and logging data at 1 Hz. Log files contain a number of summary measures which are all documented by Zephyr. Confidence measures for Heart Rate were used to verify signal quality at the time of each reading. Heart Rate measures with a signal confidence of less than 85 percent were omitted from analysis.

Procedure

A study facilitator was assigned to each car for the duration of the study. Their purpose was to ensure the safety of the driver, provide in-car training, and deliver task cues to participants. All facilitators had current driver's licenses. Three of the facilitators held Commercial Drivers' Licenses and two had significant driving-related research experience. All participant interactions by study facilitators followed a written script. Prior to experimentation, each facilitator was required to demonstrate mastery of the research protocol, including pre- and post-trip subject training and interactions.

Participants were scheduled in groups of up to six people. Upon arrival to the study location, they were given a consent form and intake questionnaire to complete (see Appendix A). Each participant was then given a heart rate monitor to be worn for the duration of the study. After receiving instructions on the proper fit of the heart monitor, each participant was given privacy to put it on. Once the experimenter had verified that the participant had correctly attached the monitor, they were directed to the car where they would complete their first condition. Prior to data collection in the vehicles, participants were able to familiarize themselves with the course by driving one circuit (see Figure 4). Each loop took approximately seven – nine minutes depending on stop lights, driving speed, and traffic at

stop signs. After familiarization with the course, participants received instructions about the DRT task and were given the opportunity to practice while sitting in a parked vehicle.

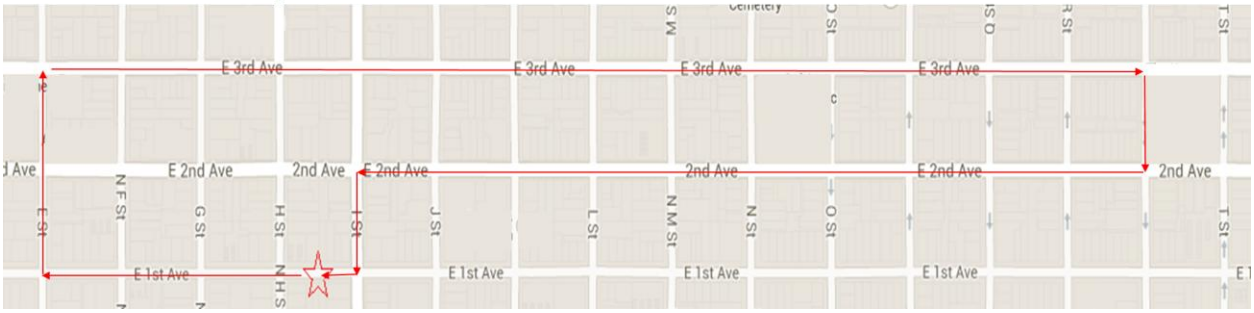

Figure 4. Course map

During experimentation, a high and low workload baseline task was given. The low workload baseline task consisted of routine driving in a single-task condition. The high workload baseline consisted of driving while performing the Operation Span (OSPAN) task used by Strayer et al. (2013, 2014). The OSPAN task consisted of listening to a math problem being read aloud, responding "yes" if the problem was true or "no" if it was false, and then being read a word aloud for later recall. The purpose of the OSPAN task was to create a high cognitive workload environment against which to compare the cognitive demand of the other tasks. For standardization, the OSPAN task was delivered via a 12 minute audio recording. An example interaction of the OSPAN task is as follows.

> Experimenter: Nine plus three divided by two equals twelve.
> Subject: False
> Experimenter: Cat
> Experimenter: Six minus five times three equals three.
> Subject: True
> Experimenter: Radio
> Experimenter: Thirteen minus ten plus three equals ten.
> Subject: False
> Experimenter: Cape
> Experimenter: Recall
> Subject: Cat… Radio… Cape

Condition **A** was a single-task drive which provided a baseline for the assessment of cognitive workload in the other conditions. In condition **B** participants drove while concurrently performing the OSPAN task. Conditions **C** through **H** were in-vehicle system interactions. Each voice interaction condition corresponded with one of the six vehicles (i.e. condition **C** corresponded to vehicle one, condition **D** corresponded to vehicle two, etc.) A schematic of the counterbalancing structure for the first 18 participants is shown in Figure 5. Complete instructions for each condition can be found in Appendix B.

Following the counterbalancing scheme presented in Figure 5, each participant completed just one drive in the single-task condition and one drive in the OSPAN condition. These two baseline assessments were collected using the same vehicle. Thus, a total of 36 (6 in each

vehicle) low workload baseline drives and 36 (6 in each vehicle) high workload baseline drives were completed. The order of the eight conditions was counterbalanced across subjects.

Subject Number	Condition
1.	**A B C** D E F G H
2.	**A B D** E F G H C
3.	**A B E** F G H C D
4.	**A B F** G H C D E
5.	**A B G** H C D E F
6.	**A B H** C D E F G
7.	H **A B C** D E F G
8.	C **A B D** E F G H
9.	D **A B E** F G H C
10.	E **A B F** G H C D
11.	F **A B G** H C D E
12.	G **A B H** C D E F
13.	G H **A B C** D E F
14.	H C **A B D** E F G
15.	C D **A B E** F G H
16.	D E **A B F** G H C
17.	E F **A B G** H C D
18.	F G **A B H** C D E
...	

Condition Key	
A	Single Task
B	OSPAN
C	Vehicle 1
D	Vehicle 2
E	Vehicle 3
F	Vehicle 4
G	Vehicle 5
H	Vehicle 6
Underlined conditions occurred in the same vehicle – shown in **bold**	

Figure 5. Counterbalancing scheme

Prior to driving in each vehicle, participants were given instructions on how to complete the calling and music functions tasks in the vehicle, and practiced with the system until they could complete the tasks without error. A complete transcription of each of the instructed voice interactions is provided in Appendix B. The Voice interactions with each of the six vehicles were functionally equivalent; the only thing that differed between vehicles was the precise sequence of commands that were required to complete the tasks. Each of the six tasks that were completed occurred at a specific geographical location on the course (see Figure 6). When the participant reached pre-specified locations on the course, the facilitator gave an instruction to begin the indicated task. Participants were not told where on the course the new tasks would be given, but the task onset location remained constant for all interactions. If the participant was unable to complete a task before the next one was to begin, they were told to abandon that task and move on. If participants failed to understand the task, instructions were repeated. All tasks began with the press of a steering wheel mounted button to initialize the voice command systems. Once initiated, each of the tasks was completed through auditory + vocal system interactions. System interactions alternated between completing a phone related task and a music functions task. Tasks were as follows:

Task 1: “Call from your contacts Joel Cooper on his cell”
Task 2: “Tune your radio to 99.5 FM,” once completed: “tune your radio to 1320 AM”
Task 3: “Dial your own phone number”

Task 4: “Play your CD*,” once completed: “tune your radio to 98.7 FM”
Task 5: “Call from your contacts David Strayer on his cell”
Task 6: “Tune your radio to 103.5 FM,” once completed: “play your CD”

*The Toyota Entune System did not allow this function at the time of testing so the alternative “Play your Satellite Radio” was used.

