MIT/CSRC Project on Assessing the Demands of Voice Based In-Vehicle Interfaces

Technical Report 2015-14

Phase II Experiment 3 - 2015 Toyota Corolla (2015b)

Finial Report: November 28, 2015
Initial Submission: November 16, 2015

Bruce Mehler, Bryan Reimer, Jonathan Dobres, & Joseph F. Coughlin

Abstract – Experiment 3 is the third in a series of three studies designed to develop data to support exploring the generalizability of, and extend upon, the findings on the demands of production level voice-command systems from the MIT AgeLab's Phase I CSRC work that was undertaken in a 2010 Lincoln MKS. Self-report, eye glance, physiology (heart rate and skin conductance), driving performance, and task performance data were collected in a 2015 Toyota Corolla across a final analysis set of 48 participants meeting the age and gender distribution of NHTSA (2013) guidelines. Interactions considered consisted of baseline single-task highway driving, voice-based destination address entry into the navigation system, voice-based point of interest (POI) selection, and voice-based phone contact calling. Manual radio tuning and the auditory-vocal n-back were also included as reference tasks.

Lead Project Contact:

Bryan Reimer, Ph.D.
Phone (617) 452-2177
reimer@mit.edu

REPORT DOCUMENTATION PAGE

RECEIVING ORGANIZATION	REPORT DATE	REPORT TYPE + DATES COVERED
Toyota Collaborative Safety Research Center	November 28, 2015	Technical Report

TITLE AND SUBTITLE	FUNDING NUMBERS
Assessing the Demands of Voice Based In-Vehicle Interfaces: Phase II Experiment 3 - 2015 Toyota Corolla (2015b)	
AUTHOR(S) Bruce Mehler, Bryan Reimer, Jonathan Dobres, & Joseph F. Coughlin	

PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	PERFORMING ORGANIZATION REPORT NUMBER
MIT AgeLab Massachusetts Institute of Technology 77 Massachusetts Avenue, E40-279 Cambridge, MA 02139 USA	Technical Report 2015-14

SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	SPONSORING/MONITORING AGENCY REPORT NUMBER
Toyota Collaborative Safety Research Center Collaborative Safety Research Center Toyota Technical Center USA, Inc. 1555 Woodridge Ave Ann Arbor, MI 48105	

SUPPLEMENTARY NOTES

DISTRIBUTION/AVAILABILITY STATEMENT	DISTRIBUTION CLASSIFICATION

ABSTRACT

Experiment 3 is the third in a series of three studies designed to develop data to support exploring the generalizability of, and extend upon, the findings on the demands of production level voice-command systems from the MIT AgeLab's Phase I CSRC work that was undertaken in a 2010 Lincoln MKS. Self-report, eye glance, physiology (heart rate and skin conductance), driving performance, and task performance data were collected in a 2015 Toyota Corolla across a final analysis set of 48 participants meeting the age and gender distribution of NHTSA (2013) guidelines. Interactions considered consisted of baseline single-task highway driving, voice-based destination address entry into the navigation system, voice-based point of interest (POI) selection, and voice-based phone contact calling. Manual radio tuning and the auditory-vocal n-back were also included as reference tasks.

SUBJECT TERMS	NUMBER OF PAGES
Voice-command, visual-manual demand, TEORT, driver-vehicle interface, DVI, HMI, driver distraction, attention management, workload, cognitive demand, guidelines, glance behavior, heart rate, skin conductance, driving performance.	62

Recommended Citation: Mehler, B., Reimer, B., Dobres, J., & Coughlin, J.F. (2015). *Assessing the Demands of Voice Based In-Vehicle Interfaces - Phase II Experiment 3 - 2015 Toyota Corolla (2015b).* MIT AgeLab Technical Report 2015-14 (November 28, 2015). Massachusetts Institute of Technology, Cambridge, MA.

Glossary of Terms (or Acronyms)

ANOVA	analysis of variance
CAMP	Crash Avoidance Metrics Partnership
CAN	controller area network
CSRC	Toyota Collaborative Safety Research Center
DVI	driver vehicle interface
ISO	International Organization for Standardization
M	mean
MIT	Massachusetts Institute of Technology
NHTSA	National Highway Transportation Safety Administration
SCL	skin conductance level
SD	standard deviation
SE	standard error
SWR	steering wheel reversals
The Alliance	Alliance of Automotive Manufacturers
TEORT	total eyes-off-road time - the sum of all glances off the forward roadway during a specific period

MIT/CSRC Project on Assessing the Demands of Voice Based In-Vehicle Interfaces

Phase II Experiment 3 - 2015 Toyota Corolla (2015b)

Contents

Introduction

Intent of Present Report

The experiment considered here was developed as the third in a series of three studies intended to explore the generalizability of the findings on production level voice-command systems from our Phase I CSRC work that was undertaken in a 2010 Lincoln MKS (Reimer, Mehler, Dobres & Coughlin, 2013; Mehler, Reimer, Dobres, McAnulty, Mehler, Munger & Coughlin, 2014). The primary focus of the current set of studies is to consider what may be learned about voice systems by developing comparable data across multiple vehicles. The interpretive portion of the report is intentionally brief. A more comprehensive analysis and discussion is envisioned in future reporting considering findings from all of the studies. For additional background on the overall project, please see the introduction of the technical report on Phase II Study 1 (2015-6) (Mehler, Reimer, Dobres, McAnulty, & Coughlin, 2015).

Protocol for Phase II Study 3

The protocol and driving route for Phase II Studies 2 & 3 were identical (see Figure 1) and, as in Study 1, the voice-command based tasks considered consisted of: full address entry into the navigation system, selection of specified points of interest (POIs), and phone contact calling. Further, reference tasks consisted of visual-manual radio tuning (single button press preset station selection and the more intensive radio reference running task) and the n-back audio-vocal-cognitive calibration reference task at the 0-, 1-, and 2-back levels.

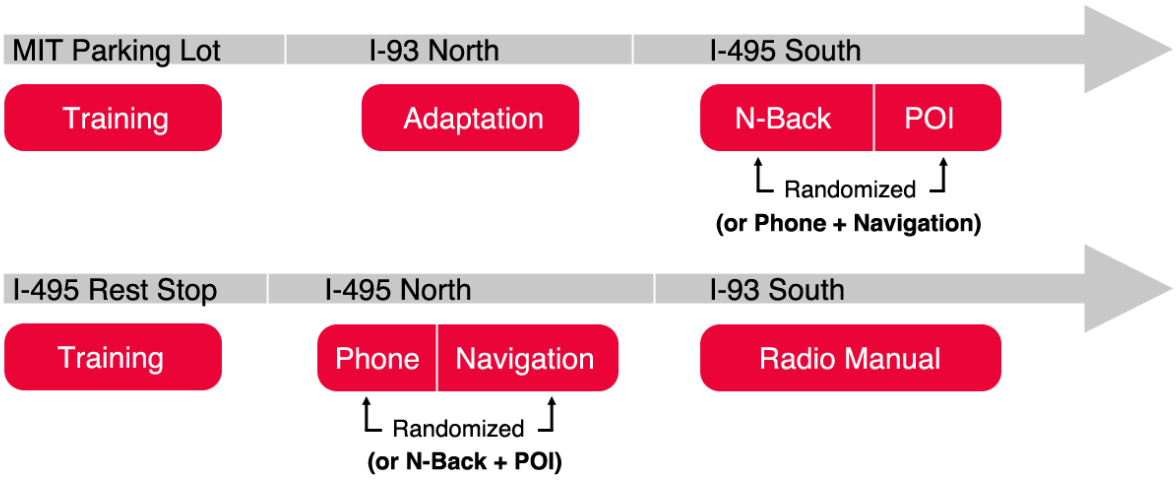

Figure 1: Overview of the Experimental Protocol for Phase II Studies 2 & 3

Contact phone calling was identical to that undertaken in previous Phase II studies (i.e. calling 4 specified contacts, 2 “easy” and 2 “hard”) (see Appendix E). The full address destination entry task consisted of the same 3 addresses employed in Study 2 (2 study specified addresses followed

by the participant's home address) (see Appendix E). As in Study 2, adjustments were made to the POI targets to match the "categories" and address book location of those in Study 2. Radio Manual Tuning was identical to that undertaken in Study 2.

Methods

Participants

Consistent with Phase II Studies 1 & 2, the research plan for Study 3 called for obtaining 48 usable participant cases equally balanced by gender and across four age groups (20-24, 25-39, 40-54 and 55-69 years). This corresponds to the age distribution recommended by NHTSA (2013) for assessment visual-manual driver distraction for in-vehicle electronic devices, with the exception of not recruiting 18 & 19 year-olds. Participants again needed to meet the following criteria:

- A valid driver's license for more than three years
- Driving on average three or more times per week
- Being in self-reported reasonably good health for their age and meeting a set of health exclusion criteria (see Appendix F)
- Clearly understanding and speaking English
- No police reported accident in the past year
- Not actively using any medications causing drowsiness
- Not having been a participant in an AgeLab on-road driving study in the past 6 months

Apparatus

Voice-Interface

The infotainment interface studied was the standard production Entune Premium Audio with navigation available on the 2015 Toyota Corolla. The software for the voice system was updated on May 14, 2015. Listed version information is as follows:

Model ID: 15TDANNA-CA01
SAM ID: AEAICDOSBG
Hard No. 86100-02101

Operating System: CU.10.71
Device Driver: CU.10.71
Audio: CU.10.71
Navigation: CU.10.71
Bluetooth: CU.10.71
GUI: CU.10.71
Middle ware: CU.10.71
Kernel: CU.10.71
Voice Recognition: CU.10.71

Figure 2: Layout of the 2015 Toyota Corolla infotainment DVI.

The layout of the 2015 Toyota Corolla infotainment DVI showing the main components of the system interface appears above in Figure 2. A closer view of the center console screen appears in Figure 3 below.

Figure 3: Closer view of the infotainment console in the center cluster of the 2015 Toyota Corolla. Screen shows call being placed after voice-based selection of a number from a phone contact list.

Figure 4: Location of the push-to-talk button on the steering wheel.

Figure 5: Engagement with the manual radio interface used as a visual-manual demand reference task. Note the vertical orientation of the touch screen located preset "buttons" on the left side of the screen.

Figure 6: Some steps in the Entune interface allowed users to make selections either verbally or by touching a choice on screen. Some other vehicles tested required tactile responses.

Instrumentation

The vehicle was instrumented with a custom data acquisition system for time synchronized recording of data from:

- vehicle information via the controller area network (CAN) bus,
- a Garmin 18X Global Positioning system (GPS) unit,
- a MEDAC System/3™ physiological monitoring unit to provide EKG and skin conductance level (SCL) signals,
- video cameras,
- a wide area microphone to capture driver speech and audio from the vehicle's speech system.

The five video cameras provided views intended to capture the driver's face for primary glance behavior analysis, the driver's interactions with the vehicle's steering wheel and center console, the forward roadway (narrow and wide-angle images), and a rear roadway view. Data were captured at:

- 10 Hz for the CAN bus and GPS,
- 30 Hz for the face and narrow forward roadway cameras,
- 15 Hz for the remaining cameras,
- 250 Hz for the physiological signals to support EKG feature extraction for heart beat interval detection.

Phone connectivity was supported by pairing a Samsung Galaxy S4 smartphone (model SCH-1545) to the vehicle's embedded system via the vehicle's Bluetooth wireless interface.

Procedure

As was the case in the Phase I studies, Phase II Studies 1 - 3 considered a range of voice-command based tasks along with established visual-manual reference tasks (radio tuning) and auditory-vocal calibration reference tasks (multi-level n-backs). In specific, the following tasks were studied:

- Voice-based interface tasks
    - › Full address destination entry
    - › Point of interest (POI) destination entry
    - › Cancel navigation (for each of above)
    - › Stored phone contact calling (single & multiple phones per contact)

- Visual-manual tasks (radio tuning)
  - Single button press preset selection – Radio Easy
  - Specified station manual radio tuning – Radio Hard
- N-back (auditory-vocal-cognitive calibration reference task)
  - 3 demand levels (0, 1, & 2-back)

Details of each of the tasks and the steps required to complete each task with the infotainment system in the 2015 Toyota Corolla are provided in the Appendix E. Training procedures were similar to those provided to the "trained" group in Phase I, Study 2 (Mehler, et al., 2014). Task evaluation took place under actual highway driving conditions on Interstate 495 (I-495) outside the greater Boston area and on I-93 south of the interchange with I-495, heading back towards Boston. The highway sections of I-495 utilized consisted of three travel lanes in each direction, are bordered largely by forest, and have a posted speed limit of 65 mph.

As in previous work, participants received training on how to complete each of the tasks prior to being asked to consider engaging with them while driving. Training took place in a parking lot at MIT and at a rest stop on I-495. The content of the training in each location depended on a counter-balanced ordering of tasks across the sample (see Figure 1). The n-back and POI entry tasks were paired together as were the contact phone calling tasks and destination address entry into the navigation system. This was done so that the secondary task portions of the south-bound and north-bound portions of the experimental drive on I-495 were of relatively equivalent duration. Half the participants experienced the n-back and POI entry tasks during the first portion of the I-495 drive and half on the second. The ordering of the tasks with-in a pairing was randomized across the sample. The manual radio tuning reference tasks were always presented last on the return route on I-93. Placing the manual radio reference tasks last was considered reasonable in the task ordering as the basics of manual radio turning is a relatively well learned task across most drivers.

Questionnaire based assessment of the participants' experiences with the tasks (workload, etc.) were obtained at the rest stop for tasks completed up to that point. Experiences related to the remaining tasks were obtained back at MIT in the parked vehicle prior to reentering the research building where additional questionnaire based evaluations were obtained and final debriefing took place.

Data Reduction and Analysis

Single task driving reference periods were calculated for 2 minutes of "just driving" prior to a recorded audio message indicating the start of a new task period on the I-495 portion of the drive (see Figure 1). There were four such baseline periods per participant. These were just prior to: the n-back, destination address entry, contact phone calling, and POI entry task periods (8 minutes total). Metrics were calculated and the mean values across the baseline periods were

used as an overall baseline/“just driving” reference. For task periods, values for each dependent measure were calculated per trial and mean values across trials were used for analytic purposes. All trials with usable data were included regardless of whether errors due to a user or system issues occurred. Including trials with errors in the analysis was seen as more representative of the actual user experience than only considering error-free trials. (See Reimer et al. 2013 for a consideration of the extent to which including trials with and without errors impacted the overall pattern of results in the first MKS study.)

As in the Phase I studies, eye glance measures were quantified following ISO standards (ISO 15007-1, 2002; ISO 15007-2, 2001) with a glance to a region of interest defined to include the transition time to the object / region. In the manual coding of video images, the timing of glance is labeled from the first video frame illustrating movement to a “new” location of interest to the last video frame prior to movement to a “new” location. Glance data were manually coded based on video of the driver following the taxonomy and procedures outlined in Reimer et al. (2013, Appendix G). Software that allowed for rapid frame-by-fame review and coding is now available as open source (Reimer, Gruevski, & Couglin, 2014). Each task period of interest was independently coded by two evaluators. Discrepancies between the evaluators (the identification of conflicting glance targets, missed glances, or glance timings that differed by more than 200ms) were mediated by a third researcher. (Smith, Chang, Glassco, Foley et al. (2005) is recommended for a discussion of the importance of multiple coders.)

Physiological data were handled in the same manner as the in the Phase I studies. In specific, R-wave peaks in the EKG signal were identified to calculate inter-beat intervals and calculate instantaneous heart rate using software developed at the MIT AgeLab. Consistent with existing standards (Task Force, 1996), automated detection of R-wave peaks were visually reviewed and misidentified and irregular intervals manually corrected. Another MIT AgeLab developed data processing package removed high-frequency noise in the skin conductance signal as per Reimer and Mehler (2011) and substantive identified motion artifacts were manually edited.

Statistical analyses were performed in R (R Core Team, 2014) and an alpha level of 0.05 was used for statistical significance assessment. Due to the non-normal distribution of the data and/or the use of ratio data (percentages) for several dependent measures, in many cases non-parametric statistics - the Wilcoxon signed rank test and the Friedman test - were used (similar to the t-test and repeated-measures ANOVA, respectively). These tests have been shown to be more robust against Type I error in cases where data are non-normal (Conover & Iman, 1981; Friedman, 1937).

Results

Analysis Sample

A total of 70 participants were enrolled in the study. A breakdown of the development of the analysis sample is provided in Figure 7 below. To be included in the analysis sample, participants had to demonstrate the ability to complete each task under controlled conditions in the parking lot and engage in each task type during the drive. Further, driving performance data from the CAN bus and video recordings of sufficient quality to code eye glance behavior had to be available. Usable physiological recordings were considered desirable, but were not required in this sample. In addition, cases were excluded if non-optimal weather conditions (e.g. heavy rain) or heavy traffic was encountered. Finally, the research associate in the vehicle was able to withdraw participants from the study due to erratic or otherwise unsafe driving behavior.