Figure 6. Course map and task locations

All data collection occurred during the workweek during daylight hours between 9am and 5pm. Peak traffic hours of 8-9 and 5-6:30 were avoided in order to ensure that all participants experience roughly the same levels of traffic during testing. However, given that this was an on-road study, there were small natural variations in roadway traffic that naturally occurred. Testing only occurred on days without active rain or snowfall, though some snow accumulation was present on the sidewalks and lawns of homes bordering the test route.

Following each drive, participants completed a written form of the NASA TLX, a subjective workload rating scale. This survey, developed by Hart and Staveland (1988), was used to measure subjective workload after the completion of each driving condition (see Appendix C). An additional two questions were added to the survey to gather information about the usability of each car system. Participants responded to the eight items on a 21-point Likert scale ranging from “Very Low” to “Very High.” After participants completed all eight of the experimental conditions, they filled out an exit survey which asked them to identify their favorite and least favorite systems. Most participants listed a single favorite and least-favorite vehicle system; however, some participants listed multiple vehicles for each question. In the case where multiple answers were given, responses were weighted by dividing each indicated vehicle system by the number of indicated systems. For example, if a participant listed two systems as their favorite, then each of the systems received half a point; if three where listed, each listed system received one third of a point.

Upon completion of the study, participants were asked to fill out a final questionnaire and were compensated for their time (see Appendix D).

Results

Three core workload measures were analyzed in this study. They were: Heart Rate, NASA TLX, and DRT Reaction Time. Due to the light vehicle instrumentation, primary driving performance data were not available. Once standardized, each of these measures were equally weighted and used to populate the Workload Rating Scale following the protocol developed by Strayer et al. (2013), and consistent with Strayer (2014). Each of these

measures was collected over the full sequence of voice interactions within each vehicle, and therefore characterize the overall level of workload associated with the full drive, including voice interactions to place a call, select music, and downtime between tasks. Measurements were thus derived during the entire testing block and not just during task intervals. Handling the data in this way accurately characterizes the full interaction experience by collapsing across momentary mental workload and task completion times. Prior to each analysis, data for each dependent measure were combined through averaging.

Heart Rate

Heart Rate readings with a confidence of less than 85 percent were not included in this analysis. This filtering excluded just under 15 percent of all collected heart data leaving an average of six minutes of usable heart data for each condition. Excluded data were likely the result of extraneous movements by the participant and ill-fitting chest straps.

A one-way repeated measures ANOVA was used to test for differences in Heart Rate among the eight experimental conditions. The overall test was significant, $F(7, 245) = 5.97$, $p < .001$, partial $\eta^2 = .146$, indicating that the measurement of heart-rate in the vehicle was sensitive to the experimental conditions (see Figure 7). The range of the mean Heart Rate values between the low and high workload baseline conditions was 3.17. Pairwise comparisons indicated that mean Heart Rate was significantly lower during music selection and call placement using Toyota's Entune system than with Hyundai's Blue Link, Chevrolet's MyLink, Chrysler's Uconnect, and Mercedes' COMAND systems (all p's < .05). On the flip side, music selection and call placement using Cheverolet's MyLink system led to a mean Heart Rate that was greater than all other vehicle systems *except* Mercedes COMAND (all p's < .05). In short, Toyota's Entune system elicited the lowest mean Heart Rate and Chevrolet's MyLink system elicited the highest mean Heart Rate, while all other systems were statistically undifferentiated.

NASA TLX

The six subscales of the NASA TLX were combined through an equally weighted average. The resulting aggregate scores were then subjected to a one-way repeated-measures ANOVA. Results indicated a highly significant main effect of experimental condition, $F(7, 245) = 56.3$, $p < .000$, partial $\eta^2 = .62$ (see Figure 8). Pairwise comparisons of the eight experimental conditions revealed a pattern that was very similar to that obtained from the Heart Rate measure reported above. In general, music selection and call placement on all of the systems elicited responses that were differentiated from the low and high workload baselines. Toyota's Entune and Hyundai's Blue Link systems were rated slightly more demanding than the low workload baseline; the Chrysler, Ford, and Mercedes systems were rated somewhat more demanding; and finally, music selection and call placement on Chevrolet's MyLink system were rated the most demanding, but still substantially less than the high workload baseline.

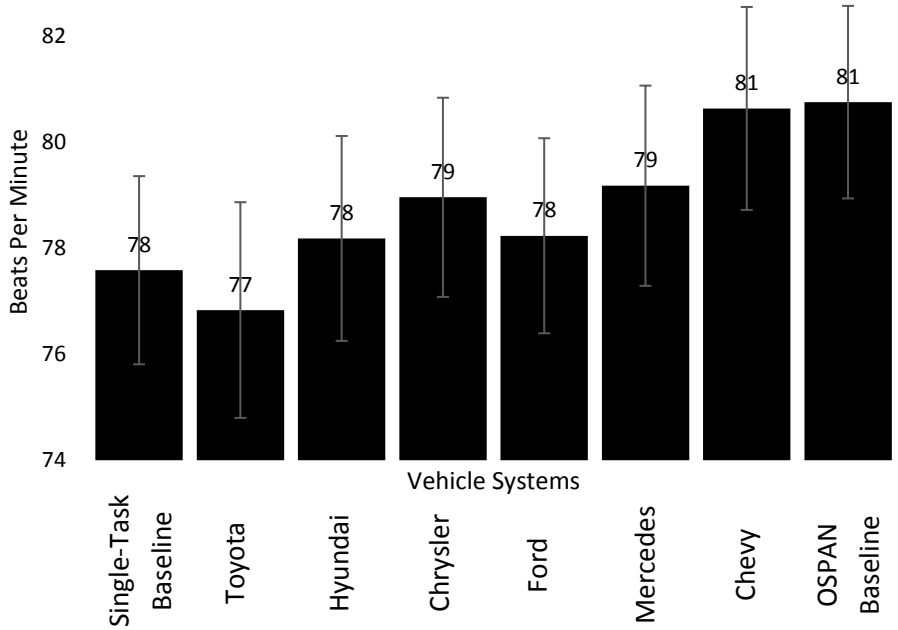

Figure 7. Mean Heart Rate for each of the 8 research conditions. Error bars represent the Standard Error of the Mean.