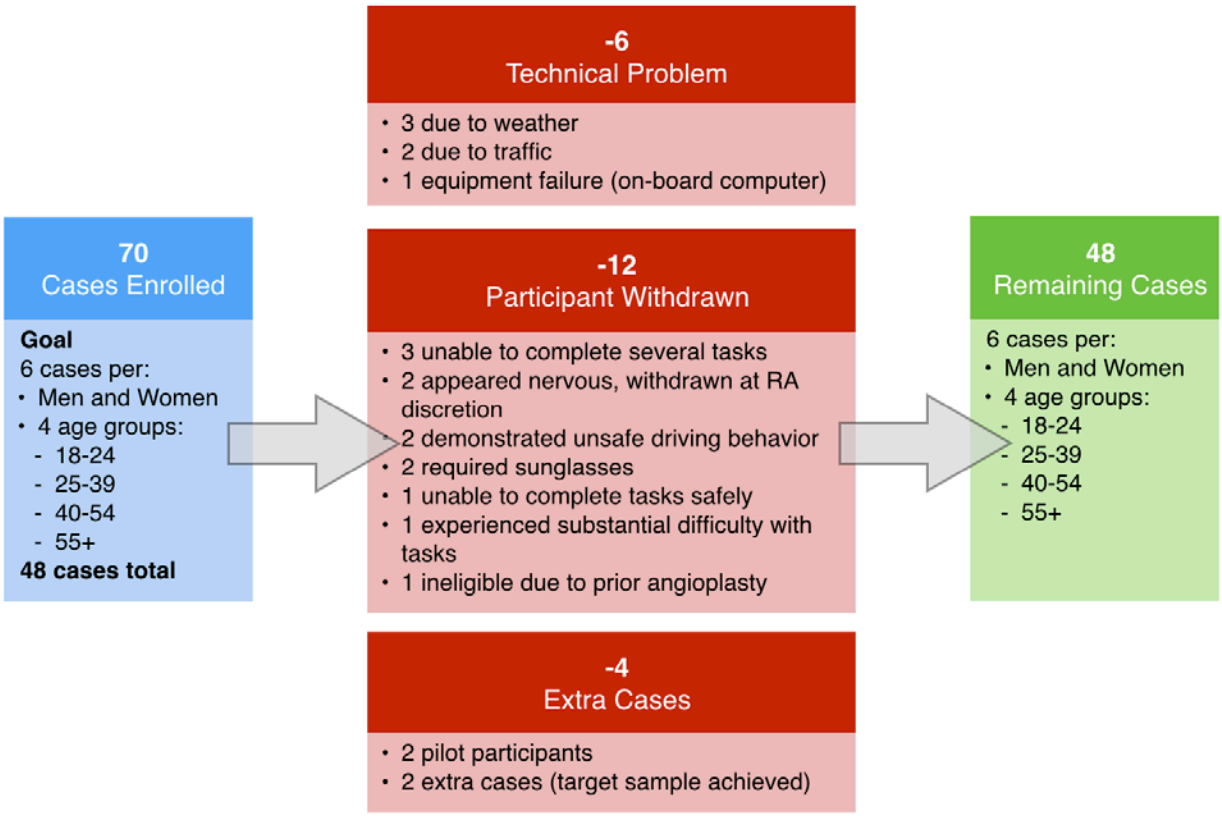

Figure 7: Summary presentation of the development of the analysis sample for the study.

The analysis sample of 48 was balanced between the two genders and distributed across the four NHTSA-recommended age groups (18-24, 25-39, 40-54, and 55+), six participants per group. The one variation from the NHTSA age guidelines was that 18 and 19 year olds were not recruited. Demographic summary statistics are given in Table 1. Age distributions did not differ

significantly between genders (t(46) = 0.094, p = 0.925). The age distribution across Phase II Studies 1, 2 and 3 was quite similar as detailed in the table.

Table 1: Comparison of summary statistics by gender for Phase II Studies 1, 2, and 3.

Study 1	**Mean Age**	**SD**	**Minimum**	**Maximum**	**N**
Female	41.8	16.6	22	68	24
Male	39.6	16.4	21	68	24
					48
Study 2					
Female	38.9	15.6	20	65	24
Male	39.8	15.3	20	67	24
					48
Study 3					
Female	39.8	17.0	20	69	24
Male	40.3	16.7	20	67	24
					48

Organization of Results

This report presents data on self-reported workload, task completion time, off-road glance metrics (mean single glance duration, percentage of glances greater than 2 seconds, cumulative total glance time), physiological metrics (heart rate and SCL), and driving performance metrics (mean and SD of velocity, steering wheel reversals). In the sections below, figures provide graphical summaries of the data and present relevant statistical analysis. An error analysis follows. A set of tables in Appendix A provide descriptive statistics (means & standard errors) for each of the dependent measures covering baseline driving, address entry, and POI entry (Table 6), contact phone calling and radio tuning (Table 7), and the n-back tasks (Table 8). Additional tables provided in Appendix B detail descriptive statistics broken down by trial for the address entry, POI selection, and contact phone calling tasks. Appendix C provides descriptive statistics for address entry and POI selection selectively collapsed across trials to support comparison of data across other studies where the total number of trials differed.

Self-Reported Workload

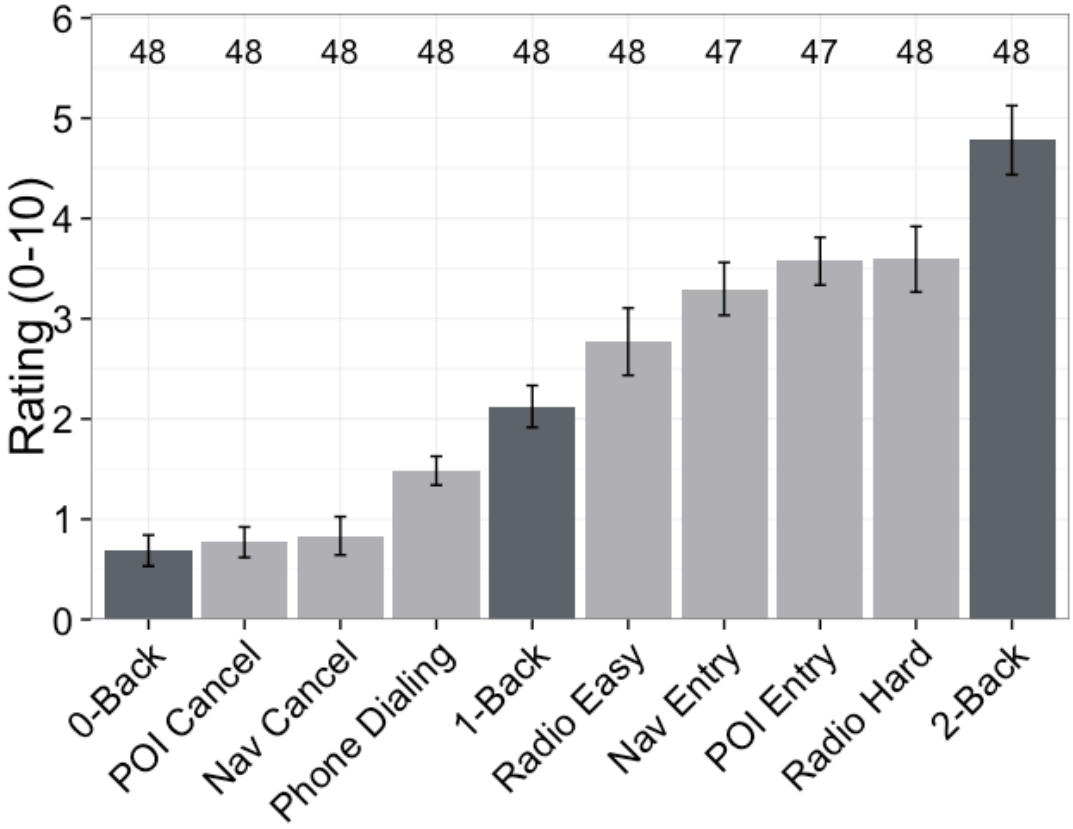

Figure 8: Self-reported workload ratings for each task under study, in ascending order. Bars represent mean performance while error bars represent the mean-adjusted standard error. Numbers at the top of each column represent the number of data points available per task. Darker bars indicate the n-back cognitive loading tasks.

Participants were asked to rate how much workload they experienced while engaged in each task while driving on a scale of 0 (low) to 10 (high). Summary statistics are presented graphically in Figure 8. Mean and standard error values are detailed in Appendix A. Across the sample, ratings for a total of two tasks were not given (across two different participants). Friedman tests across tasks consider ratings from the participants for which all ratings are available.

Workload ratings differed significantly across all tasks ($X^2(9) = 209.8$, $p < 0.001$). Similarly, the three escalating levels of the n-back resulted in significantly different workload ratings ($X^2(2) = 84.6$, $p < 0.001$). Workload ratings also differed significantly across the different interface tasks ($X^2(6) = 140.6$, $p < 0.001$). Note that the three levels of the n-back fall in a range across the perceived levels of interface task difficulty, with the 0-back and 2-back bookending the scores, and the 1-back falling in the middle range of the ratings. This finding has been consistent across all studies in Phase II of this research program.

Task Completion Time

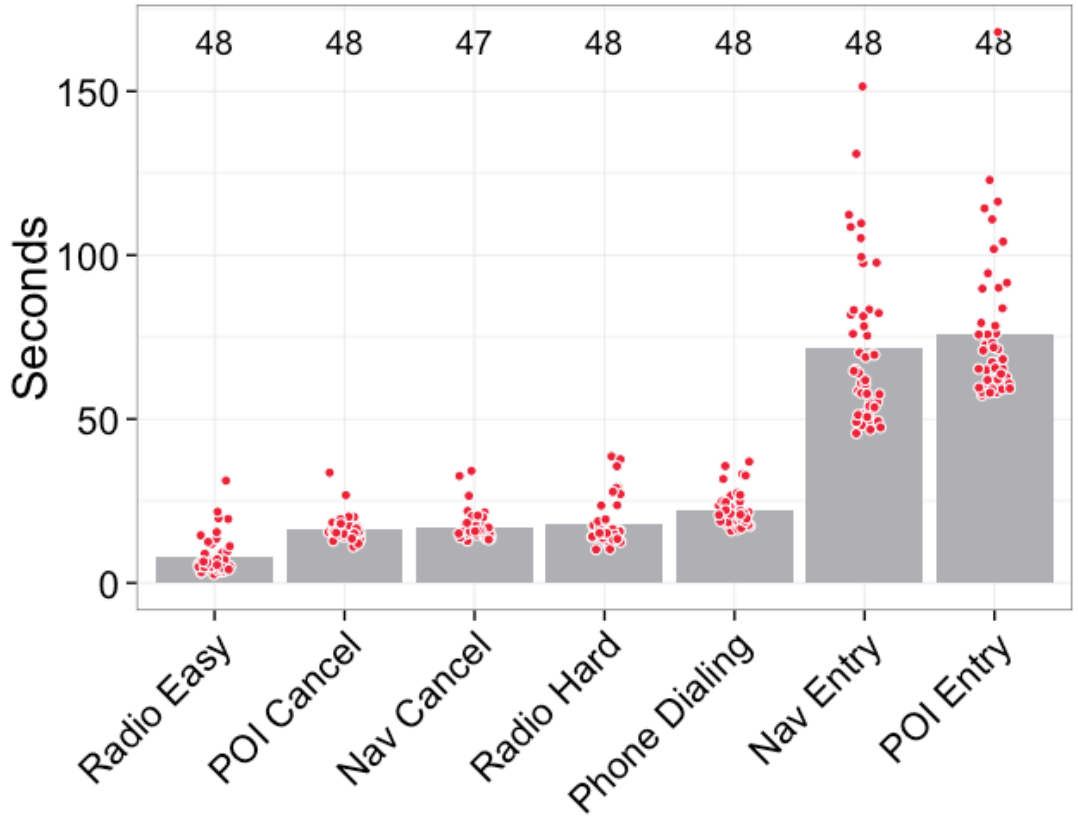

Figure 9: Task completion times for each task under study. Note that completion times for the n-back tasks are not included, as these always had a fixed duration of 30 seconds.

Task completion times varied significantly across task types ($X^2(6) = 234.0$, $p < 0.001$). While many tasks were completed, on average, in less than 30 seconds (manual radio tuning (17.9s), voice contact phone calling (22.3s), and cancelation of destinations in the navigation and POI systems), navigation full address entry and the POI entry tasks required markedly longer times to complete (71.4s and 75.8s, respectively). This general pattern has been quite consistent across all of the infotainments systems studied in this project, with the radio easy task being the shortest task, the radio hard manual tuning reference task and the phone contact calling tasks being intermediate, and full address entry and POI selection being appreciably longer. It can be observed that phone contact calling, address entry, and the POI selection tasks were, on average, nominally shorter in the Corolla than the Impala (26.3s; 88.1s; 93.6s) and the CLA (27.4s; 72.6; 99.0s).

Physiology

Heart Rate

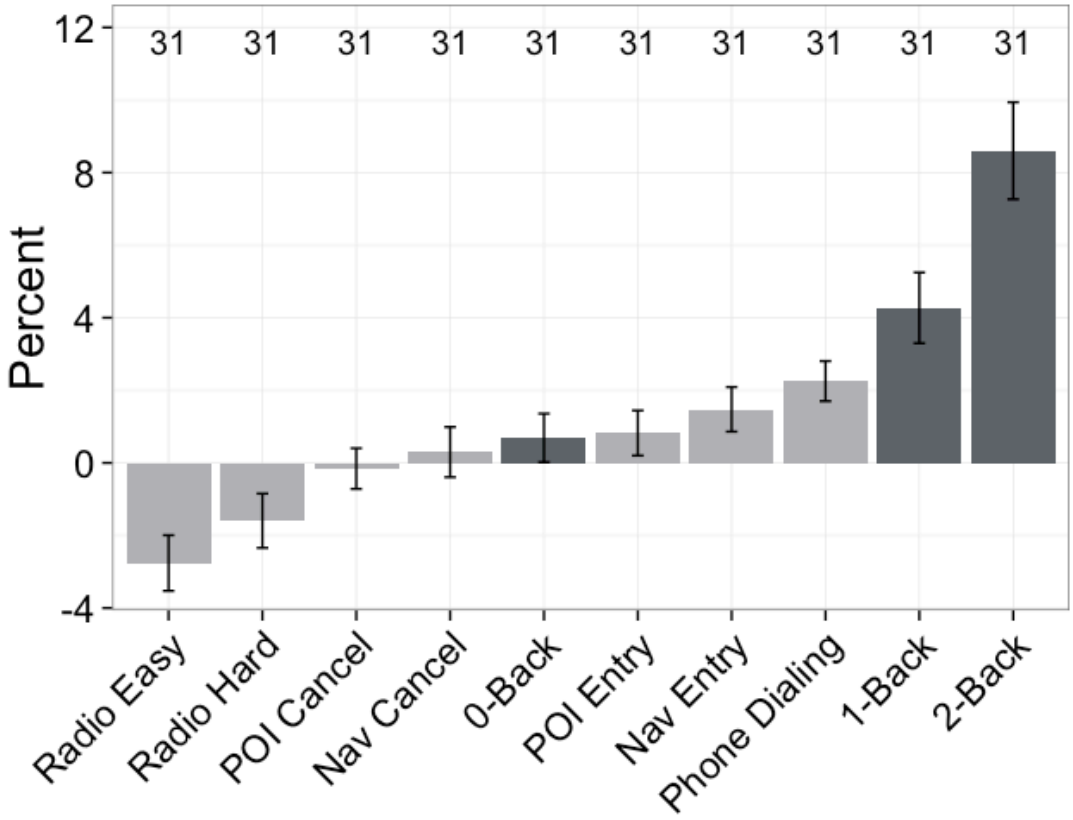

Figure 10: Percentage change in mean heart rate relative to mean baseline driving for each task under study. Labeling as in Figure 8.

Measures of mean heart rate were normalized as the percentage change from the mean heart rate observed during baseline single-task driving periods. (Alternate representation in beats per minute is presented in Appendix D.) Statistical testing on heart rate data is limited to 31 participants in the analysis due to recording issues and/or excessive artifact in several cases. Changes in heart rate differed significantly across all task periods ($X^2(9) = 79.6$, $p < 0.001$). Changes were also significantly different across the levels of the n-back ($X^2(2) = 25.8$, $p < 0.001$), as well as when considering only the interface-based tasks ($X^2(6) = 35.8$, $p < 0.001$). The n-back tasks produced significantly greater changes in heart rate compared to the interface tasks (interface M = 0.05%, n-back M = 4.52%; W = 244, $p < 0.001$, Wilcoxon test of mean interface heart rate vs. mean n-back heart rate).

The stair-step pattern seen in the different levels of the n-back task, as well as the mean percentage change values observed, are generally consistent with values observed in the previous on-road studies in the Lincoln MKS (Mehler et al., 2014), Chevrolet Impala, and Mercedes CLA.

The mean change in heart rate values in those studies for the 0-, 1-, and 2-back were 3.2%, 7.7%, and 10.9% in the Lincoln; 3.4%, 7.1%, and 10.5% in the Impala, 1.5%, 5.5%, and 10.5% in the CLA. The corresponding values here in the Corolla were 0.7%, 4.3%, and 8.6%, thus showing similar mean values, though slightly lower than in previous studies.

In the Corolla, as in the other vehicles studied in this project, heart rate changes, where present, were generally modest compared to the n-back reference tasks. On average, heart rate increased by 2.3% during the voice-based phone contact calling task, which is similar to what was seen in the Impala (1%) and CLA (2.1%). These values fall well below the apparent heart rate arousal level associated with the 1-back task (4.3%).