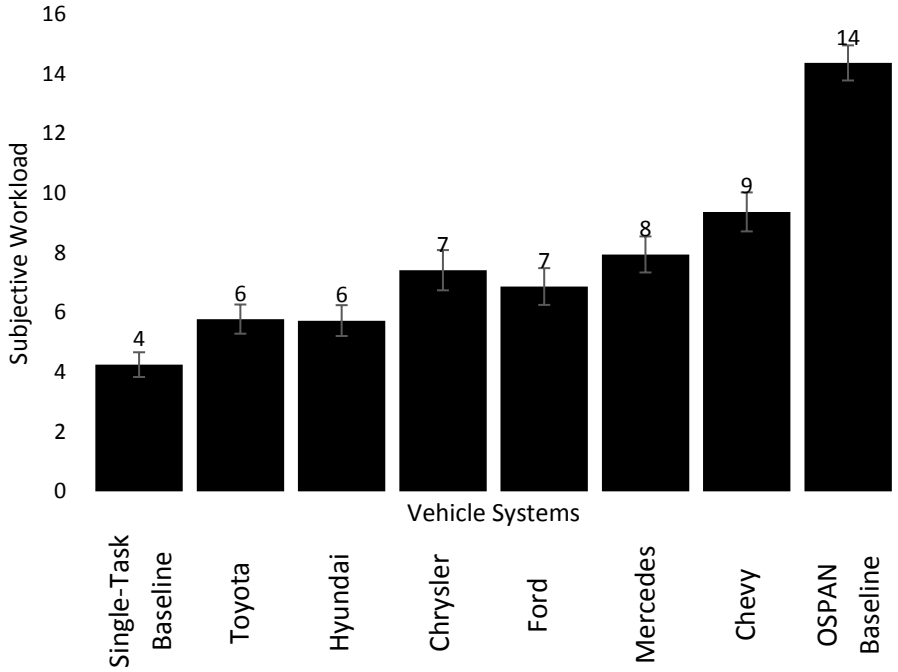

Figure 8. NASA TLX for each of the 8 research conditions. Error bars represent the Standard Error of the Mean.

DRT Reaction Time

In order to clean the DRT data for analysis, all non-responses prior to the first response were removed and all non-responses after the last response were also removed. Additionally, all

responses which fell under 100ms or beyond 2500ms were removed. This cleaning procedure was identical to that specified by the ISO draft standard. Mean reaction time was then calculated from the remaining data. Excluding either unrealistically fast or slow responses removed an average of 8.6 responses from each condition for each subject. This standard procedure left an average of 25.4 valid responses for each subject in each condition.

A one-way repeated measures ANOVA indicated that the music selection and call placement tasks significantly affected reaction time, $F(7, 245) = 16.1$, $p < .000$, partial $\eta^2 = .315$ (see

Figure). Pairwise comparisons indicated a very similar pattern to that seen in the Heart Rate and NASA TLX measures. One exception was that mean reaction time while selecting music or placing a call using the MyFord Touch system was significantly slower than with all other systems. Otherwise, the same consistent pattern was observed: reaction times while interacting with any of the voice based systems was significantly slower than the low workload baseline but faster than the high workload baseline. Again, the one exception was that reaction times while using the MyFord Touch system were nearly as delayed as those observed in the high workload baseline.

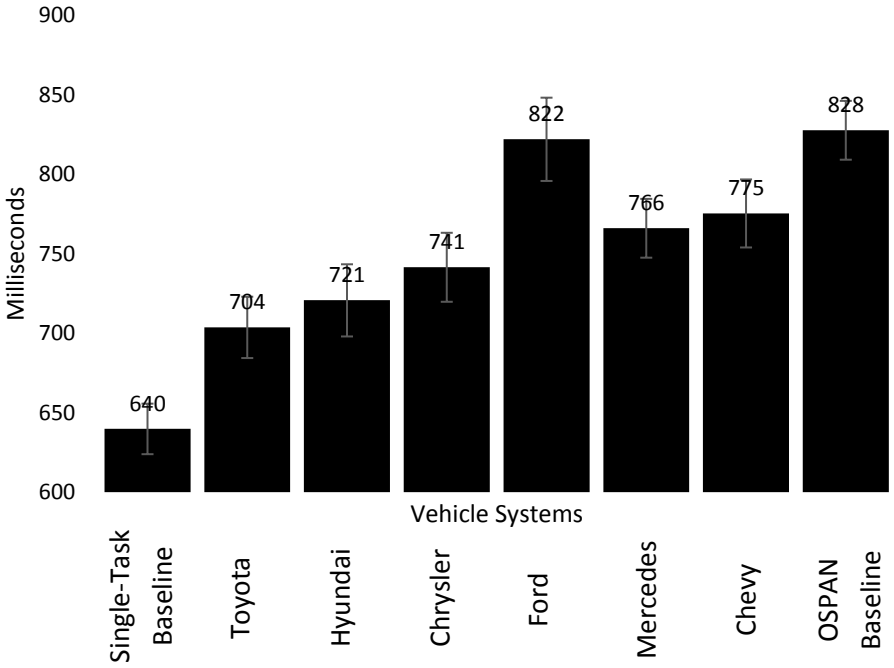

Figure 9. DRT Reaction Time for each of the 8 research conditions. Error bars represent the Standard Error of the Mean.

Workload Rating Scale

Based on the combinatorial procedure that was presented by Strayer et al. (2013), Heart Rate, NASA TLX, and DRT Reaction Time data were Z-transformed and linearly combined with equal weighting to generate a summary variable (Please refer to Strayer et al. [2013]

for a complete description of this analytic approach). This summary variable was then analyzed using a one-way repeated measures ANOVA revealing a significant overall effect of condition, $F(7, 245) = 49.4$, $p < .000$, partial $\eta^2 = .585$. A table with the mean Z-scores for each of the measures and experimental conditions is presented below.

Table 1. Z-Scores for Heart Rate Heart Rate, NASA TLX, and DRT Reaction Time

Measure	Single-Task	Toyota: Entune	Hyundai: Blue Link	Chrysler: Uconnect	Ford: MyFord Touch	Mercedes: COMAND	Chevy: MyLink	Ospan
Heart Rate	-0.49	-0.49	-0.03	0.05	-0.14	0.13	0.51	0.43
TLX	-0.93	-0.56	-0.49	-0.08	-0.19	0.06	0.41	1.77
DRT	-1.02	-0.46	-0.26	-0.15	0.62	0.18	0.29	0.79

Pairwise comparisons for the measures cleanly distinguished six groups (see Figure 10). As expected, the single-task and OSPAN conditions were statistically different from all of the voice based system interactions. On the low end, Toyota's Entune system produced moderately more mental workload than the single-task condition. Based on our prior findings, this resulted in a workload estimate that is similar to listening to the radio or an audio book. Music selection and call placement using Hyundai's Blue Link system led to a significant increase in workload from the Toyota system, a level which was similar to holding conversation over a cell phone or with a passenger. Music selection and call placement using the Chrysler, Ford, and Mercedes systems led to a level of workload that was similar to the error free speech-to-text system evaluated in Phase 1. Finally, music selection and call placement using Chevrolet's MyLink system led to a level of workload that was greater than any of the other system interactions.