Skin Conductance Level

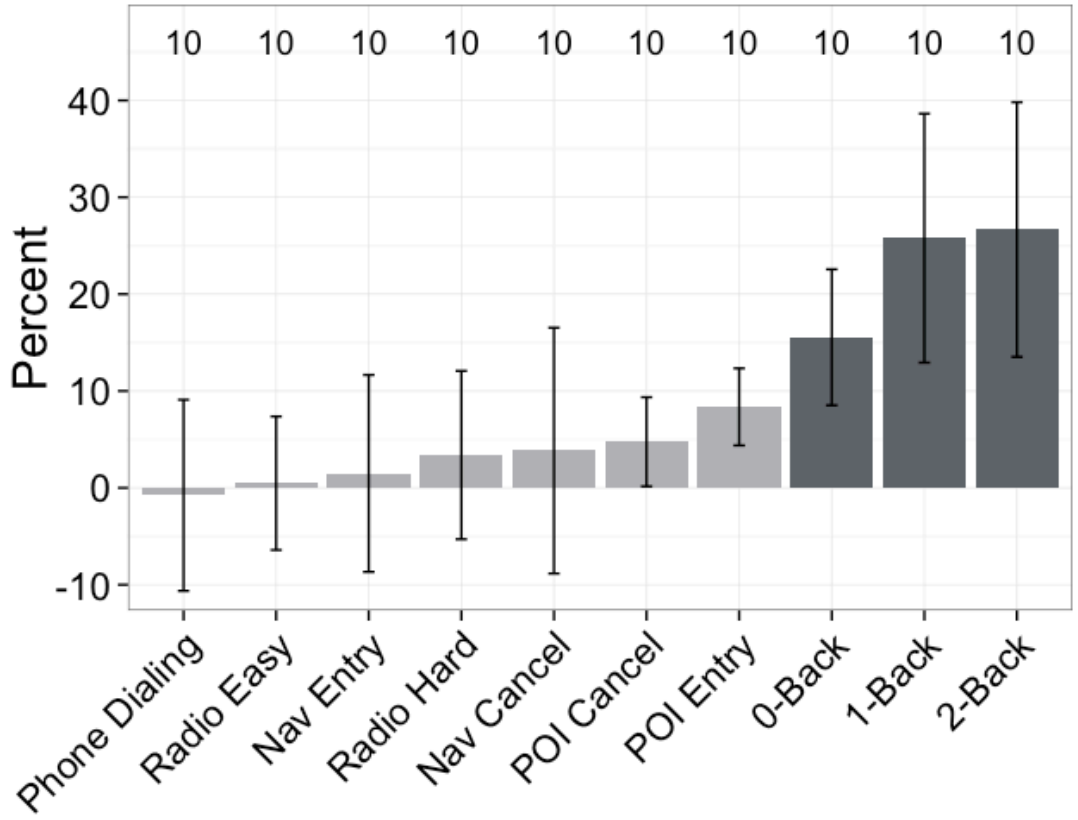

Figure 11: Percentage change in mean skin conductance level relative to mean baseline driving in all tasks periods under study. Labeling as in Figure 8.

Skin conductance measures could not be analyzed for 38 participants due to technical issues and/or high levels of motion artifact in the recordings. This unusually high level of unusable cases was traced to a subtle fault in a skin conductance sensor that lead to intermittent signal quality. This was not detected until a signal quality review was conducted on recorded data near the end of the study period. An analysis of the remaining 10 participants follows (note that

Appendix D, which displays raw skin conductance measures, contains slightly more data points since the figure above shows only participants with complete data across all tasks).

As with heart rate, skin conductance level measurements were normalized against measurements from the baseline driving reference period. (Alternate presentation in physiological units is presented in Appendix D.) While the data do demonstrate clear patterns in skin conductance levels, owing to the small available sample and relatively high variability of this metric, these measures do not achieve statistical significance when considering all tasks ($X^2(9) = 8.8$, $p = 0.457$), interface tasks alone ($X^2(6) = 4.8$, $p = 0.570$), or n-back tasks alone ($X^2(2) = 0.2$, $p = 0.905$).

Keeping the limitations of the current SCL dataset in mind, as observed for heart rate, SCL showed a general increase across the increasing levels of working memory demand with the n-back task with mean values of: 15.54% 0-back, 25.78% 1-back, and 26.27% 2-back (see in Table 8 in Appendix A for full descriptive statistics). The corresponding mean values in the Impala sample were: 7.64% 0-back, 9.00% 1-back, and 14.43% 2-back. In the CLA these were: 11.15% 0-back, 15.48% 1-back, and 17.14% 2-back. Considering the DVI tasks alone, changes in skin conductance from baseline in the Corolla ranged from -0.77% during the Phone Dialing task to 26.67% during 2-Back. In the Impala, mean changes in skin conductance for the comparable tasks ranged from 11.64% for the Phone Dialing task to 14.43% for the 2-Back. A 13.7% increase in SCL over baseline in the Impala was observed for the manual radio hard task. In the CLA, changes ranged from -10.84% during the easy radio tuning task to 2.07% during POI entry.

Glance Behavior

Following NHTSA guidelines (2013), glance behavior is quantified in the section that follows considering glances off-the-forward-roadway. Descriptive statistics (mean and standard deviation) for both off-the-forward-roadway measures and glance-to-device measures are provided in the tables in Appendix A. A consideration of total glance-time-to-device is presented in Appendix D. Glance coding was carried out as per the description in the section on *Data Reduction and Analysis*.

Mean Single Off-Road Glance Duration

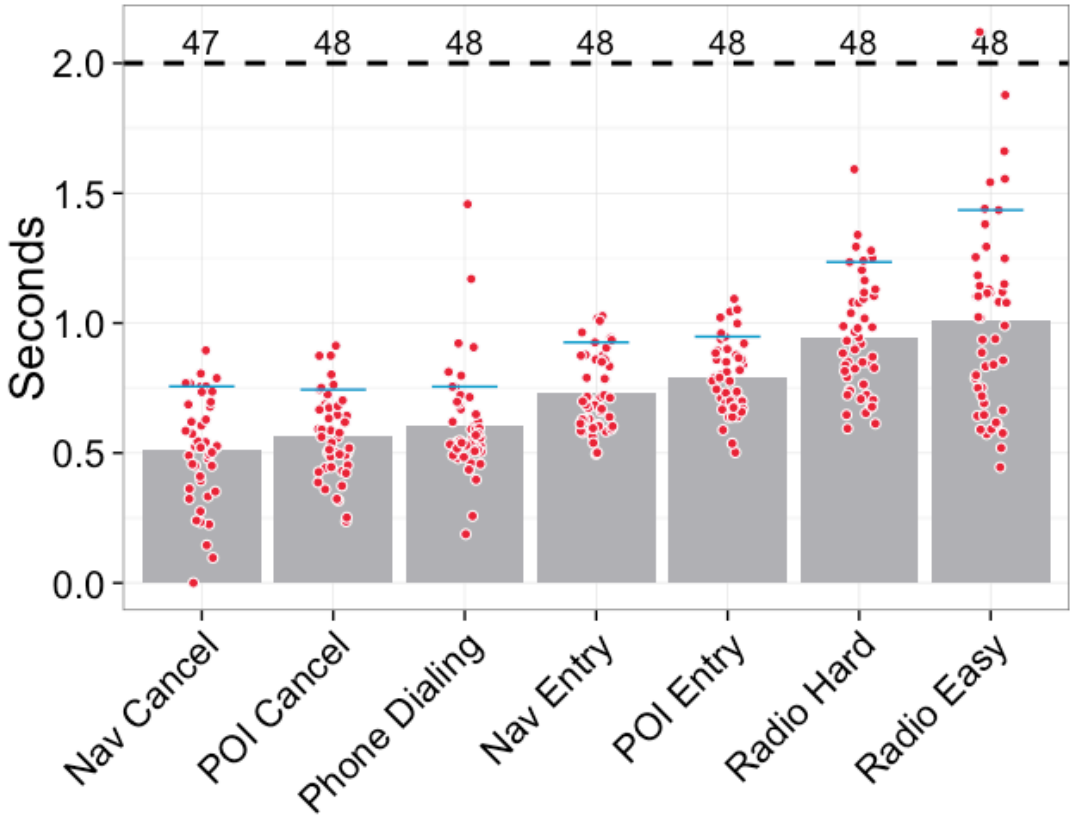

Figure 12: Mean single off-road glance duration during each task under study. Points represent individual participant performance and bars represent task means. Short line segments indicate the 87.5th percentile of performance (42/48 participants), while the large dashed line represents NHTSA's recommended criterion for this metric. For any task, if the short line segment is below the large dashed line, that task meets NTHSA's recommended criteria for a visual-manual interface if it was to be applied to these tasks and data the collection methodology employed.

Mean single glance duration differed significantly across tasks ($X^2(6) = 191.2$, $p < 0.001$), though the actual range of mean glance durations was relatively small, with a mean of 0.51s for the Navigation Cancel task and a mean of 1.01ss for the Radio Tuning (Easy) task. All tasks met the NHTSA-recommended glance duration threshold of 2.0s.

The relative ordering of address entry and the manual radio hard tuning task are consistent with what was seen in the Lincoln MKS, Chevrolet Impala, and Mercedes CLA, with mean single glance durations being shorter during voice-based address entry than during the manual radio tuning reference task. Mean values in the Corolla were 0.73s address entry and 0.94s radio hard tuning and corresponding values in the other vehicles were, respectively: CLA (0.72s; 0.81s), the Impala (0.77s and 0.88s for the corresponding number of trials), and MKS (0.82s and 1.03s).

Percentage of Single Off-Road Glances Greater than 2.0s

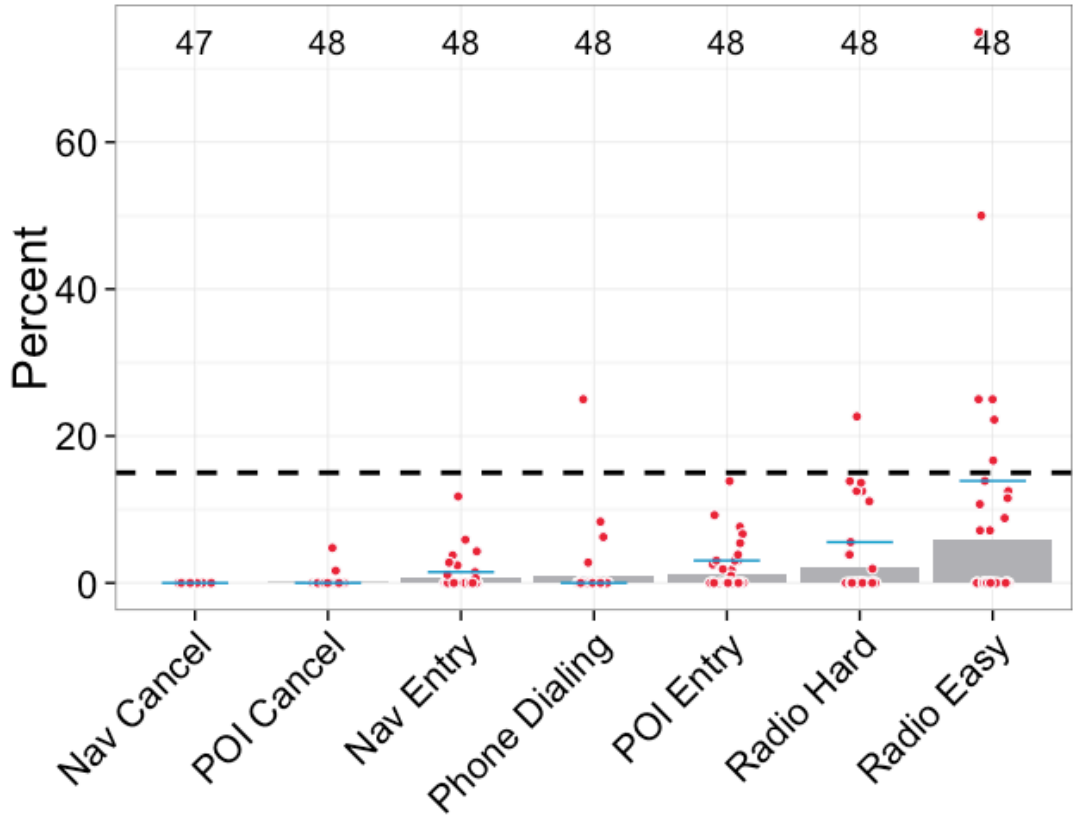

Figure 13: Percentage of long duration off-road glances (in excess of 2.0s) in all tasks under study. Labeling as in Figure 12.

The percentage of long duration glances differed significantly across tasks ($X^2(6) = 40.4$, $p < 0.001$), though, as in previous studies, the range was fairly small (0.0% during the Navigation Cancel task to 5.95%% during Radio Tuning (Easy) task), with many participants having no long duration glances during task periods. All tasks under consideration met the NHTSA-recommended guideline of no more than 21/24 (87.5percentile) of the sample showing more than 15% of long duration off-road glances, if this guideline for visual-manual interfaces was applied to these tasks and data collection methodology.

Total Eyes Off-Road Time (TEORT)

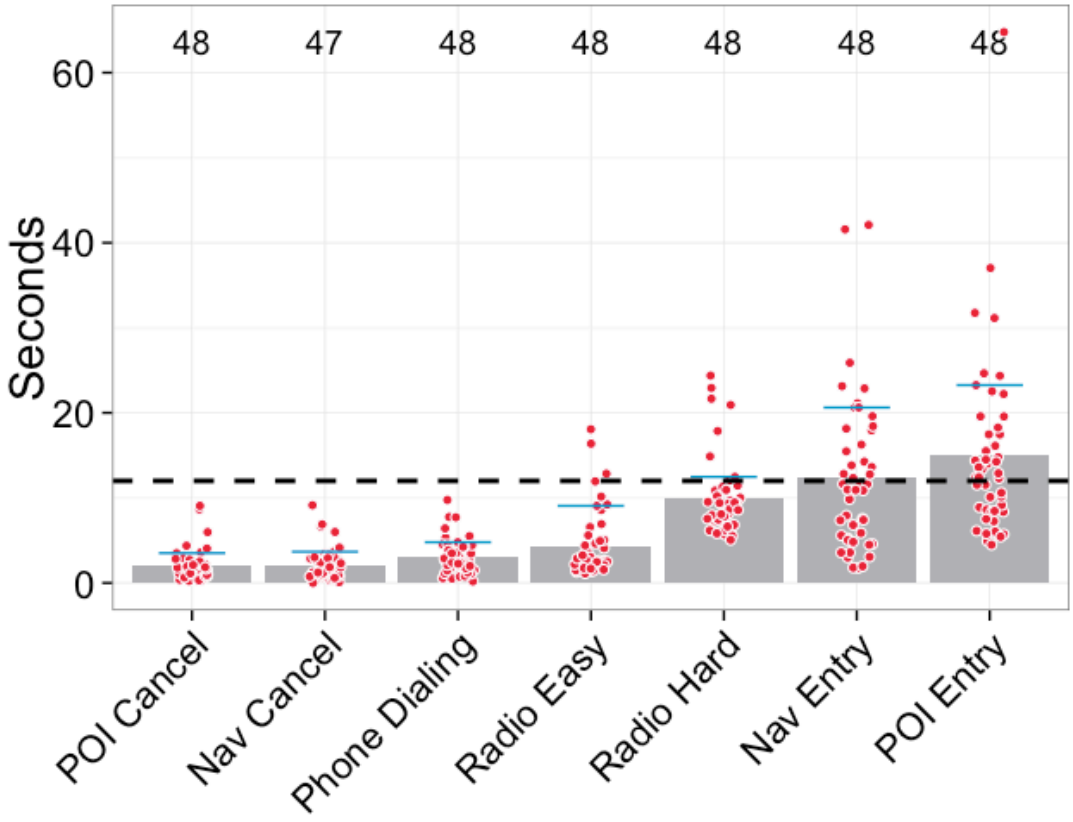

Figure 14: Cumulative off-road glance time for each task period under study. Labeling as in Figure 12.

TEORT, the cumulative duration of off-road glances during a task period, differed significantly across tasks ($X^2(6) = 203.6$, $p < 0.001$). While many tasks required less than 10 seconds of off-road glance time on average, the Navigation Entry and POI Entry tasks required more (12.40s and 14.98s, respectively). The corresponding values for the Impala Navigation Entry and POI Entry tasks (trials 1-3) were 22.8s and 24.6s, and in the CLA these were 19.0s and 35.4s, respectively. It can be observed then that both total task time and TEORT was nominally lower in the Corolla interface than in the Impala and CLA for these relatively comparable voice-involved tasks.

As was the case for the data collected in the Impala and the CLA, the Navigation Entry and POI Entry tasks would not meet the NHTSA (2013) recommended criterion of less than 12.0s of off-road glance time for 21/24 participants (87.5percentile) if this guideline for visual-manual interfaces was applied to these voice-involved multi-modal tasks and data collection methodology. The values for this 2015 model year interface are, however, much closer to this threshold than what was observed in the 2014 Impala and CLA models, and the 2010 Lincoln MKS.

While NHTSA's guidelines assess glance behavior in terms of the total time a driver's eyes are directed away from the forward roadway (TEORT), the earlier Alliance (2006) guidelines consider the total time during a task that a driver's eyes are directed to DVI related off-road glances and specify a 20 second criterion. This alternate way of looking at total task associated glance time is presented in Figure 28 in Appendix D. Note that the "radio hard" task as tested under these highway conditions falls within the guidelines established by the Alliance. Additional points related to how the radio manual tuning reference task was employed in the research will be consider in the Discussion.