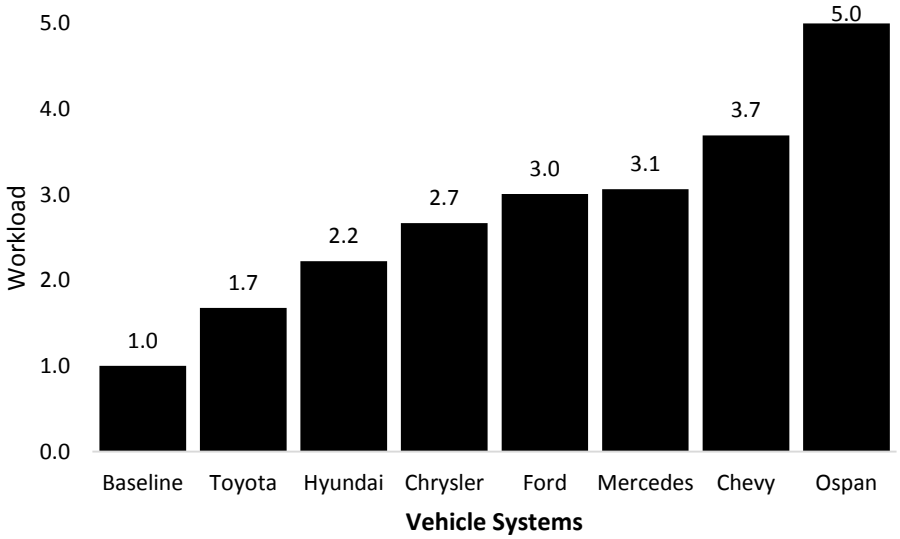

Figure 10. Mental Workload Scale for each of the 8 research conditions

Supplementary Measures

Given the somewhat surprising variability in the cognitive ratings across the range of OEM offerings, the research team explicitly considered five additional factors. These are: Task

Completion Time, Task Errors, System Dialogue Steps, Subjective Evaluations, and Vehicle Workload. These are presented here because they provide additional insight into the validity and reliability of the rating system.

Task Completion Time. A one-way repeated measures ANOVA indicated a significant effect of System on task completion time, $F(5, 175) = 24.8$, $p < .000$, partial $\eta 2 = .415$ (see Figure 11). Pairwise comparisons indicated that voice interactions using Toyota's Entune system were completed in the least amount of time (Call Placement: 20 seconds; Music Selection: 22 seconds), while the same tasks using Chevrolet's MyLink system took considerably longer (Call Placement: 29 seconds; Music Selection: 43 seconds). At the vehicle level, task completion time was significantly correlated with the Workload Rating Scale ($r(10) = .96$, $p < .001$), suggesting that the amount of time require for actual subjects to complete each voice task was an important element in the measured cognitive demand of the systems.

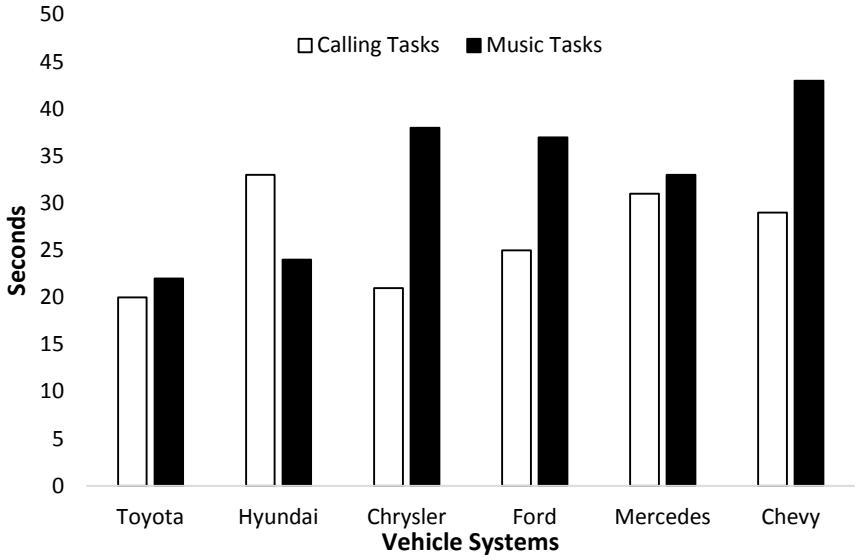

Figure 11. Task Completion Times by system and task

Task Errors. The average number of errors per task was calculated for the music selection and call placement tasks across each of the six vehicle systems. Errors were classified as system interactions where a system misunderstanding was present. For example, the participant may have said "Tune to 90.1," but the system heard "Tune to 98.1." These nonparametric data were assessed using a Friedman Chi Square test. Results indicated a significant difference across the six different vehicle systems ($\chi 2(5) = 27.8$, $p < .001$), as well as a significant difference between the two voice tasks ($\chi 2(1) = 18.7$, $p < .001$) (See Figure 12). One interesting outcome from this analysis is the clear dichotomy between system performance on the call placement and Music Tasks. In every system, with the exception of the Hyundai, errors were more common in the music selection task. Indeed, the music selection task is what appears to truly separate the systems in terms of their error rates.

The ordered relationship between error counts and mental workload suggests that a strong component of mental workload is likely the error proneness of the system. Without exception, the rank ordering of systems based on interaction errors results in the same ordering as the mental workload scale. That is, in both cases, the rank ordering across the

vehicle systems was: Toyota, Hyundai, Chrysler, Ford, Mercedes, and then Chevrolet. At the system level, the correlation between Task Errors and the Workload Rating Scale was significant ($r(10) = .85$, $p < .001$). This indicates that a significant element of cognitive load during voice interaction is driven by system errors.

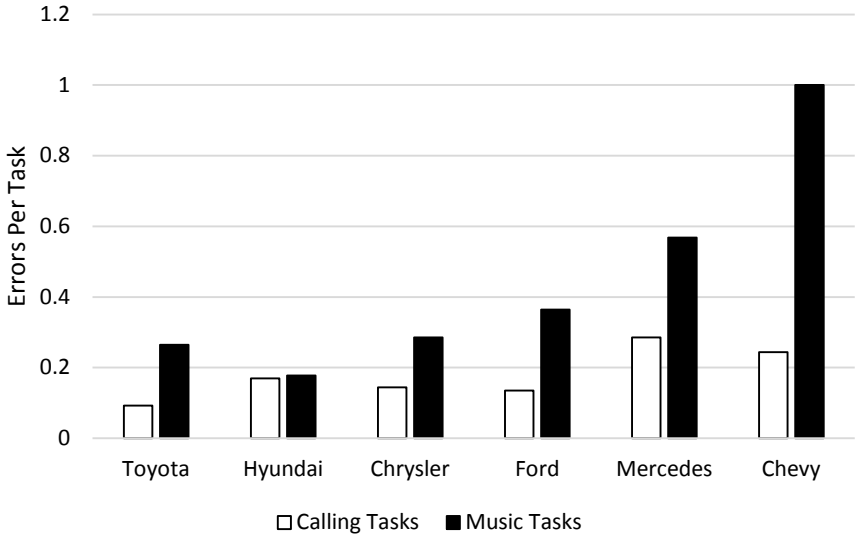

Figure 12. Mean Error Count per interaction by system and task

System Dialogue Steps**.** One possibility in the above analyses is that tasks with more required steps for completion lead to longer completion times and an increase in the number of Task Errors. In order to gain insight into this potential relationship, the number of required steps to complete each task was coded. The exact command sequence for each of the tasks and systems is given in Appendix E. From this, a distillation of the required words to complete each task as well as the turn-takes during each interaction were extracted. These are presented in Table 2. Interestingly, neither the total interaction words nor the total turn takes correlated with the computed Workload Rating Scale ($r(10) = .14$, $p > .05$; $r(10) = .44$, $p > .05$). This suggests that the cognitive demand incurred by voice interactions with vehicle systems is not necessarily predominated by the structure of the voice interactions. To illustrate, the Chevrolet MyLink system required fewer turn takes and total words than some of other systems but led to the highest score on the Workload Rating Scale. However, the Toyota Entune system, which elicited the least amount of mental workload also required the fewest turn takes and words to complete the various tasks, suggesting that a well-structured and concise interaction may be an important part of the solution, but not the only requirement. Clearly, the error proneness of the system is also critical.

Subjective Evaluations. After completing each of the eight experimental conditions, participants were asked to identify their favorite and least favorite vehicle systems. A simple tally of these data indicated that subjective preferences followed the workload rating scale almost perfectly. The Toyota system was rated as the most preferred, followed by the Hyundai system, then the Chrysler and Ford systems, with the Chevrolet system being ranked lower than the others. This ordering was remarkably similar to that observed on the other variables. The one exception to this fit was the Mercedes COMAND system, which ranked as the least favorite system. Thus, on the whole, expressed preferences for the system were highly related to the measured cognitive demands for the interactions.

Preferred systems were the least cognitively demanding, while systems that were least preferred were generally the most cognitively demanding.

Table 2. Task completion turns and required word count.

	Dial Number			Call Contact		
	Turns	User Words	System Words	Turns	User Words	System Words
Toyota	4	7	15	3	4	4
Hyundai	7	8	30	4	5	19
Chrysler	4	6	30	2	4	12
Ford	6	7	20	3	4	9
Mercedes	6	8	10	4	4	12
Chevy	6	7	25	3	4	9

	Play CD			Tune Radio		
	Turns	User Words	System Words	Turns	User Words	System Words
Toyota	1	1	0	1	4	0
Hyundai	3	1	8	3	1	9
Chrysler	2	3	4	2	4	4
Ford	3	1	5	3	1	7
Mercedes	1	1	0	3	3	0
Chevy	3	1	5	3	4	8

Vehicle Workload. One potential issue that arises in the above analysis of workload associated with the music selection and call placement tasks is that some of the observed workload across the different vehicles might have been driven by different control requirements for the vehicles themselves. That is, some of the vehicles may have been more demanding to drive than others. In order to address this potential issue, a standardized aggregate score was created for each pair of baseline drives which combined the DRT reaction time, the Heart Rate data, as well as NASA TLX data. This value was averaged across the low and high workload baselines (we reasoned that any difference in workload associated with just driving would be reflected in both of the baseline driving conditions). The logic for this aggregation is identical to the logic which supports the Workload Rating Scale presented above. Identical to the workload rating scale, vehicles that were consistently more or less cognitively demanding to drive would yield standardized workload ratings above or below zero.

A one-way ANOVA indicated that none of the vehicles was any more or less demanding to drive than any of the others, $F(5,30) = .694$, $p = .635$. Based on these results, we feel confident that the observed differences in mental workload associated with the six systems evaluated in this report were the result of system interactions and not related to difference in baseline vehicle driving demand. For comparison, the Z-transformed scores associated with just driving as well as driving while interacting with the voice systems are presented in Figure 13 below.

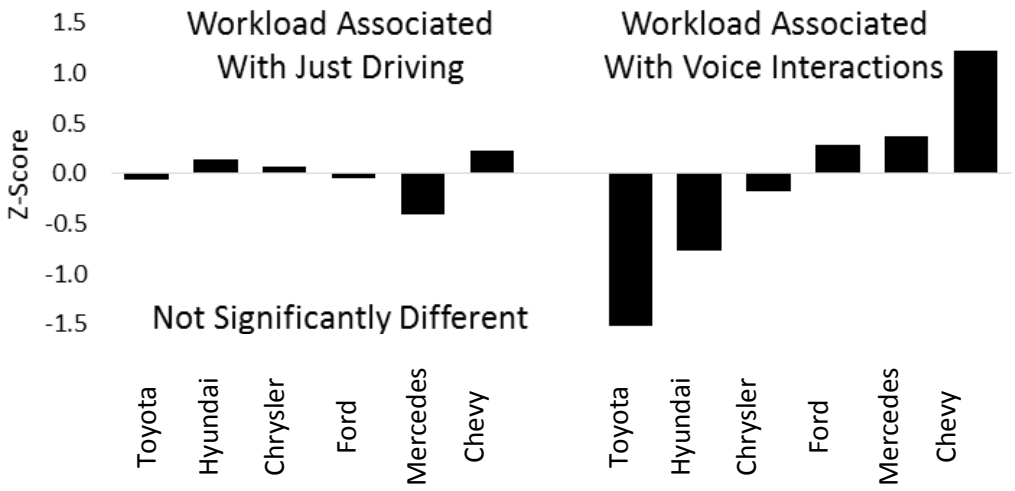

Figure 13. Workload of driving each vehicle compared to workload of interacting with each vehicle's system.

Discussion

General Findings

The purpose of this study was to evaluate the mental demands of simple auditory-vocal vehicle interactions across five 2013 and one 2012 model year OEM infotainment systems. This research was designed to address three novel questions, which were: How cognitively demanding are auditory/vocal vehicle interactions with actual OEM systems? How similar/ dissimilar is the cognitive demand associated with different OEM systems? How do the cognitive demands of OEM speech interactions compare to the cognitive demands of the various tasks evaluated in Strayer et al. (2013, 2014)?

Results obtained in this investigation indicate that simple auditory-vocal interactions with vehicles may significantly elevate mental workload in drivers. On the standardized rating scale developed by Strayer et al. (2013), and refined in 2014, a mean demand score of approximately 3 was observed across the six OEM systems. A score of 3 is midway between the workload associated with the single-task baseline and the OSPAN mental math condition, and indicates a moderate level of cognitive load. Workload ratings for all voice interactions were greater than that observed in the single-task driving condition and less than that observed in the OSPAN task. Thus, all systems imposed some demand, but no system imposed more demand than the OSPAN math task.