Number of Glances Off-Road

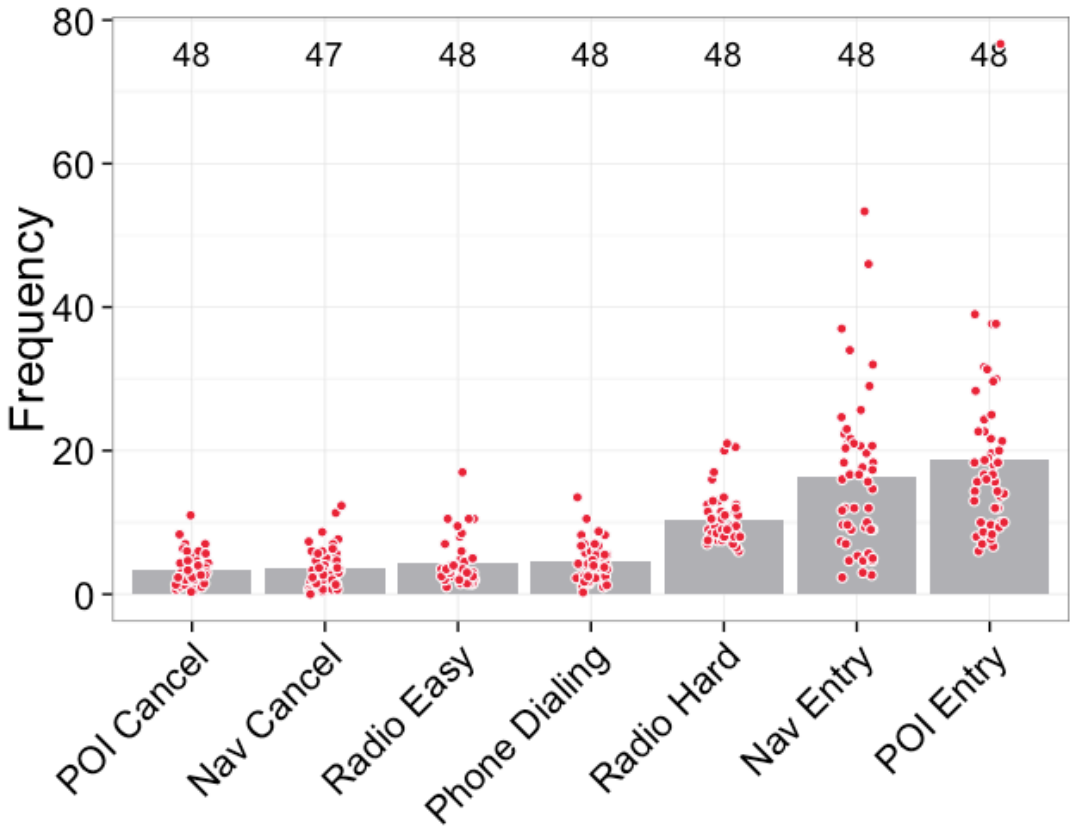

Figure 15: Number of glances off-the-forward-roadway for each task period under study. Labeling as in Figure 12.

The number of glances off-the-forward-roadway during a task period differed significantly across tasks ($X^2(6) = 198.2$, $p < 0.001$). Not surprisingly, the distribution of the number of such off-road glances was quite similar to the TEORT distribution (see Figure 14), with the total number of glances being highest for Navigation Entry (M = 16.44) and POI Entry (M= 18.83). This same relationship between number of glances and TEORT was observed in the Impala and CLA.

Glance Metric Monte Carlo Analysis

While this study examines the performance of a total of 48 participants (6 per age/gender cell), NHTSA's guidelines for the evaluation of visual-manual distraction recommend a balanced sample of 24 participants (3 per age/gender cell). It is possible that the larger sample size resulted in a pattern of pass/fail criteria that is not representative of what would be found with a smaller sample. To investigate this possibility, a Monte Carlo analysis was performed, in which 3 participants are sampled from every cell of 6, to produce a randomized subsample. Participants were sampled with replacement, meaning that a participant could be selected for inclusion in the same subsample multiple times, thus broadening the range of possible results.

Two thousand randomized subsamples were created, and their pattern of pass/fail criteria compared to the full reference sample. Table 2 presents the percentage of subsamples that agree with the pass/fail findings in the main reference sample. Subsample agreement was 100% for all metrics and tasks except for cumulative off-road glance time for the radio tuning tasks (96.65% for easy tuning and 62.65% for hard tuning), and for the percentage of long duration (>2.0s) glances during radio tuning easy tasks (61.95%).

Table 2: Percentage of subsamples with pass/fail criteria in agreement with the main reference sample.

	Percentage of Glances > 2.0s	**Mean Single Off-Road Glance Duration**	**Cumulative Off-Road Glance Time**
Navigation Cancel	100.00%	100.00%	100.00%
Navigation Entry	100.00%	100.00%	100.00%
Phone Dialing	100.00%	100.00%	100.00%
POI Cancel	100.00%	100.00%	100.00%
POI Entry	100.00%	100.00%	100.00%
Radio Easy	61.95%	100.00%	96.65%
Radio Hard	100.00%	100.00%	62.65%

Orienting Behavior

Analysis of the Corolla study data employed the same single coder methodology as the previous studies in which the coder watched muted video recorded from the camera mounted on the vehicle dashboard, facing the driver. This camera recorded the participant's face and upper torso, providing a clear view of their posture and head orientation. Participant's behaviors were coded according to the guide below. It should be noted that the analysis does not explicitly distinguish glances for visual confirmation from glances associated with orienting response (OR) behavior, and it is recognized that this is a partial confounding factor in the coding.

Table 3: Coding guide for orienting behavior including color codes used in graphic figures.

Category	Color Code	Description
Unknown	gray	Participant did not perform the task or its corresponding data are missing.
None	dark green	Participant exhibits no orienting response (OR) towards the center console display. This means there is no head tilting or leaning of the body towards the device (center console display).
Slight	light green	Participant exhibits some mild OR towards the device. The participant leans his/her head towards the device periodically throughout the task, or briefly leans his/her body toward the device.
Moderate	yellow	Participant exhibits a fair amount of OR towards the device. This means the participant leans his/her head or body towards the device or speaks directly at the device for a sustained period of time.
Prioritizing	red	Participant exhibits a clear and sustained OR toward the device. This means that the participant fully leans his/her head towards the device or repositions his/her body toward the device. The participant may also appear to be speaking directly at the device while also glancing for prolonged periods of time at the screen.

In contrast to the Impala, in which a small display in the center of the instrument cluster was actively used in addition to the center console display in the voice-based interactions, the Corolla infotainment DVI was similar to the CLA and the MKS in which visual-manual interaction during voice-involved tasks was limited to the press to speak button on the steering wheel and the center console display.

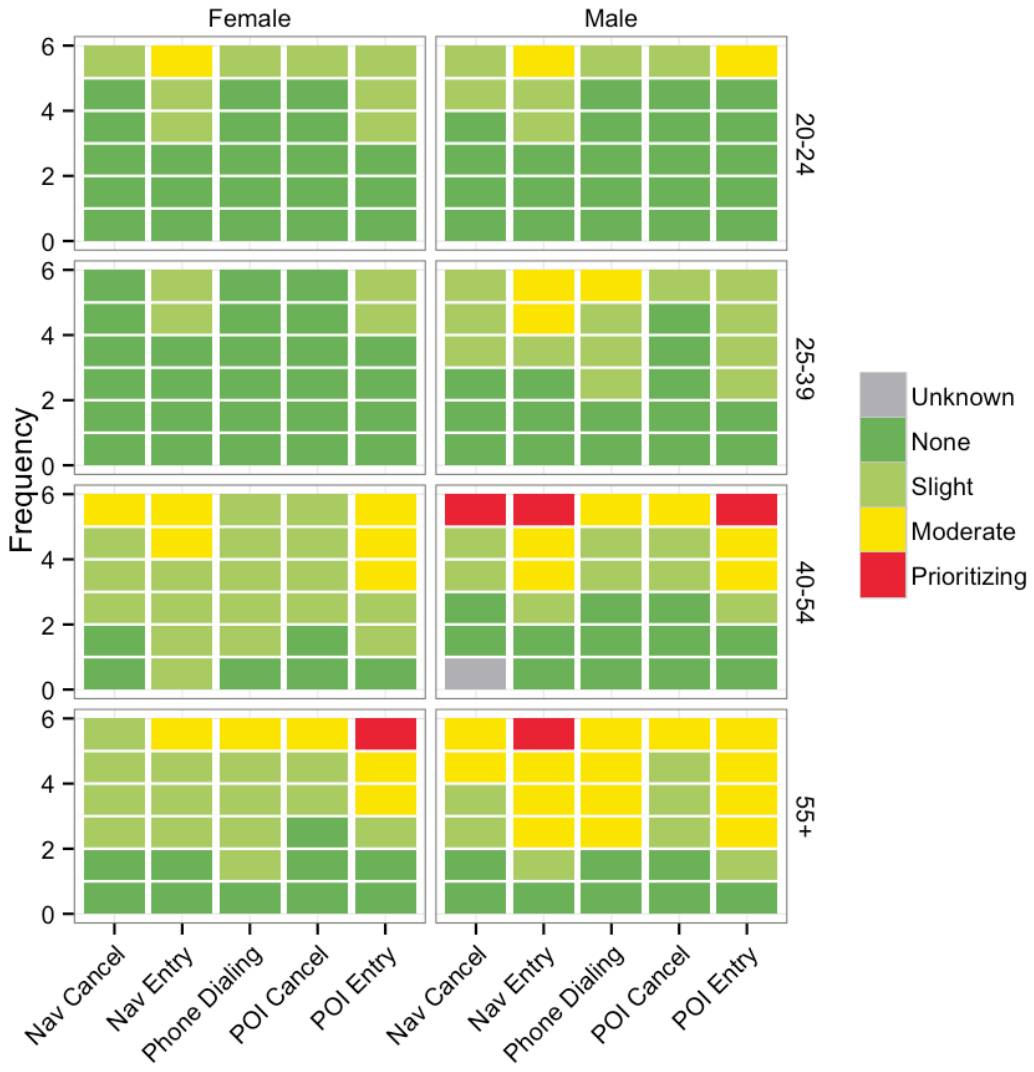

Figure 16: Orienting responses, divided across age groups and genders, as assessed by a single coder. This graphic visualizes each participant's greatest degree of orienting within a given task type.

As was the case in the second MKS sample, the Impala, and CLA, there was no significant effect of gender ($F(1, 40) = 1.05$, $p = 0.313$) on the presence of orienting behavior toward the center stack region. In contrast to previous studies, an effect of age is apparent, with orienting responses becoming more pronounced as age increases ($F(3, 40) = 6.69$, $p = .001$). These factors did not interact significantly ($F(3, 40) = 0.73$, $p = 0.052$), although some might interpret this as a trend. To the extent that a trend is present, it is in the opposite direction from what was seen in Phase I Study 1 in which older women tended to show more of an orienting response to the center console. The directionality in this data is for older males to show the stronger orienting response. The relatively small cell size per age and gender grouping (N=6) argues for a caution in interpreting such trends. The apparent age effect does, however, stand out more clearly as indicated by the statistical test values.

Driving Performance

Mean Velocity

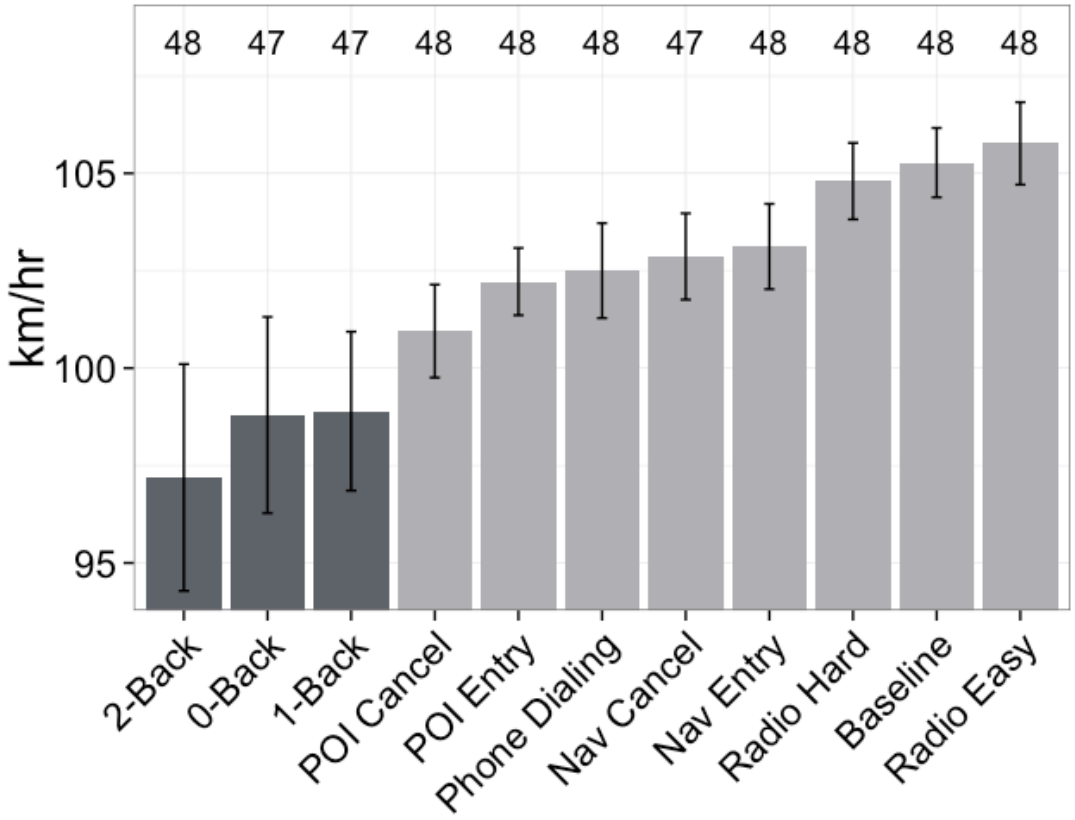

Figure 17: Mean vehicle velocity during all task periods under study. Labeling as in Figure 8.

Mean vehicle velocity was not significantly different across the various interface tasks ($X^2(6) = 9.3$, $p = 0.155$), the three levels of the n-back task ($X^2(2) = 1.1$, $p = 0.587$), or all tasks combined ($X^2(9) = 12.3$, $p = 0.199$). Mean velocity was nominally at its highest during the radio easy task and at its lowest during the 2-back task (105.8km/hr and 97.2km/hr, respectively; alternately 65.74mi/hr and 60.39mi/hr). N-back periods and interface periods were not significantly different in a direct comparison ($W = 1307$, $p = 0.259$).

Variability in Velocity

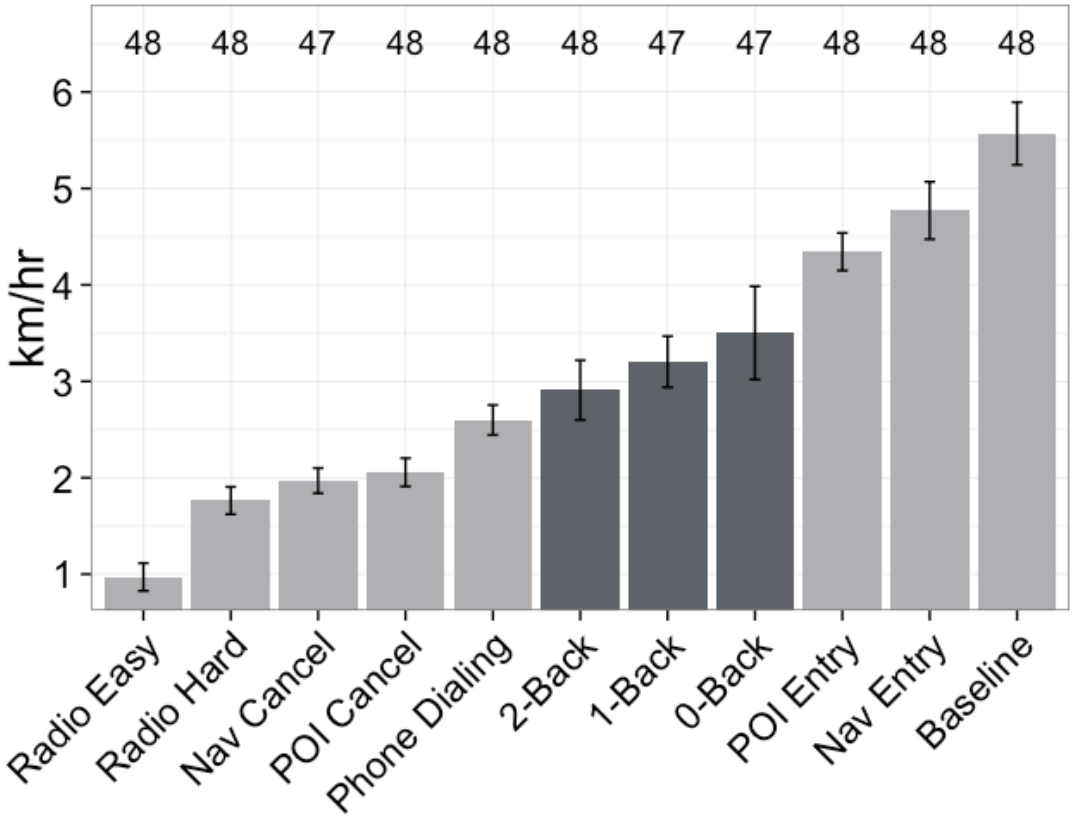

Figure 18: Standard deviation in velocity in all task periods under study. Labeling as in Figure 8.

Standard deviation values for longitudinal velocity decreased during each of the secondary tasks relative to baseline (single task) driving in this study in a pattern quite similar to that observed in the Impala and CLA. As in previous studies, Figure 18 shows that variability in velocity was found to at least nominally drop across all tasks relative to baseline driving. The smallest drops were for the relatively longer tasks – destination address entry in the navigation system and POI selection. For the Corolla, an overall significant main effect across baseline and all tasks ($X^2(9) = 199.4$, $p < 0.001$) appeared and a significant effect across baseline and the interface tasks ($X^2(6) = 186.7$, $p < 0.001$). There was no main effect across the n-back tasks ($X^2(2) = 0.8$, $p = 0.667$). As noted in our previous reports, there are known limitations in attempting to compare standard deviation values across tasks with significantly different time durations. It seems highly likely that in the present data, the variability of velocity variable may be more closely reflective of task time than providing a sensitive metric directly assessing task demand.

Steering Wheel Reversals

Steering wheel reversals were considered as a control metric and classified as proposed in the final report of the European Union AIDE project (deliverable D2.2.5, section 7.12) (Östlund et

al., 2005). (See also SAE (2015) standard document J2994 for additional discussion of this metric.) Major steering wheel reversals captures the number of steering wheel inputs exceeding an angular reversal gap of 3°. For minor steering wheel reversals, an angular reversal gap of 0.1° was used. The rate of steering wheel reversals per minute was obtained by dividing the raw reversal count by the task trial duration.