In the best case, evaluated voice commands using Toyota's Entune system imposed modest additional demands as compared to the single-task baseline condition. In the worst case, those same activities using Chevy's MyLink system imposed mental demands that were approaching the high workload baseline (OSPAN mental math). Not surprisingly, the most critical element of workload appeared to be the duration of the interaction. For the tasks selected in this analysis, Toyota's Entune system required the least amount of interaction time while Chevrolet's MyLink required the most.

Most of the auditory-vocal interactions evaluated in this research were more demanding than the three conversation tasks evaluated by Strayer et al. (2013). The exception, of course, was that call voice interactions using Toyota's Entune system were less demanding than the book-on-tape condition used by Strayer et al. (2013). In general, most systems elicited about the same, or less, mental workload than the speech-to-text task evaluated by Strayer et al. (2013). The exception here was that the Chevy MyLink system was significantly more demanding (see Strayer et al., 2014 for additional research contrasts). Overall, observed mental workload during voice interactions ranged from being as low as listening to audio media to higher than any of the non-math tasks earlier measured.

A primary contributing factor related to mental workload was the total time required to complete each system interaction. Task completion time is clearly a high level measure that encompasses a variety of subordinate factors. These include dialogue requirements and accuracy of speech comprehension, among others. In this research we measured the system verbosity in conjunction with the optimal command structure. Additionally, the accuracy of speech comprehension was captured through the measurement of Task Errors. In general we found that short and concise interactions were related to reduced cognitive demand, but that a more important factor in total task time was the number of errors that arose during the interaction. A comparison of the task step data with the workload ratings data and the error data reveals that some systems with fairly concise interaction steps elicited elevated mental workload due to an increase in comprehension errors, while other systems that required more task steps – but made fewer errors – fared better.

Mental workload for activities evaluated in this investigation was highest for the Chevrolet MyLink system. The score of 3.7 on the standardized scale is among the highest that has been measured (however, see the Siri condition in Strayer et al. 2014). We believe that this high level of workload was elicited by system errors and the prolonged duration of the task. In many circumstances, participants were unable to complete the music functions task at all during the drive. For many of the drivers, the first reaction was to simply give up trying to use the system. Given the circumstances, this is perhaps the safest decision. We feel that it is unrealistic to expect drivers to persist in failed attempts to use their voice to achieve a goal which can be accomplished manually by the flick of a switch and the press of a button. Drivers will probably not use voice commands if they do not require less effort than their manual counterparts. Voice systems which fail to understand the driver will not be used, and if they are used, will result in high cognitive demands and frustration.

Another finding worth additional discussion was that the Mercedes system was subjectively rated as the least favorite voice system by participants. Objectively, however, it imposed no more workload than the systems offered by Ford or Chrysler. One reason the Mercedes may have received unfavorable reviews was because of the rigidity of the commands that it required. If a driver did not say a command in a very specific manner, the system would not understand the command. In addition, to complete the radio tuning and CD tasks the system required reactivation after every step. For example, to play a radio station, the driver had to press the voice button and say "radio." He or she then had to press the button again and say "FM" or "AM." The final step was to push the button yet again and say the name of the radio station. In other vehicles, these steps were often combined into one simple command, e.g. "play 99.5 FM." Despite the cumbersome nature of the command procedure, the system itself seemed to respond to proper commands fairly quickly. Whereas in some systems the driver had to wait until the audio system finished speaking before

saying a command, the Mercedes played a short tone to show it was listening. Additionally, in some of the cars the driver also had to wait for the system to verbally repeat the command just given, whereas in the Mercedes no audible feedback was given.

Cognitive workload in this evaluation appeared to have been primarily driven by interaction time. Interaction time was in turn driven by dialogue requirements and comprehension errors. In order to maximally reduce mental demand, system interactions should be as short and accurate as possible. Well-executed voice systems have the potential to keep a driver's eyes on the road without imposing significant cognitive demand. However, poorly executed voice systems may have the opposite effect by imposing high levels of mental demand on drivers with the potential to also incur long glances away from the roadway in order to check system status and understanding. Based on this it is clear that voice interactions can be made sufficiently simple and accurate to reduce cognitive demand in the vehicle to levels approaching the widely accepted tasks of listening to the radio or a book on tape.

Results indicated that Heart Rate was generally less sensitive than we had anticipated based on Mehler, Reimer, & Wang (2011). One potential explanation for the relative insensitivity of Heart Rate in the current study is the way in which tasks were evaluated. It is possible that for Heart Rate to be sensitive to mental workload a sustained task engagement is needed. The discrete task engagement design used in the current study may not have allowed Heart Rate to track with task difficulty in the expected manner. In order to evaluate this hypothesis it would have been necessary to evaluate changes in Heart Rate as participants engaged and disengaged in each of the voice tasks. Given the hardware used in the current study, such an analysis was simply not possible. Alternatively, it may be the case Heart Rate is sensitive to some types of cognitive workload and not others, or that Heart Rate can only provide a general reflection of workload and is not as sensitive to fine gradations in workload that might arise from subtle task differences.

A measurement challenge that arose during this evaluation regards the most appropriate method for evaluating cognitive workload using the same task among different vehicles. Traditionally, cognitive workload evaluation has inferred a momentary measure of load by averaging workload across a continuous interaction. Examples of this approach can be seen in research on cell phone conversations and driving (See Drews, Pasupathi, & Strayer, 2008; McKnight & McKnight, 1993; Rakauskas, Gugerty, & Ward, 2004) and research looking at general cognitive load and driving (Harbluk, Noy, & Eizenman, 2002; Harms, 1991). In these cases, data are often treated as if cognitive demand is constant during an entire experimental condition. Averaged differences in performances from a baseline condition are then interpreted as the effect of the additional load. This general approach cannot be reasonably applied to the comparison of real world systems which differ in their expected interaction time. Measuring real-world systems using a fixed interaction duration would not factor in expected differences in task completion time and would return a workload value at the moment of task interaction, no matter how prolonged or brief the task. There are a number of potential methods for factoring exposure duration into the final estimate of workload. This research did so by fixing task locations within the evaluation drive so that participants could experience some single-task driving if they finished the current task prior to the scheduled onset of the next task. An alternative approach would be to rescale performance data in a manner that factors in task completion time. In order to

fully evaluate the cognitive demands of simple voice interactions across different systems some accounting of task interaction time must be factored in to the assessment.

Limitations

Currently, the association between cognitive driver distraction and safety risk is not well understood. Driver distraction has been defined as "The diversion of attention away from activities critical for safe driving toward a competing activity, which may result in insufficient or no attention to activities critical for safe driving" (Regan et al., 2011). As such, the Workload Rating Scale used in this research may be considered as a cognitive driver distraction scale. However, because of the complex manifestations of cognitive driver distraction, it is not clear whether and how these observed differences might result in changes to real world safety risk. As of yet, there is no unambiguous correspondence between variations in mental workload and the actual risk of a crash. Clearly, additional research is needed to gain a better understanding of the crash risks of various cognitive tasks. At a minimum, a confident understanding of the risk of two previously measured tasks on the Workload Rating Scale will allow relative risks to be extrapolated for the other tasks.