Major Steering Wheel Reversals

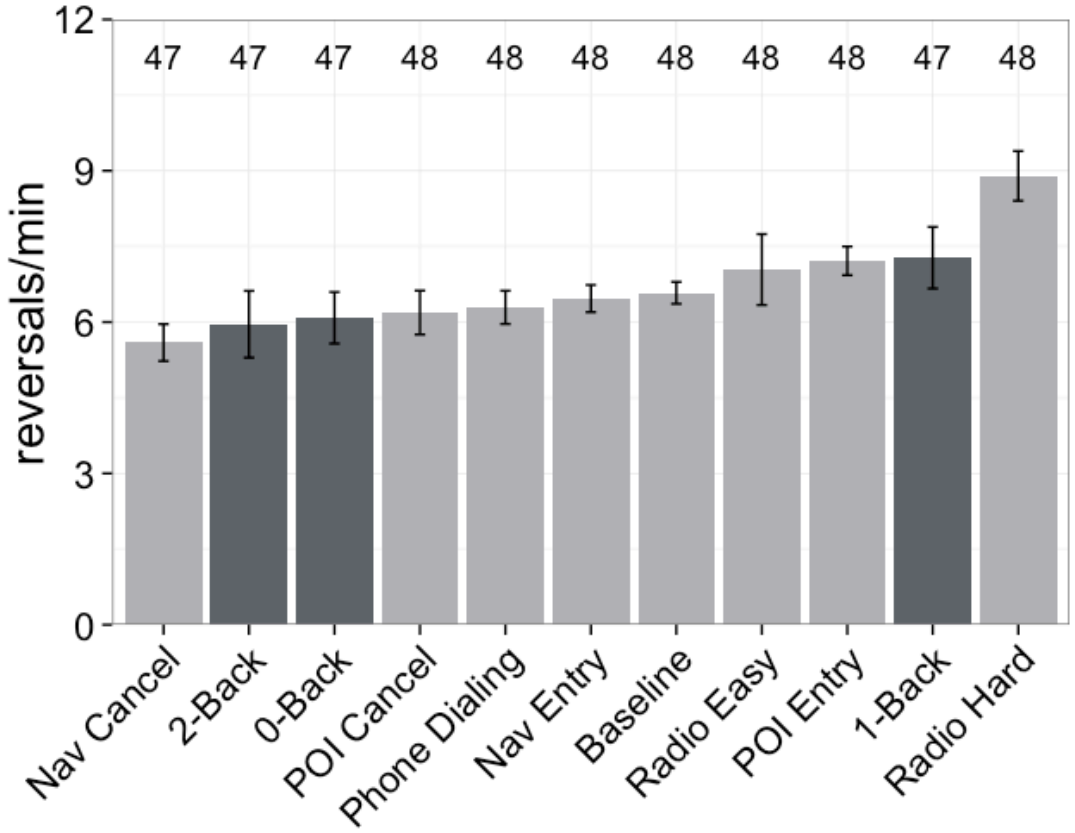

Figure 19: Major steering wheel reversal rates in all task periods under study. Labeling as in Figure 8.

Major steering wheel reversals differed significantly across all tasks ($X^2(9) = 29.2$, $p < 0.001$), an effect likely driven by differences between interface tasks ($X^2(6) = 33.1$, $p < 0.001$) rather than n-back tasks ($X^2(2) = 2.6$, $p = 0.271$). Major steering wheel reversal rates were lowest during the navigation cancel task (5.59 events/min), which is relatively brief. Major steering wheel reversal rates were highest during the manual radio (hard) tuning reference task (8.90 events/min). The finding that major steering wheel reversal rates were, on average, highest for the manual radio tuning reference task is in line with what was observed in the MKS, Impala, and CLA.

Minor Steering Wheel Reversals

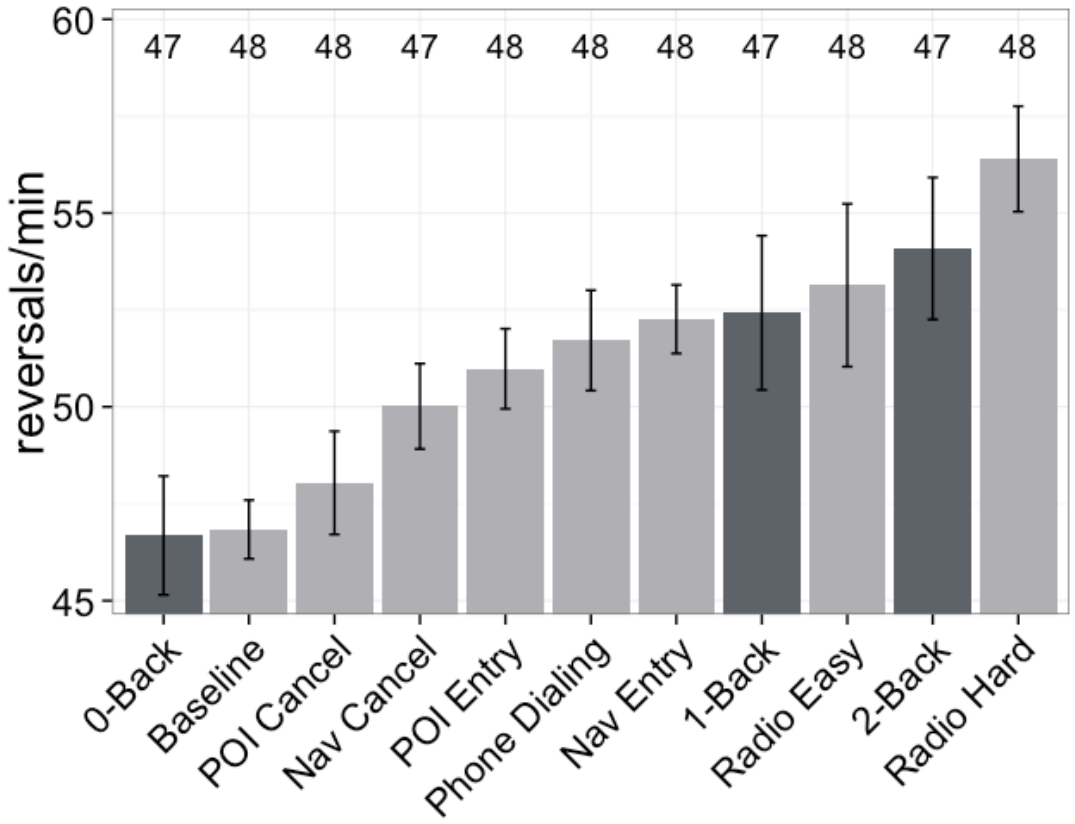

Figure 20: Minor steering wheel reversal rates in all task periods under study. Labeling as in Figure 8.

Minor steering wheel reversal rates differed significantly across all tasks ($X^2(9) = 24.8$, $p = 0.003$), considering just interface tasks alone ($X^2(6) = 16.1$, $p = 0.013$), and considering the n-back tasks independently ($X^2(2) = 7.5$, $p = 0.023$). Minor steering wheel reversal rates were lowest for the 0-back task, which were quite similar to baseline driving (46.68 and 46.84 reversals/min), as was generally the case in the MKS, Impala, and Corolla. In line with what was observed in the HASTE studies (Engström, Johansson, & Östlund, 2005), as a pure auditory-vocal surrogate task, the 2-back level of the n-back was associated with the next to the largest minor steering wheel reversal rate value. However, it can also be observed that manual radio tuning, the classic actual visual-manual task, as opposed to a surrogate visual-manual task), showed the largest value.

Task Performance / Error Analysis

The first part of this analysis considers for each individual task trial whether a trial was error free or if a system or user-based error occurred. An example of a user error is a participant giving an incorrect command during a voice-entry task, resulting in the task moving forward incorrectly or not moving forward at all. A representative system error is the system misinterpreting a voice

command that was in the correct form and understandable by human observers. Two evaluators independently coded each trial for errors (the research associate observing the participant during the drive and a second staff member who reviewed video and audio recordings of the interaction). A third member of the research staff mediated any discrepancies. For purposes of the binary classification of whether a user or system error occurred during a trial, the categorization followed was made that if a user error and system error occurred in the same trial, to code the trial as a user error regardless of the number of each type of error in the trial. Consequently, the rate of system errors may be somewhat underrepresented in this analysis.

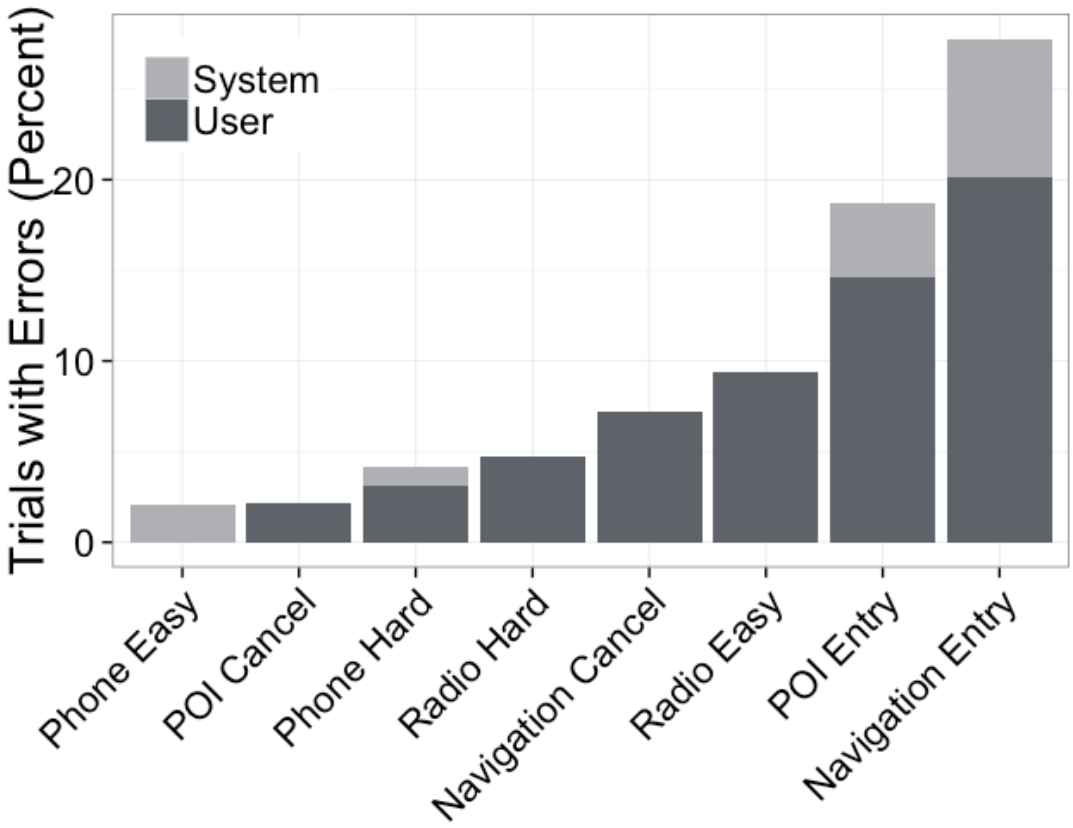

Figure 21: Error rates for each task across all trials and participants (i.e. percentage of trials in which either a user or system error, or both, occurred).

As shown in Figure 21, the Navigation Entry and POI Entry trials had much higher error rates (27.8% and 18.8%, respectively) than what was observed to contact phone calling trials (3.1%), paralleling what was observed in the Impala and CLA. Although this pattern was similar across vehicles, the total percentage of trials with errors was lower for the more involved tasks in the Corolla. The comparative percentage of trials with errors for Navigation Entry and POI Entry, respectively in the other vehicles were: Impala (24.0% and 31.8%), CLA (37.5% and 56.9%).

The next analysis is a more fine-grained characterization of the extent to which participants experienced difficulty completing a task. Individual trials were classified as: 1) completed without error or backtracking, 2) completed with backtracking, 3) completed with one instance

of the research associate providing a prompt to assist the participant, 4) completed with more than one prompt by the research associate, or 5) failure to complete the task. An example of "backtracking" is the situation where the system did not recognize or misinterpreted a street name, but the dialog allowed another opportunity for entry by asking for confirmation or indicating that it did not understand. In other words, a backtracking classification indicates that the system successfully supported error recovery (arising from either user error or system recognition error) and did not require the participant to begin the entire task again from the start. Backtracking could also occur when a participant recognized that they made an error (such as giving a wrong street name) and used an option provided by the system to correct the error. If the research associate judged that a participant was not progressing through a task on their own, one or more limited prompts was provided. The intent was to provide participants, as needed, with further assistance in learning how to use the system so that they might gain additional familiarity and potentially be more successful on subsequent trials. If a participant restarted a task more than twice or otherwise failed to progress in the interaction despite assistance, then they were guided through terminating the trial and moved-on. Failure to progress could be due to either user or system errors. Trials that failed to progress or were terminated due to either user or system errors were categorized as a failure.

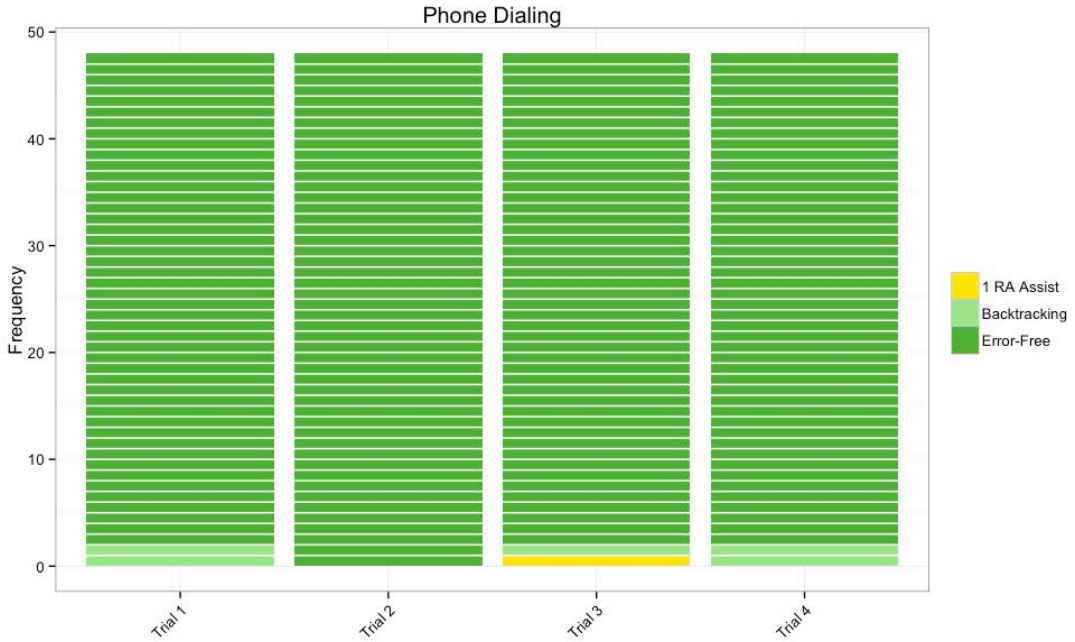

Figure 22: Graphical counts of error level for each trial of the contact phone calling task. Note that each column fully represents the 48 participants under consideration.

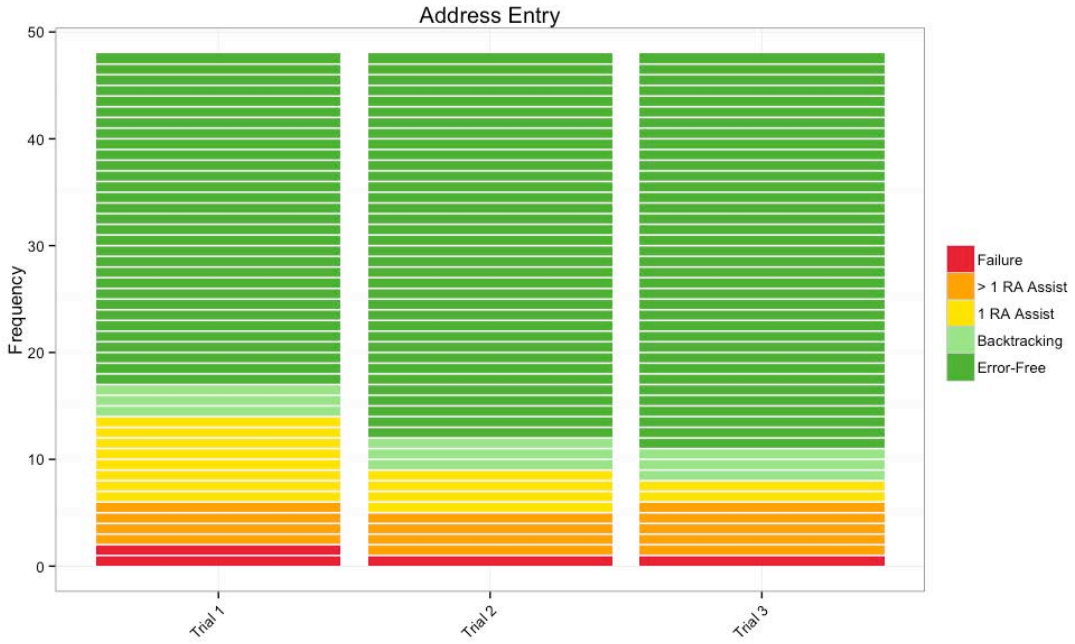

Figure 23: Graphical counts of error level for each trial of the address entry task. Note that each column fully represents the 48 participants under consideration.

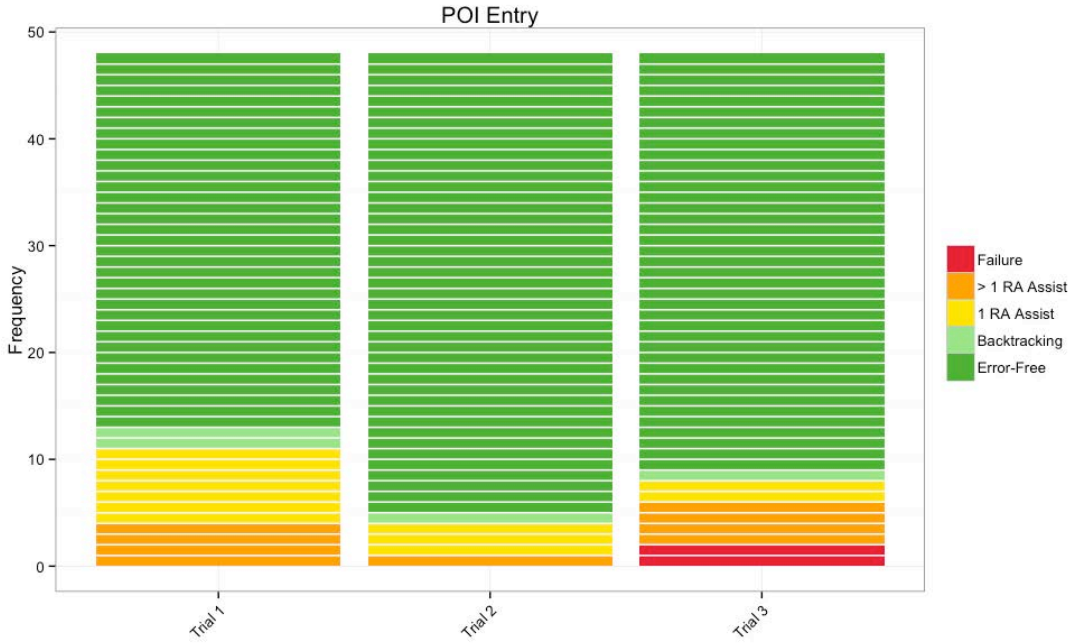

Figure 24: Graphical counts of error level for each trial of the address entry task.

Figure 22 (previous page) highlights overall error levels for the contact phone calling tasks. Figure 23 and Figure 24 show error levels for all participants across individual trials of the Navigation Entry and POI Entry tasks, respectively. While a formal statistical analysis of this data is difficult, a clear trend is observed for the Navigation Entry task, in that error rates generally decrease as trials progress. No particular trend is evident for the POI Entry task, or the

Phone Dialing task, which had few errors overall. This is in contrast to the pattern observed in the Impala, in which all three task types showed some evidence of decreasing error rates over time. Further, the CLA showed a clearer trend of decreasing errors over time for Phone Dialing.

Table 3 and Table 5 provide a numerical breakdown of error rates by level of assistance required and source of error. The "N/A" column indicates navigation system cancel trials that did not occur because a participant failed in the immediately preceding activity to enter a destination address or select a POI – hence, trials in which there was no navigation destination to cancel.

Table 4: Number of tasks/participants requiring a given level of assistance to complete a task.

	N/A	Failure	> 1 RA Assist	1 RA Assist	Back-tracking	Error-Free	TOTAL Trials
Phone Easy	0	0	0	0	2	94 (97.9%)	96
Phone Hard	0	0	0	1	3	92 (95.8%)	96
Nav Entry	0	4	13	14	9	104 (97.2%)	144
Nav Cancel	5	0	0	9	1	129 (89.6%)	144
POI Entry	0	2	9	12	4	117 (81.3%)	144
POI Cancel	5	0	2	1	0	136 (94.4%)	144
Radio Easy	0	1	2	4	2	87 (90.6%)	96
Radio Hard	11	0	1	3	0	81 (84.4%)	96
TOTAL	21	7	27	44	21	840 (87.5%)	960

Table 5: Task performance by trial tabulated by error type.

	Error Free	System Error	User Error	N/A	TOTAL Trials
Phone Easy	94 (97.9%)	2	0	0	96
Phone Hard	92 (95.8%)	1	3	0	96
Nav Entry	104 (97.2%)	11	29	0	144
Nav Cancel	129 (89.6%)	0	10	5	144
POI Entry	117 (81.3%)	6	21	0	144
POI Cancel	136 (94.4%)	0	3	5	144
Radio Easy	87 (90.6%)	0	9	0	96
Radio Hard	81 (84.4%)	0	4	11	96
TOTAL	840 (87.5%)	20	79	21	960

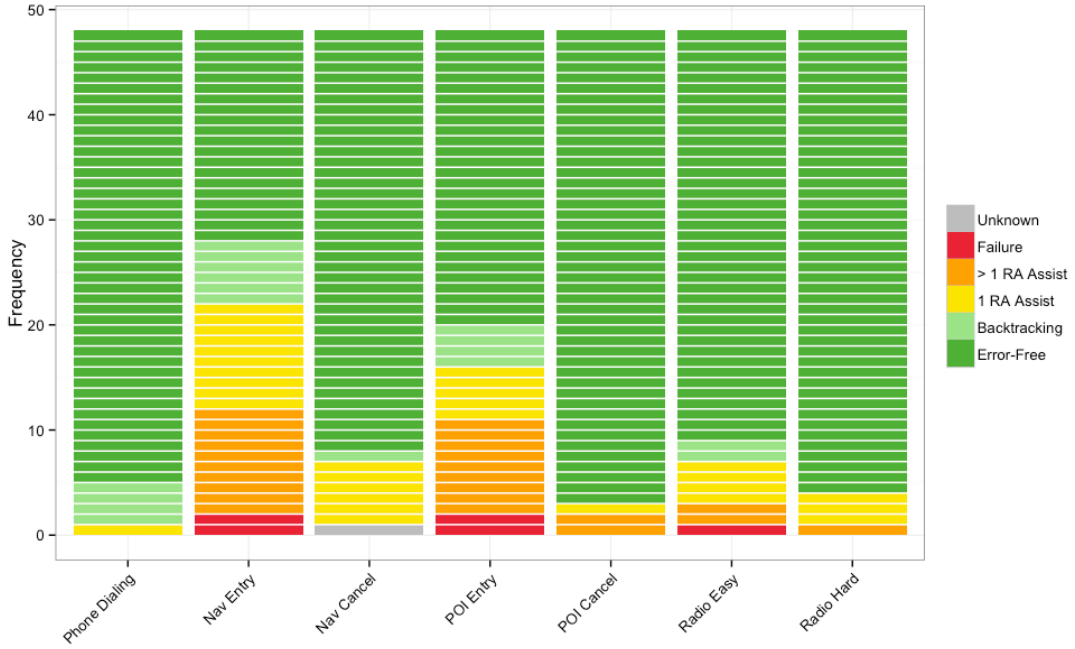

Figure 25: "Worse case" experience by task type at the individual participant level.

Across the 48 participants, color bars in Figure 25 indicate the highest level of difficulty experienced per participant during each task type. For example, the plot for radio hard indicates that 44 participants completed all trials error-free and 3 participants required 1 RA assist for at least one trial of interaction with the manual tuning of the radio interface, while 1 participant required more than 1 RA assist. The likelihood of experiencing some level of difficulty working with the interface while on-road was clearly much higher, during this relatively brief exposure, for the address entry into the navigation system and the POI entry tasks than for voice-involved phone contact calling.

Discussion

The combined results of Phase II Studies 1 through 3 (2014 Chevrolet Impala, 2014 Mercedes CLA, and 2015 Toyota Corolla) show a number of commonalities as well as differences across the voice-involved infotainment system implementations studied. All of the implementations show relatively low measures of subjective and objective demand associated with the specification of a contact to call from a saved contacts list. In contrast, the more content rich tasks of specifying a full address (city name, street name, street number) for entry into a navigation system or location of a unique point-of-interest (POI) such as a specific restaurant or museum, can be much more demanding and associated with a relatively high rate of errors in users who have been trained, but are relatively new to using a given system.

Phone Contact Calling

Using the voice-based phone contact calling interface in the Corolla was clearly an easily learned and executed task. After training in a parking lot, during on-road highway driving, out of 192 calling trials across the 48 participants in the analysis sample, 186 calls were completed on the first try, 5 were successfully completed with back-tracking by the participant, and only 1 required an assistive prompt on the part of the RA. Self-reported workload for placing calls using the interface was, on average, given a relatively low score of 1.5 on a 0 (low) to 10 (high scale), which was intermediate between the 0-back (0.7) and 1-back (2.1) auditory-vocal cognitive reference task. In contrast, the radio manual reference task was given a mean workload rating of 3.6 and the 2-back level of n-back task was rated at a 4.8. As another point of reference, selecting a radio station preset by visually locating and making a single touch screen "button press" was rated at a 2.8.

From an objective standpoint, in addition to having a low error rate, the task did not take very long to complete – 22.3 seconds on average. For comparison, the manual radio tuning task had a mean completion time of 17.9 seconds in the sample. Glance metrics easily met both Alliance (2006) and NHTSA (2013) guidelines for visual-manual tasks under the actual highway driving conditions considered. The mean TEORT for this task was 3.0 seconds. Available physiological data (heart rate and skin conductance level) showed, on average, modest reactivity. No dramatic shifts in driving behavior, relative to other tasks, stood out. Mean velocity was nominally lower than during the overall baseline (-3.2 kph). Major steering wheel reversal rates were lower than observed during the manual radio tuning reference task and very similar to what was observed during single task baseline driving. Minor steering wheel reversal rates increased relative to baseline, in-line with what would be expected of a largely auditory-vocal cognitive task as per wat was observed in the HASTE studies (Engström, Johansson, & Östlund, 2005); however, it can be noted that minor steering wheel reversal rates also increased during the manual radio tuning task.

Considering ease of use across system implementations, the error profile for phone contact calling in the Impala was only modestly higher than that seen in the Corolla. Out of the 192 calling trials in the Impala, 184 were completed on the first try, 3 required an assistive prompt on the part of the RA, and 2 were failures. In contrast, in the CLA, during the first calling trial, 12 out of 48 drivers in the analysis sample required some level of RA assistance to place a call and 3 additional drivers had to backtrack to successfully complete the call. As detailed in the CLA report, the sample as a whole did improve in placing calls with further exposure. The RA assistance rate dropped to 6, then 2, and 2, respectively over trials two through four. Self-reported workload ratings for phone contact calling across the vehicles was: Corolla (1.5), Impala (1.6), CLA (2.1). This would seem to indicate that while the method of placing calls in the CLA was not as immediately intuitive to this U.S. based sample of drivers, it could fairly readily be learned by most participants with some support.

POI & Address Entry Tasks

In contrast to voice-based phone contact calling, using currently available embedded in-vehicle voice systems for the more involved tasks of entering a full address (city name, street name, street number) or selecting a specific point-of-interest (POI), is clearly more demanding. In the Corolla implementation studied, self-reported workload for both of these task types was relatively similar and in the same general range as manual radio tuning reference task: address entry (3.3), POI (3.6), radio hard (3.6). Note that all three of these task types were rated as intermediate between the1-back (2.1) and 2-back (4.8) levels of the n-back auditory-vocal cognitive reference task. The reasons for the higher self-reported workload ratings for these tasks relative to voice-base phone contact calling are undoubtedly multi-dimensional. Both task types involved more interactive steps, took appreciably longer to complete (address entry 71.4s; POI 75.8s), required more visual engagement (TEORT) (address entry 12.4s; POI 15.0s), and the likelihood of experiencing a user or system error were substantially higher.

Comparing ease of use across system implementations, on average, the Corolla system was given nominally lower self-reported workload ratings. Mean address entry self-report workload ratings were: Corolla 3.3, CLA 3.4, Impala 3.9. The difference in ratings between the Corolla and CLA is modest, but is in-line with a lower overall error rate (backtracking, RA assist, plus failures) for the Corolla (27.8% vs. 36.1%) and lower TEORT (12.4s vs. 19s). The more notable differential for address entry is between the Corolla and the CLA vs. the Impala. The overall error profiles for address entry showed some differences across the vehicles, but were not dramatically different. Task completion time, on average, was similar in the Corolla and CLA (71.4s and 72.6s, respectively); mean address entry task completion time in the Impala was longer at 88.1 seconds. As we discuss in detail in the Impala technical report (Phase II Study 1), the average delay time between when a user finished issuing a voice command / verbal input string and the system responded, was significantly longer than in other systems we have evaluated (see also McWilliams, Reimer, Mehler, Dobres, & McAnulty, 2015). Whereas other systems tested showed mean delay times close to or below 2.5 seconds, mean delay times in the Impala averaged around 8 seconds and, in some instances, exceeded 15 seconds. This delay was quite noticeable and likely lead to a greater level of frustration with the basic function of the interface independent of its basic ability to process verbally input commands and information.

Another possible contributing factor to the higher reported demand in the Impala implementation relative to the Corolla and CLA was an overt mode switch in the latter part of the task. After the user input the destination address verbally, the system spoke the address and asked for a verbal confirmation that it was correct. The Impala DVI then displayed the address on the instrument cluster display and told the user to complete the selection from the “radio display”. This involved shifting attention from the forward orientation to the display in the instrument cluster to the center stack display, touching the address on the touch screen, and then touching the “GO” button on the touch screen. These additional visual-manual engagement steps seemed

"unnecessary" duplication as the user had already verbally confirmed the identification of the desired address. Undoubtedly, there were functional aspects related to the various subsystems employed in the overall DVI that necessitated these additional steps; however, it made for a less than optimal interaction for the user. This set-up also resulted in much greater visual demand than was seen in the Corolla and CLA. TEORT in the Impala address entry was, on average, 21.2 seconds. In net, both these implementation characteristics appear to overcome what might have offered some overall advantage by placing more of the multimodal interaction in the screen in the instrument cluster.

A somewhat different pattern across vehicles was seen for POI selection. Mean self-reported workload ratings for the POI task were: Corolla 3.6, CLA, 4.5, Impala 4.1. The differential between the Corolla and Impala was likely due to the same factors between them in address entry, i.e. the longer voice system response time in the Impala and the mode switch to a primary visual-manual task in a separate display area for completion of POI entry into the navigation system. In the CLA, however, POI entry appeared to be a more challenging task to learn than basic address entry. User backtracking, some level of RA assistance to complete the task, or task failure was present in over 50% of the POI trials in the CLA. While it is plausible that the POI interface offered a degree of flexibility and depth that experienced users might well find useful, the likelihood of initial frustration for novice users seems high. However, two qualifiers should be kept in mind in regard to the CLA. First, the complexity / flexibility of the menu structure in the CLA was largely independent of voice-interface characteristics, so the findings for this task relate to the voice interface per se in a limited manner. Second, the POIs used in the CLA evaluation were not the same as those used in the Corolla and the CLA, so it is possible that even though they were selected with the intention of providing the same relative level of selection demand as those used in the other two vehicles, there is the possibility that they were in some way fundamentally more challenging than the POIs used in the other two vehicles. While this possibility should be allowed for, the finding that all three POIs showed similarly elevated levels of difficulty for participants does suggest that there is something about the basic POI interface structure that was involved.

It appears that POI entry is generally a somewhat more demanding task in the DVIs tested than full address entry into the various navigations systems. It seems likely that this is due to the menu structure / search logic of the tasks rather than due to any intrinsic characteristics of the basic voice interface in each vehicle. Interestingly, the amount of RA prompting to aid participants through the POI tasks was relatively similar between the address entry and POI tasks in the Corolla and Impala. It was only in the CLA that the POI task logic seemed to be more difficult for a higher percentage of the users to pick-up than for address entry. Of the three vehicle interfaces, POI entry in the Corolla had fewer outright failures and required less RA assistance for the samples to work through the tasks.

While it can be very reasonably argued that voice-based interfaces may be less demanding that corresponding primarily visual-manual interfaces for accomplishing a given task, that does not

change the fact that the total level of demand of a specific type of voice-command involved interface needs to be taken into account. The totality of data collected in this project and related work would tend to suggest that using a voice interface to select a contact from a saved phone list is a relatively low demand task compared to other secondary task undertaken while driving, such as manual tuning of a radio. Other tasks, such entry of an address into a navigation system or selecting a POI, are likely less demanding if done through a voice-involved interface than through a primary visual-manual input system. Nonetheless, they are not unsubstantial in their broad level of demand, although generally falling in the same range as manual radio tuning across a number of factors. These tasks clearly need to be optimized to minimize overall demand as much as possible since current implementations certainly do not fully allow drivers to keep their hands on the steering wheel and eyes solely on the road. On the other hand, such interfaces do generally offer clear advantages in terms of visual and manual demand in contrast to "classical" primary visual-manual alternatives.

Additional Observations

Throughout this report, the manual radio tuning reference task (radio hard) is used as a comparison point in assessing the impact of tasks on various metrics. It is important to keep in mind that the task as used in this context is based on NHTSA's (2013) characterization of the task and explicitly involves two "button presses" (hard or touch screen depending on availability) and manual rotation of the fine tuning knob to a specified station (see Appendix E for additional detail). In a number of current production infotainment systems, the same goal state may frequently be obtainable with a single button press, followed by the fine tuning step. In the context of our studies, the task is designed to involve a comparable type and amount of demand across vehicles rather than assessing the typical amount of demand involved in fine tuning the radio in a particular manufacturer's DVI. Thus, steps involved in the radio tuning task employed in this study do not necessarily reflect the most likely manner in which most drivers would engage with the interface or by which the manufacture may have intended the radio to be used. In essence, a finding that the 87.5 percentile threshold for a testing sample for the radio hard task in these results is above NHTSA's (2013) recommended guideline of 12 seconds for TEORT, does not necessarily mean that the DVI as provided by the OEM does not meet the intent of the guideline. That being said, it is also interesting to note that as a manual reference task, the two button press and fine tuning knob adjustment task employed across the vehicles studied to date generally produces TEORT data either right at or just over the NHTSA recommended threshold if it were collected under the simulation testing conditions specified.