Drivers in this research were also not experienced with each of the in-vehicle systems that were evaluated. We would expect that drivers who routinely use their voice activated system features would show a gradual reduction in the cognitive demand required to use that system. However, the general naivety of users in this research is useful in its own right as it reflects a driver's workload, success, and frustration during the critical first exposure. Indeed, if the system does not work right the first time, a single exposure is all that a driver might ever have, as frustration and failure might keep a driver from ever returning to the task.

Each of the participants in this research experienced all six of the different voice systems during a single three hour block. This design made it possible for participants to anchor their subjective ratings within their immediate experience. This was likely a strong contributing factor to the finding that subjective workload ratings were highly consistent with the objective measures. However, one drawback of having participants experience each of the six systems in a single block is that they may have had a difficult time adapting to each of the different systems, possibly getting confused about the appropriate syntax required to complete each of the various tasks. Additionally, it is also possible that participants became fatigued during the study, which may have adversely affected their performance.

This research evaluated a call placement and a music selection task that were similarly implemented across each of the vehicles included in this study. This restricted set of tasks allowed us to make direct comparisons between vehicles. However, it is unknown how the results might generalize to other in-vehicle voice commands afforded by the various vehicle systems. Critically, this research does not allow us to generalize to the full set of functions afforded by any of the systems that we evaluated. It is certainly possible that some systems may be unusually poor at some tasks but better at others. If we had evaluated a different set of tasks it is highly likely that the relative performance of the systems would have changed.

References

Delogu, C., Conte, S., & Sementina, C. (1998). Cognitive factors in the evaluation of synthetic speech. *Speech Communication*, *24*(2), 153-168.

Drews, Frank A., Monisha Pasupathi, and David L. Strayer. "Passenger and cell phone conversations in simulated driving." *Journal of Experimental Psychology: Applied* 14.4 (2008): 392.

Harbluk, J. L., & Lalande, S. (2005). Performing e-mail tasks while driving: The impact of speech-based tasks on visual detection. In *Proceedings of the 3rd International Driving Symposium on Human Factors in Driving Assessment, Training and Vehicle Design* (pp. 304-310).

Harbluk, J. L., Noy, Y. I., & Eizenman, M. (2002). *The impact of cognitive distraction on driver visual behaviour and vehicle control* (No. TP# 13889 E).

Harms, L. (1991). Variation in drivers' cognitive load. Effects of driving through village areas and rural junctions. *Ergonomics*, *34*(2), 151-160.

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

Hyman, I. E., Boss, S. M., Wise, B. M., McKenzie, K. E., & Caggiano, J. M. (2010). Did you see the unicycling clown? Inattentional blindness while walking and talking on a cell phone. *Applied Cognitive Psychology*, *24*(5), 597-607.

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

McKnight, A. J., & McKnight, A. S. (1993). The effect of cellular phone use upon driver attention. *Accident Analysis & Prevention*, *25*(3), 259-265.

Mehler, B., Reimer, B., & Wang, Y. (2011). A comparison of heart rate and heart rate variability indices in distinguishing single-task driving and driving under secondary cognitive workload. Paper presented at *the 6th international driving symposium on human factors in driving assessment, training, and vehicle design.* Lake Tahoe, California. June 27-30.

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

Paris, C. R., Gilson, R. D., Thomas, M. H., & Silver, N. C. (1995). Effect of synthetic voice intelligibility on speech comprehension. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(2), 335-340.

Rakauskas, M. E., Gugerty, L. J., & Ward, N. J. (2004). Effects of naturalistic cell phone conversations on driving performance. *Journal of safety research*,*35*(4), 453-464.

Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of experimental psychology: Applied*, *9*(2), 119.

Regan, M., Hallett, C., & Gorden, C.P. (2011). Driver distraction and driver inattention: Definition, relationship and taxonomy. *Accident Analysis & Prevention,* 43(5), 1771-1781.

Simons, D. J. (2000). Attentional capture and inattentional blindness. *Trends in cognitive sciences*, *4*(4), 147-155.

Strayer, D.L., & Drews, F. A. (2004). Profiles in driver distraction: Effects of cell phone conversations on younger and older drivers. Human Factors, 46, 640-649.
Strayer, D. L., & Drews, F. A. (2007). Cell-phone-induced driver distraction. Current Directions in Psychological Science, 16, 128-131.
Strayer, D.L., Cooper, J.M., Turrill, J., Coleman, J., Medeiros-Ward, N., & Biondi, F. (2013).Measuring Cognitive Distractionsin the Automobile. AAA Foundation for Traffic Safety, Washington, DC.
Strayer, D.L., Turrill, J., Coleman, J., & Cooper, J.M. (2014). Measuring Cognitive Distraction in the Automobile II: Assessing In-Vehicle Voice-Based Interactive Technologies. AAA Foundation for Traffic Safety, Washington, DC.

Appendix A: Intake Questionnaire

Participant Number (for lab use only) _____
Date of birth (mm/dd/yy) _____
Gender Male Female
Handedness Right Left
Date of Study (mm/dd/yy) _____

1. Do you have normal or corrected-to-normal vision?

   Yes No

2. Are you color blind? * If you don't know, please tell us.

   Yes No

3. Are you a native or equally fluent speaker of English?

   Yes No

4. Have you had your normal amount of caffeine today?

   Yes No

5. Did you get a normal amount of sleep last night?

   Yes No

   5a. If no, please specify how many hours of sleep you got last night: _____

6. Do you have a valid driver's license?

   Yes No

7. Have you ever participated in a study in which you drove or observed someone drive in an instrumented vehicle?

   Yes No

Appendix B: Instructions and Training

DRT Training

- Periodically, either a red or green light is going to turn on. When you see the green lights, please respond as quickly as you can by clicking the button on your finger against the steering wheel. Only click the button once. When you click the button the light will turn off, otherwise the light will only be one for 1 second and then it will turn off. The lights will continue to cycle between red and green at random intervals. You will not know when the next light will turn on or which color it will be. Remember to respond as quickly as you can to only the green lights. Do you have any questions?

Single Task

- In this condition, you will be driving around the course as you normally do, obeying all traffic laws and not exceeding the speed limit of 25 mph. You will continue to respond to the green light as quickly as possible by pressing your finger once against the steering wheel.
- Do you have any questions?
- *[Note: remember to fill out AAAFTS Workload Ratings Survey]*

Ospan Task

- In this condition, you will be driving around the course as you normally do, obeying all traffic laws and not exceeding the speed limit of 25 mph. In addition to driving, you will do a verbal math and memory task. Once we begin driving you will hear math problems being read to you. After the math problem is read, please respond "yes" if the answer is true and "no" if the answer is false. For example, you will hear "is 2*1+1=5" and you would respond "no" because it does not equal 5. After you answer either yes or no we will give you a word to remember for later recall, for example "dog". When you get to the end of a set of math problems we will say "recall" and this is your cue to recall the list of words from that set in the order in which you heard them. If you don't remember all the words, you might say "The first word is dog, I don't remember the second word, the third is cat." The math problems and words cannot be repeated so do your best to listen and respond as accurately as you can.
- You will continue to respond to the green light as quickly as possible be pressing your finger once against the steering wheel.
- Do you have any questions?
- *[Note: Play example file labeled "Ospan Example" on desktop]*
- *[Note: remember to fill out AAAFTS Workload Ratings Survey]*

In-Vehicle System Interaction (IVIS)

- In this condition, you will be driving around the course, driving as you normally do, obeying all traffic laws and not exceeding the speed limit of 25 mph. In addition to driving, you will be doing a number of voice-controlled tasks. For example, you may be asked to call someone, dial a number, or change the radio station. Once you have

completed the task, continue driving until you are given instructions for the next task. If we interrupt your current task with instructions for a new task, please abandon your current task to complete the new one. We will go over examples of how to use the system.

- You will continue to respond to the green light as quickly as possible be pressing your finger once against the steering wheel.
- *[Note: remember to fill out AAAFTS Workload Ratings Survey]*

Training (12-15 minutes): Complete 5 examples for each action (radio, CD, call, dial)

Radio Tasks:

Play:	CD	96.3 FM
	93.3 FM	CD
	CD	97.1 FM
	1160 AM	CD
	CD	107.9 FM

Phone Tasks:

Call:	John Doe, Mobile	Dial:	801-000-1234
	Chris Hunter		your own phone number
	Mike Earl, Work		555-555-5555
	James miller		your own phone number
	Kirk Baird, Home		801-123-4567

Appendix C: NASA TLX Survey

Condition (circle one): Participant Number:_____

Single Task	Car 2 IVSI	Car 5 IVSI
Ospan 1	Car 3 IVSI	Car 6 IVSI
Car 1 IVSI	Car 4 IVSI	

How mentally demanding was the task?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Very Low Very high

How physically demanding was the task?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Very Low Very high

How hurried or rushed was the pace of the task?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Very Low Very High

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

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Perfect Failure

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

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Very Low Very High

How insecure, discouraged, irritated, stressed, and annoyed were you?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Very Low Very High

Only answer for IVSI tasks

How intuitive, usable, easy to use was this system?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Not at all Very Much

How complex, difficult, confusing was this system?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Not at all Very Much

Appendix D: Exit Questionnaire

WIRB 20131987
#11469905.0

Participant # _____.

Answer the following questions on a scale of **1-7**, where an answer of 1 is Not At All and an answer of 7 is extremely.

If you owned one of the vehicles featured in today's study, how likely would you be to use your voice to do the following?

1. _____ Control the radio
2. _____ Control a navigation system
3. _____ Dial a number from your contacts
4. _____ Dictate a text message or email
5. _____ Update social media such as Facebook or LinkedIn

6. What was your FAVORITE voice system that you used in this study and why?

7. What was your LEAST favorite voice system that you used in this study and why?

Appendix E: Optimal System Interaction Dialogue

Ford MyFORD Touch

Dial

System: “Please say a command”
Subject: “Dial”
System: “Start saying a phone number”
Subject: “801 - 520 - xxxx”
System: “801 - 520-xxxx” say dial, delete, or continue speaking the digits”
Subject: “Dial”

Contacts Call

System: “Please say a command”
Subject: “Call Joel Cooper Cell”
System: “Calling Joel Cooper on cell”

Play CD

System: “Please say a command”
Subject: “CD”
System: “CD…”

Tune Radio

System: “Please say a command”
Subject: “98.7”
System: “Tuning to 98.7”

MYLINK

Dial

System: “Please say a command”
Subject: “Dial”
System: “Please say a phone book name, you may also say a number, and then say dial”
Subject: “801 - 520 - xxxx”
System: “801 - 520 - xxxx”
Subject: “Dial”

Contacts Call

System: “Please say a command”
Subject: “Call Joel Cooper cell”
System: “Calling Joel Cooper on cell”

Tune Radio

System: “Please say a command”
Subject: “Tune to FM 98.7”
System: “Tuning to FM 98.7”

Play CD

System: “Please say a command”
Subject: “CD”
System: “CD”

Chrysler UConnect

Dial

Subject: “Dial”
System: “Say the phone number or say the full number and phone type you want to call”
Subject: “801 - 520 - xxxx”
System: “Dialing 801 - 520 - xxxx, press the phone button to end the call”

Contacts Call

Subject: “Call Joel Cooper Cell”
System: “Calling Joel Cooper mobile, press the phone button to end the call”

Play CD

Subject: “Change to CD”
System: “Changing source to disk”

Tune Radio

Subject: “Tune to 98.7 FM”
System: “Tuning to 98.7 FM”

Toyota Entune

Dial

Subject: “Dial 801 - 520 - xxxx”
System: “801 - 520 - xxxx, please say dial, correction, remove, or continue adding numbers”
Subject: “Dial”
System: “Dialing”

Contacts Call

Subject: “Call Joel Cooper cell”
System: “Joel Cooper, mobile, dialing”

Play CD (CD function not available, satellite function was substituted)

Subject: “Satellite”

Tune Radio

Subject: “Tune to 98.7 FM”

Mercedes COMAND

Dial

Subject: “Dial number”
System: “Please say the number”
Subject: “801-520-xxxx”

System: “801-520-xxxx”
Subject: “Okay”
System: “Dialing”

Contacts Call

Subject: “Call”
System: “Please say the name”
Subject: “Joel Cooper, Cell”
System: “Joel Cooper, mobile, accepted, Joel Cooper, mobile, dialing”

Play CD

Subject: “CD”

Tune Radio (See 22_C5 4:45)

Subject: “Radio”
Subject: “FM”
System: “99.5”

Hyundai BlueLink

Dial

System: “Please say a command after the beep”
Subject: “Dial number”
System: “Which number would you like to dial? Or say international”
Subject: “801 - 520 - xxxx”
System: “801 - 520 - xxxx, add numbers or say dial, correction, delete”
Subject: “Dial”
System: “Dial”

Contacts Call

System: “Please say a command after the beep”
Subject: “Call Joel Cooper cell”
System: “Joel Cooper on cell phone, would you like to call this contact?”
Subject: “Yes”

Play CD

System: “Please say a command after the beep”
Subject: “CD”
System: “CD”

Tune Radio

System: “Please say a command after the beep”
Subject: “98.7”
System: “98.7 FM”