As was done for the Impala and CLA, a Monte Carlo analysis of the glance behavior considering the 24 subject sample size recommended in the NHTSA (2013) guidelines was run on the Corolla dataset. The simulations indicate that a sample of 24 drivers are highly likely to achieve results similar to those seen in the complete sample of 48. Consistent outcome agreement (100%) was obtained for all simulations involving voice-involved tasks. The only variation in outcomes

with the simulations was for radio tasks, specifically for percentage of glances greater than 2 seconds for the radio easy task and cumulative off-road glance time (TEORT) for both the radio easy and radio hard tasks. The latter value (62.65%) indicates that the task tends to generate data that hovers very close to the current recommended threshold value. This same pattern has appeared in the majority of the tests we have run on this task across vehicles, perhaps suggesting that this threshold level is a little low in representing the intended manual radio tuning reference threshold, at least when tested under actual on-road driving conditions.

Limitations

As we stated in previous reports, while the range of measures of demand and apparent workload collected in the present work was fairly extensive (self-report, task completion time, peripheral physiological arousal in the form of heart rate and skin conductance level, multiple glance behavior metrics, driving performance measures, and task performance), the assessment was not exhaustive in terms of necessarily capturing all aspects of attentional engagement. For example, while we have demonstrated that high working memory demands are associated with elevations in peripheral physiological arousal and can use the auditory-vocal n-back as a reference point for scaled levels of cognitive processing demand of this type, these measures do not necessarily capture other aspects of cognitive absorption in a task. Thus, this series of studies did not directly assess the extent to which potential low arousal associated "look but did not see" states might have been engendered by the voice-based interactions considered.

Conclusions

This work continues to highlight the highly multi-modal nature of voice-command interfaces as they continue to evolve in current production systems. As was documented in our initial Phase I study and throughout the range of vehicles considered in our Phase II and related work, including voice input in a user interface does not generally translate into elimination of demand for drivers to glance off the forward roadway during engagement with the task. For tasks such as address entry and POI selection, voice-command based engagement frequently involves a significant number of glances away from the forward roadway, although mean single glance duration and percentage of glances longer than 2 seconds are well within guidelines recommended by NHTSA (2013) for visual distraction if they were applied to these tasks. While there is visual engagement associated the voice-involved systems studied, TEORT time for a task such as phone contact calling can be relatively brief using a voice-involved interface and clearly offer advantages over primary visual-manual interface alternatives across a range of metrics (Mehler, Kidd, et al. 2015; Reimer, et al., 2015).

In the Toyota Corolla considered in this report, and as we have observed in other vehicles, frequency of successful task completion can be relatively high for easier voice-recognition tasks such as phone contact selection, but error rates continue to be elevated, at least for new users, for

more complex tasks such as address entry and POI selection. Thus, the demand associated with voice-involved interfaces needs to be evaluated taking into account both the type of task and the characteristics of a given implementation. While the relatively high level of frequency of errors encountered in this research in address entry and POI selection is far from optimal, across the vehicles studied to date, there is a range of demand observed, suggesting that design insight can be gained from studying these differences.

Acknowledgements

Acknowledgement is extended to the Toyota Collaborative Safety Research Center (CSRC) which provided the primary funding for this project. In addition, we are particularly grateful for the valuable, constructive comments provided by James Foley and Kazutoshi Ebe of CSRC during the development of the study.

Supplemental / matching support was provided United States Department of Transportation's Region I New England University Transportation Center at MIT.

This work would not have been possible without the support of AgeLab staff including: Alea Mehler, Hale McAnulty, Hillary Abraham, Dan Brown, Tom McWilliams, and Anthony Pettinato in study management, participant recruitment, the development and refinement of data analysis and extraction tools, data collection, and exhaustive reduction and coding of eye glance and other data.

The interpretive aspects of this report reflect the views of the authors, who are also responsible for the factualness and accuracy of the information presented herein.

References

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician, 35*(3), 124–129.

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F, 8(2),* 97-120.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32*(200), 675–701.

ISO 15007-1. (2002). Road vehicles - Measurement of driver visual behaviour with respect to transport information and control systems - Part 1: Definitions and parameters. Geneva, Switzerland: International Standards Organization.

ISO 15007-2. (2001). Road vehicles - Measurement of driver visual behaviour with respect to transport information and control systems - Part 2: Equipment and procedures. Geneva, Switzerland: International Standards Organization.

McWilliams, T., Reimer, B., Mehler, B., Dobres, J., & McAnulty, H. (2015). A secondary assessment of the impact of voice interface turn delays on driver attention and arousal in field conditions: a consideration of 4 vehicle systems and a smartphone. *Proceedings of the 8th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, Salt Lake City, Utah, June 22-25, 2015.

Mehler, B., Kidd, D., Reimer, B., Reagan, I., Dobres, J. & McCartt, A. (2015). Multi-modal assessment of on-road demand of voice and manual phone calling and voice navigation entry across two embedded vehicle systems. *Ergonomics.* DOI:10.1080/00140139.2015.1081412.

Mehler, B., McAnulty, H., & Reimer, B. (2015). *Assessing the Demands of Today's Voice Based In-Vehicle Interfaces (Phase II) - Project Plan 3: Study Design for Corolla HMI Assessment (Study Code 2015b).* MIT AgeLab Technical Report No. 2015-5A (May 22, 2015). Massachusetts Institute of Technology, Cambridge, MA.

Mehler, B., Reimer, B., Dobres, J., McAnulty, H., Mehler, A., Munger, D., & Coughlin, J.F. (2014). *Further evaluation of the effects of a production level "voice-command" interface on driver behavior: replication and a consideration of the significance of training method.* MIT AgeLab Technical Report No. 2014-2 (July 9, 2014). Massachusetts Institute of Technology, Cambridge, MA.

Mehler, B., Reimer, B., Dobres, J., McAnulty, H., & Coughlin, J.F. (2015). *Assessing the Demands of Voice Based In-Vehicle Interfaces: Phase II Experiment 1 - 2014 Chevrolet Impala (2014b).* MIT AgeLab Technical Report No. 2015-6. Massachusetts Institute of Technology, Cambridge, MA.

Mehler, B., Reimer, B. & Dusek, J.A. (2011). *MIT AgeLab delayed digit recall task (n-back).* MIT AgeLab White Paper Number 2011–3B. Massachusetts Institute of Technology, Cambridge, MA.

Mehler, B., Reimer, B., & McAnulty, H. (2014). *Assessing the Demands of Today's Voice Based In-Vehicle Interfaces (Phase II) - Project Plan 1: 2014 Chevrolet Impala (Study Code 2014b).* MIT AgeLab Technical Report No. 2014-16 (September 28, 2014). Massachusetts Institute of Technology, Cambridge, MA.

Mehler, B., Reimer, B., McAnulty, H., Dobres, J., Lee, J. & Coughlin, J.F. (2015). *Assessing the Demands of Voice Based In-Vehicle Interfaces - Phase II Experiment 2 - 2014 Mercedes CLA (2014t).* MIT AgeLab Technical Report 2015-8. Massachusetts Institute of Technology, Cambridge, MA.

National Highway Traffic Safety Administration. (2013). *(Issued Guidelines) Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices (Docket No. NHTSA-2010-0053).* Washington, DC: U.S. Department of Transportation National Highway Traffic Safety Administration (NHTSA).

Östlund, J., Peters, B., Thorslund, B., Engström, J., Markkula, G., Keinath, A., Horst, D., Juch, S., Mattes, S., & Foehl, U. (2005). Adaptive Integrated Driver-Vehicle Interface (AIDE): Driving performance assessment - methods and metrics. (Report No. IST-1-507674-IP). Information Society Technologies (IST) Programme, Gothenburg, Sweden.

R Core Team. (2013). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Reimer, B., Gruevski, P., & Coughlin, J.F. (2014). MIT AgeLab Video Annotator, Cambridge, MA. https://bitbucket.org/agelab/annotator

Reimer, B., & Mehler, B. (2011). The impact of cognitive workload on physiological arousal in young adult drivers: A field study and simulation validation. *Ergonomics, 54*(10), 932–942.

Reimer, B. & Mehler, B. (2013). *The effects of a production level "voice-command" interface on driver behavior: summary findings on reported workload, physiology, visual attention, and driving performance*. MIT AgeLab White Paper No. 2013-18A. Massachusetts Institute of Technology, Cambridge, MA.

Reimer, B., Mehler, B., Dobres, J. & Coughlin, J.F. (2013). *The effects of a production level "voice-command" interface on driver behavior: reported workload, physiology, visual attention, and driving performance.* MIT AgeLab Technical Report No. 2013-17A (November 18, 2012). Massachusetts Institute of Technology, Cambridge, MA.

Reimer, B., Mehler, B., Reagan, I, Kidd, D., & Dobres, J. (2015). Multi-modal demands of a smartphone used to place calls and enter addresses during highway driving relative to two embedded systems. Arlington, VA: Insurance Institute for Highway Safety.

SAE (2015). SAE International Surface Vehicle Recommended Practice, "Operational definitions of driving performance measures and statistics," SAE Standard J2944, June 2015.

Smith, D.L., Chang, J., Glassco, R., Foley, J., & Cohen, D. (2005). Methodology for capturing driver eye glance behavior during in-vehicle secondary tasks. *Transportation Research Record: Journal of the Transportation Research Board, 1937(1),* 61-65.

Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (1996). Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *European Heart Journal, 17*, 354-381.

Appendix A: Descriptive Statistics (Summary Tables)

Baseline Driving, Destination Address Entry & POI

Table 6: Means (and standard errors) for each Navigation Entry (destination address) and POI task, and variable of interest. Cells marked "N/A" indicate no applicable data for that task and variable (for example, there are no baseline periods for any of the "percentage change" metrics, as these are normalized against the baselines).

	Baseline	Navigation Entry	Navigation Cancel	POI Entry	POI Cancel
Self-Reported Workload	N/A	3.30 (0.35)	0.83 (0.19)	3.57 (0.31)	0.77 (0.12)
Task Completion Time	N/A	71.41 (3.45)	16.99 (0.64)	75.80 (3.20)	16.44 (0.52)
Mean Glance Time (to Device)	N/A	0.81 (0.04)	0.24 (0.04)	0.97 (0.03)	0.30 (0.04)
Mean Glance Time (Off-Road)	N/A	0.73 (0.02)	0.51 (0.03)	0.79 (0.02)	0.56 (0.02)
Glances Longer than 2.0s (to Device)	N/A	0.93 (0.37)	0.00 (0.00)	2.22 (0.75)	0.21 (0.18)
Glances Longer than 2.0s (Off-Road)	N/A	0.71 (0.30)	0.00 (0.00)	1.31 (0.41)	0.13 (0.10)
Total Glance Time (to Device)	N/A	8.32 (1.17)	0.81 (0.22)	10.46 (1.30)	0.88 (0.23)
Total Glance Time (Off-Road)	N/A	12.40 (1.28)	2.14 (0.28)	14.98 (1.51)	2.13 (0.26)
Number of Glances to Device	N/A	9.39 (1.34)	1.20 (0.30)	10.97 (1.31)	1.11 (0.24)
Number of Off-Road Glances	N/A	16.44 (1.59)	3.54 (0.41)	18.83 (1.76)	3.32 (0.32)
Heart Rate	75.58 (2.46)	76.44 (2.40)	76.24 (2.51)	76.15 (2.53)	75.40 (2.48)
Change in Heart Rate	N/A	1.48 (0.71)	0.30 (0.66)	0.82 (0.87)	-0.16 (0.75)
Skin Conductance Level	7.17 (0.90)	7.71 (1.10)	7.97 (1.12)	7.81 (1.00)	7.53 (0.94)
Change in SCL	N/A	1.49 (7.17)	3.85 (9.84)	8.36 (7.42)	4.76 (8.22)
Mean Velocity (mph)	65.41 (0.48)	64.08 (0.67)	63.92 (0.62)	63.52 (0.65)	62.73 (0.93)
Mean Velocity (kph)	105.27 (0.77)	103.12 (1.07)	102.87 (0.99)	102.22 (1.04)	100.95 (1.50)
Change in Mean Velocity	N/A	-1.97 (0.95)	-1.95 (0.89)	-2.73 (0.95)	-4.02 (1.48)
Velocity Range	21.78 (1.12)	17.91 (1.61)	7.24 (0.79)	15.87 (0.78)	6.93 (0.45)
Change in Velocity Range	N/A	-13.37 (6.59)	-65.01 (2.69)	-18.06 (5.45)	-64.81 (2.39)
SD of Velocity	5.57 (0.33)	4.77 (0.34)	1.97 (0.13)	4.34 (0.23)	2.06 (0.14)
Change in SD of Velocity	N/A	-3.67 (7.96)	-59.79 (3.33)	-9.39 (6.55)	-57.90 (3.07)
Major Steering Wheel Reversals	6.58 (0.28)	6.47 (0.34)	5.59 (0.49)	7.21 (0.35)	6.19 (0.51)
Change in Major SWRs	N/A	1.81 (4.78)	-11.77 (7.89)	13.90 (5.51)	0.99 (9.44)
Minor Steering Wheel Reversals	46.84 (1.20)	52.26 (1.41)	50.01 (1.75)	50.98 (1.63)	48.04 (1.85)
Change in Minor SWRs	N/A	13.14 (2.01)	6.15 (2.26)	9.61 (2.47)	3.62 (3.39)

Units: All change scores represent % change from baseline (single task) driving and were calculated individually by subject and the resulting values averaged. Heart rate is in beats per minute, skin conductance level (SCL) is in microsiemens, velocity metrics are in kph unless otherwise indicated, and steering wheel reversals (SWRs) are normalized as number per minute.

Contact Phone Calling & Manual Radio Tasks

Table 7: Means (and standard errors) for phone contact calling and radio tasks, and variable of interest. Cells marked "N/A" indicate no applicable data for that task and variable (for example, there are no baseline periods for any of the "percentage change" metrics, as these are normalized against the baselines).

	Contact Calling	Radio Easy	Radio Hard
Self-Reported Workload	1.48 (0.18)	2.77 (0.37)	3.59 (0.41)
Task Completion Time	22.34 (0.72)	7.91 (0.82)	17.89 (0.97)
Mean Glance Time (to Device)	0.50 (0.04)	1.07 (0.06)	0.96 (0.03)
Mean Glance Time (Off-Road)	0.61 (0.03)	1.01 (0.05)	0.94 (0.03)
Glances Longer than 2.0s (to Device)	1.30 (0.75)	6.56 (2.13)	2.11 (0.72)
Glances Longer than 2.0s (Off-Road)	0.88 (0.56)	5.95 (2.02)	2.11 (0.72)
Total Glance Time (to Device)	1.63 (0.25)	4.09 (0.55)	9.72 (0.65)
Total Glance Time (Off-Road)	3.01 (0.30)	4.36 (0.57)	9.88 (0.66)
Number of Glances to Device	1.98 (0.27)	3.74 (0.42)	10.04 (0.49)
Number of Off-Road Glances	4.52 (0.39)	4.27 (0.45)	10.42 (0.51)
Heart Rate	76.81 (2.38)	73.69 (2.45)	74.55 (2.51)
Change in Heart Rate	2.25 (0.68)	-2.76 (0.79)	-1.59 (0.71)
Skin Conductance Level	7.45 (1.13)	7.32 (1.12)	7.62 (1.23)
Change in SCL	-0.77 (6.98)	0.48 (8.94)	3.39 (9.77)
Mean Velocity (mph)	63.69 (0.79)	65.72 (0.60)	65.12 (0.54)
Mean Velocity (kph)	102.50 (1.27)	105.77 (0.96)	104.80 (0.86)
Change in Mean Velocity	-2.29 (1.11)	0.89 (0.91)	-0.13 (0.81)
Velocity Range	8.68 (0.51)	3.19 (0.30)	5.99 (0.44)
Change in Velocity Range	-55.76 (3.04)	-83.37 (1.79)	-69.86 (2.68)
SD of Velocity	2.60 (0.16)	0.97 (0.08)	1.76 (0.13)
Change in SD of Velocity	-47.28 (3.59)	-80.05 (2.04)	-64.70 (3.17)
Major Steering Wheel Reversals	6.29 (0.41)	7.04 (0.77)	8.90 (0.63)
Change in Major SWRs	0.57 (7.20)	9.60 (13.47)	34.74 (8.07)
Minor Steering Wheel Reversals	51.71 (1.91)	53.14 (2.75)	56.39 (1.88)
Change in Minor SWRs	10.44 (3.04)	14.90 (5.58)	21.93 (3.50)

Units: All change scores represent % change from baseline (single task) driving and were calculated individually by subject and the resulting values averaged. Heart rate is in beats per minute, skin conductance level (SCL) is in microsiemens, velocity metrics are in kph unless otherwise indicated, and steering wheel reversals (SWRs) are normalized as number per minute.

Baseline Driving & N-Back Tasks

Table 8: Means (and standard errors) for baseline and each level of the n-back task for each variable of interest. Cells marked "N/A" indicate no applicable data for that task and variable (for example, there are no baseline periods for any of the "percentage change" metrics, as these are normalized against the baselines).

	Baseline	0-Back	1-Back	2-Back
Self-Reported Workload	N/A	0.69 (0.16)	2.12 (0.24)	4.78 (0.42)
Heart Rate	75.58 (2.46)	75.41 (2.57)	78.07 (2.55)	81.52 (2.47)
Change in Heart Rate	N/A	0.69 (0.91)	4.27 (1.29)	8.60 (1.60)
Skin Conductance Level	7.17 (0.90)	7.82 (0.99)	8.39 (0.91)	8.47 (1.11)
Change in SCL	N/A	15.54 (9.92)	25.78 (14.91)	26.67 (15.90)
Mean Velocity (mph)	65.41 (0.48)	61.39 (1.92)	61.45 (1.62)	60.39 (2.17)
Mean Velocity (kph)	105.27 (0.77)	98.80 (3.08)	98.90 (2.60)	97.19 (3.49)
Change in Mean Velocity	N/A	-5.93 (3.03)	-5.57 (2.52)	-5.69 (2.69)
Velocity Range	21.78 (1.12)	11.42 (1.48)	10.82 (0.90)	10.11 (0.98)
Change in Velocity Range	N/A	-40.01 (7.89)	-45.41 (5.20)	-45.74 (6.46)
SD of Velocity	5.57 (0.33)	3.50 (0.54)	3.20 (0.31)	2.91 (0.35)
Change in SD of Velocity	N/A	-25.23 (11.80)	-33.77 (7.44)	-35.63 (9.32)
Major Steering Wheel Reversals	6.58 (0.28)	6.09 (0.61)	7.28 (0.68)	5.96 (0.67)
Change in Major SWRs	N/A	-1.84 (10.19)	13.88 (11.53)	0.94 (12.28)
Minor Steering Wheel Reversals	46.84 (1.20)	46.68 (2.30)	52.43 (2.66)	54.09 (2.63)
Change in Minor SWRs	N/A	-0.04 (4.54)	11.95 (5.47)	16.30 (5.26)

Units: All change scores represent % change from baseline (single task) driving and were calculated individually by subject and the resulting values averaged. Heart rate is in beats per minute, skin conductance level (SCL) is in microsiemens, velocity metrics are in kph unless otherwise indicated, and steering wheel reversals (SWRs) are normalized as number per minute.

Appendix B: Results Breakdown by Trial

Destination Address Entry by Trial

Table 9: Means (and standard errors) for each trial, in sequence, for the destination address entry task (Nav Entry).

	Trial 1	Trial 2	Trial 3
Task Completion Time	71.55 (5.46)	72.53 (4.74)	70.15 (4.74)
Mean Glance Time (to Device)	0.76 (0.05)	0.90 (0.04)	0.78 (0.05)
Mean Glance Time (Off-Road)	0.70 (0.03)	0.75 (0.03)	0.74 (0.02)
Glances Longer than 2.0s (to Device)	0.92 (0.55)	1.49 (0.84)	0.38 (0.23)
Glances Longer than 2.0s (Off-road)	0.58 (0.33)	1.16 (0.70)	0.39 (0.20)
Total Glance Time (to Device)	8.31 (1.55)	8.36 (1.36)	8.29 (1.29)
Total Glance Time (Off-Road)	12.10 (1.75)	12.60 (1.53)	12.50 (1.43)
Number of Glances to Device	9.62 (1.78)	9.00 (1.48)	9.54 (1.49)
Number of Off-Road Glances	16.25 (2.12)	16.44 (1.85)	16.62 (1.82)
Heart Rate	77.35 (2.36)	76.63 (2.42)	75.34 (2.47)
Change in Heart Rate	2.65 (0.95)	1.58 (0.79)	-0.30 (0.48)
Skin Conductance Level	7.47 (1.21)	7.64 (1.12)	8.03 (1.05)
Change in SCL	0.40 (8.66)	5.07 (9.54)	11.87 (8.83)
Mean Velocity (mph)	64.55 (0.81)	61.90 (1.13)	65.78 (0.52)
Mean Velocity (kph)	103.89 (1.30)	99.61 (1.82)	105.86 (0.83)
Change in Mean Velocity	-1.24 (1.16)	-5.38 (1.60)	0.68 (0.77)
Velocity Range	18.94 (2.64)	21.27 (2.67)	13.53 (0.94)
Change in Velocity Range	-13.83 (7.94)	4.96 (12.25)	-32.66 (5.23)
SD of Velocity	4.85 (0.54)	5.68 (0.70)	3.78 (0.29)
Change in SD of Velocity	-7.77 (8.71)	18.82 (16.64)	-25.87 (6.02)
Major Steering Wheel Reversals	6.09 (0.45)	7.26 (0.52)	6.05 (0.47)
Change in Major SWRs	-3.13 (7.77)	14.37 (8.17)	-9.05 (6.33)
Minor Steering Wheel Reversals	52.20 (1.72)	53.77 (1.70)	50.81 (1.58)
Change in Minor SWRs	11.94 (3.27)	15.58 (2.98)	9.12 (2.41)

Units: All change scores represent % change from baseline (single task) driving and were calculated individually by subject and the resulting values averaged. Heart rate is in beats per minute, skin conductance level (SCL) is in microsiemens, velocity metrics are in kph unless otherwise indicated, and steering wheel reversals (SWRs) are normalized as number per minute.

POI Entry by Trial

Table 10: Means (and standard errors) for each trial, in sequence, for the POI entry task.

	Trial 1	Trial 2	Trial 3
Task Completion Time	74.59 (4.42)	66.11 (3.13)	86.71 (5.25)
Mean Glance Time (to Device)	0.98 (0.04)	0.96 (0.03)	0.99 (0.04)
Mean Glance Time (Off-Road)	0.79 (0.03)	0.75 (0.02)	0.83 (0.02)
Glances Longer than 2.0s (to Device)	2.94 (1.30)	1.39 (0.81)	2.31 (0.94)
Glances Longer than 2.0s (Off-road)	1.86 (0.65)	0.67 (0.34)	1.40 (0.58)
Total Glance Time (to Device)	10.27 (2.07)	6.64 (1.06)	14.48 (1.57)
Total Glance Time (Off-Road)	14.61 (2.22)	10.76 (1.19)	19.57 (2.02)
Number of Glances to Device	11.10 (2.13)	7.04 (1.06)	14.75 (1.57)
Number of Off-Road Glances	18.29 (2.45)	14.46 (1.44)	23.75 (2.41)
Heart Rate	76.48 (2.51)	76.02 (2.56)	75.96 (2.60)
Change in Heart Rate	1.30 (0.87)	0.68 (1.01)	0.50 (0.93)
Skin Conductance Level	7.99 (1.01)	7.87 (0.95)	7.58 (1.10)
Change in SCL	14.30 (8.43)	12.55 (6.91)	6.14 (7.92)
Mean Velocity (mph)	64.79 (0.73)	63.73 (0.60)	62.03 (1.42)
Mean Velocity (kph)	104.27 (1.18)	102.56 (0.96)	99.83 (2.28)
Change in Mean Velocity	-0.86 (1.07)	-2.48 (0.86)	-5.10 (2.10)
Velocity Range	14.28 (1.44)	14.94 (1.17)	18.39 (1.65)
Change in Velocity Range	-26.65 (8.44)	-28.56 (5.00)	-2.99 (11.22)
SD of Velocity	4.04 (0.51)	4.03 (0.34)	4.96 (0.50)
Change in SD of Velocity	-14.85 (12.42)	-22.30 (6.85)	3.51 (12.82)
Major Steering Wheel Reversals	7.12 (0.45)	7.21 (0.47)	7.30 (0.57)
Change in Major SWRs	11.38 (7.39)	15.29 (7.77)	12.81 (9.00)
Minor Steering Wheel Reversals	50.05 (1.85)	51.99 (2.08)	50.90 (1.71)
Change in Minor SWRs	7.31 (3.22)	11.27 (3.39)	8.85 (2.79)

Units: All change scores represent % change from baseline (single task) driving and were calculated individually by subject and the resulting values averaged. Heart rate is in beats per minute, skin conductance level (SCL) is in microsiemens, velocity metrics are in kph unless otherwise indicated, and steering wheel reversals (SWRs) are normalized as number per minute.

Contact Phone Calling by Trial

Table 11: Means (and standard errors) for each trial, in sequence, for the contact phone calling task.

	Trial 1 "Easy 1"	Trial 2 "Easy 2"	Trial 3 "Hard 1"	Trial 4 "Hard 2"
Task Completion Time	23.69 (1.44)	19.80 (0.71)	24.32 (1.46)	21.54 (1.15)
Mean Glance Time (to Device)	0.67 (0.07)	0.49 (0.09)	0.48 (0.06)	0.35 (0.06)
Mean Glance Time (Off-Road)	0.66 (0.04)	0.61 (0.07)	0.59 (0.03)	0.57 (0.04)
Glances Longer than 2.0s (to Device)	3.12 (2.20)	2.08 (2.08)	0.00 (0.00)	0.00 (0.00)
Glances Longer than 2.0s (Off-Road)	1.45 (0.88)	2.08 (2.08)	0.00 (0.00)	0.00 (0.00)
Total Glance Time (to Device)	2.68 (0.44)	1.24 (0.22)	1.66 (0.37)	0.92 (0.30)
Total Glance Time (Off-Road)	3.91 (0.47)	2.52 (0.30)	3.27 (0.41)	2.34 (0.37)
Number of Glances to Device	3.04 (0.43)	1.71 (0.29)	2.04 (0.41)	1.12 (0.28)
Number of Off-Road Glances	5.38 (0.54)	4.08 (0.46)	4.94 (0.49)	3.69 (0.50)
Heart Rate	77.01 (2.48)	76.72 (2.33)	77.18 (2.39)	76.26 (2.51)
Change in Heart Rate	2.06 (0.95)	1.82 (0.70)	2.39 (0.82)	1.24 (0.97)
Skin Conductance Level	7.65 (1.14)	7.46 (1.13)	7.44 (1.10)	7.26 (1.19)
Change in SCL	3.88 (7.17)	0.76 (6.35)	1.53 (7.22)	-4.04 (6.49)
Mean Velocity (mph)	65.64 (0.80)	62.66 (1.38)	62.43 (1.48)	64.03 (0.85)
Mean Velocity (kph)	105.64 (1.29)	100.84 (2.22)	100.47 (2.38)	103.05 (1.36)
Change in Mean Velocity	0.46 (1.16)	-4.09 (2.15)	-4.48 (2.23)	-2.04 (1.19)
Velocity Range	8.15 (0.79)	9.67 (1.24)	9.86 (0.99)	7.03 (0.56)
Change in Velocity Range	-59.43 (3.81)	-51.16 (6.01)	-48.25 (5.93)	-65.13 (2.93)
SD of Velocity	2.45 (0.25)	2.95 (0.42)	2.92 (0.31)	2.08 (0.18)
Change in SD of Velocity	-50.33 (5.50)	-41.94 (7.78)	-40.15 (6.86)	-58.69 (3.73)
Major Steering Wheel Reversals	6.64 (0.72)	6.44 (0.70)	6.91 (0.64)	5.18 (0.61)
Change in Major SWRs	0.57 (11.13)	5.11 (13.28)	9.48 (10.80)	-20.11 (9.99)
Minor Steering Wheel Reversals	52.03 (2.39)	51.75 (2.32)	54.16 (2.78)	48.91 (2.52)
Change in Minor SWRs	11.54 (4.48)	11.04 (4.42)	15.27 (5.13)	3.26 (4.42)

Units: All change scores represent % change from baseline (single task) driving and were calculated individually by subject and the resulting values averaged. Heart rate is in beats per minute, skin conductance level (SCL) is in microsiemens, velocity metrics are in kph unless otherwise indicated, and steering wheel reversals (SWRs) are normalized as number per minute.

Appendix C: Address Entry & POI Data for Comparison with other Samples

Collapsed Values for Trials 1 & 2 (for Comparison to MKS Data)

The studies in Phase I (carried out in the MKS) presented participants with two address to enter into the navigation system, while Study 1 in Phase II presented four addresses. The table below collapses the data for the first two address entry trials (Nav Entry) in the present study to allow direct comparison of findings across studies. The "first" two addresses are the same across studies.

The first two POI entries for the current study are also provided for comparative purposes.

Table 12: Means (and standard errors) collapsed across trials 1 & 2.

	Nav Entry	Nav Cancel	POI Entry	POI Cancel
Self-Reported Workload	3.30 (0.35)	0.83 (0.19)	3.57 (0.31)	0.77 (0.12)
Task Completion Time	72.04 (3.77)	17.47 (0.74)	70.35 (3.12)	16.72 (0.75)
Mean Glance Time (to Device)	0.83 (0.04)	0.26 (0.04)	0.97 (0.03)	0.36 (0.05)
Mean Glance Time (Off-Road)	0.73 (0.02)	0.51 (0.03)	0.77 (0.02)	0.59 (0.02)
Glances Longer than 2.0s (to Device)	1.20 (0.53)	0.00 (0.00)	2.17 (0.88)	0.32 (0.26)
Glances Longer than 2.0s (Off-Road)	0.87 (0.42)	0.00 (0.00)	1.27 (0.43)	0.20 (0.16)
Total Glance Time (to Device)	8.33 (1.24)	0.96 (0.27)	8.45 (1.45)	1.15 (0.33)
Total Glance Time (Off-Road)	12.35 (1.38)	2.36 (0.34)	12.69 (1.57)	2.39 (0.36)
Number of Glances to Device	9.31 (1.41)	1.40 (0.36)	9.07 (1.46)	1.38 (0.32)
Number of Off-Road Glances	16.34 (1.69)	3.90 (0.49)	16.38 (1.76)	3.56 (0.39)
Heart Rate	76.99 (2.38)	76.16 (2.52)	76.25 (2.51)	75.37 (2.48)
Change in Heart Rate	1.81 (0.54)	0.26 (0.49)	0.91 (0.60)	-0.19 (0.49)
Skin Conductance Level	7.56 (1.13)	8.07 (1.14)	7.93 (0.98)	7.54 (0.89)
Change in SCL	2.14 (5.17)	4.71 (7.34)	11.01 (5.12)	6.70 (5.62)
Mean Velocity (mph)	63.23 (0.82)	63.26 (0.84)	64.26 (0.55)	62.95 (0.85)
Mean Velocity (kph)	101.75 (1.32)	101.81 (1.36)	103.42 (0.89)	101.31 (1.37)
Change in Mean Velocity	-2.65 (0.74)	-2.48 (0.79)	-2.19 (0.61)	-3.85 (0.96)
Velocity Range	20.10 (2.32)	8.17 (1.11)	14.61 (0.88)	6.81 (0.56)
Change in Velocity Range	-8.81 (5.29)	-63.17 (2.30)	-22.94 (3.65)	-65.17 (1.90)
SD of Velocity	5.27 (0.45)	2.15 (0.17)	4.03 (0.28)	2.03 (0.17)
Change in SD of Velocity	1.03 (6.37)	-58.19 (2.71)	-14.08 (4.74)	-58.31 (2.41)
Major Steering Wheel Reversals	6.67 (0.36)	5.71 (0.59)	7.17 (0.39)	6.06 (0.54)
Change in Major SWRs	3.76 (3.76)	-9.84 (6.43)	13.61 (4.21)	-0.63 (6.84)
Minor Steering Wheel Reversals	52.99 (1.53)	50.35 (1.93)	51.02 (1.82)	47.92 (2.07)
Change in Minor SWRs	13.46 (1.63)	6.75 (1.97)	9.45 (1.89)	3.05 (2.57)

Units: All change scores represent % change from baseline (single task) driving and were calculated individually by subject and the resulting values averaged. Heart rate is in beats per minute, skin conductance level (SCL) is in microsiemens, velocity metrics are in kph unless otherwise indicated, and steering wheel reversals (SWRs) are normalized as number per minute.

Appendix D: Selected Graphs in Alternate Formats

Heart Rate in BPM

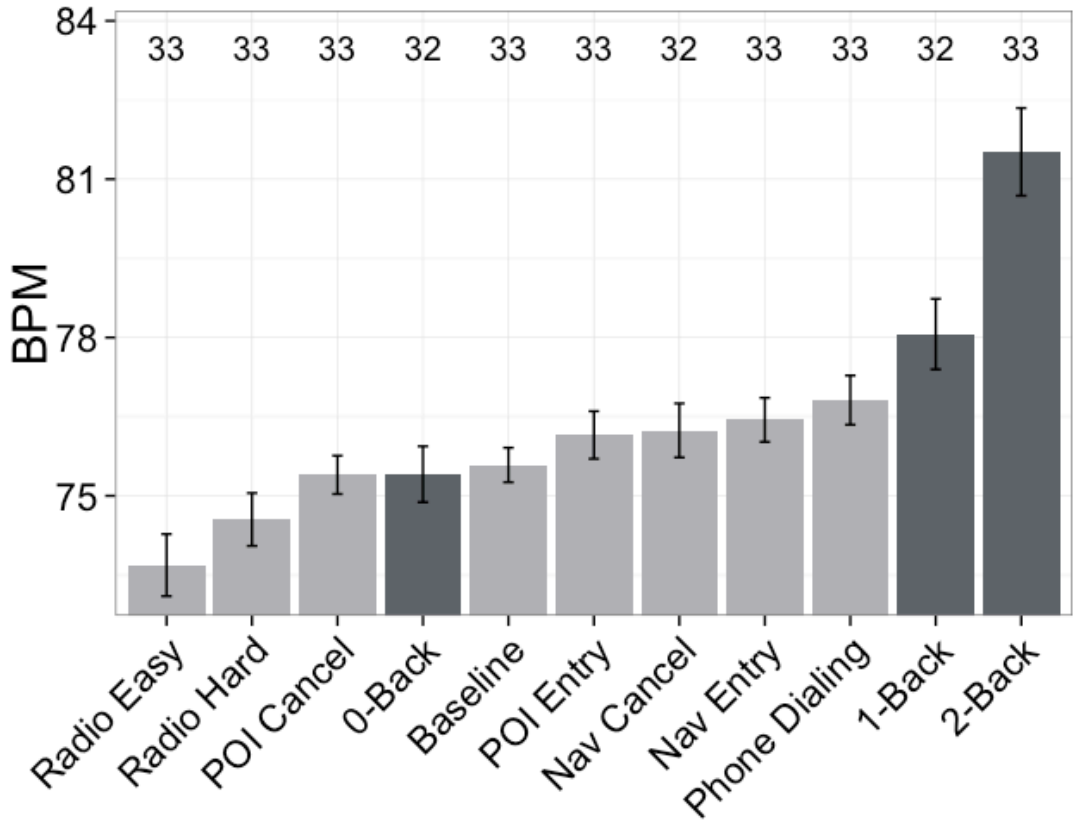

Figure 26: Mean heart rate for sample in beats per minute (BPM) for each task under study. Labeling as in Figure 8.

Skin Conductance in Absolute Units

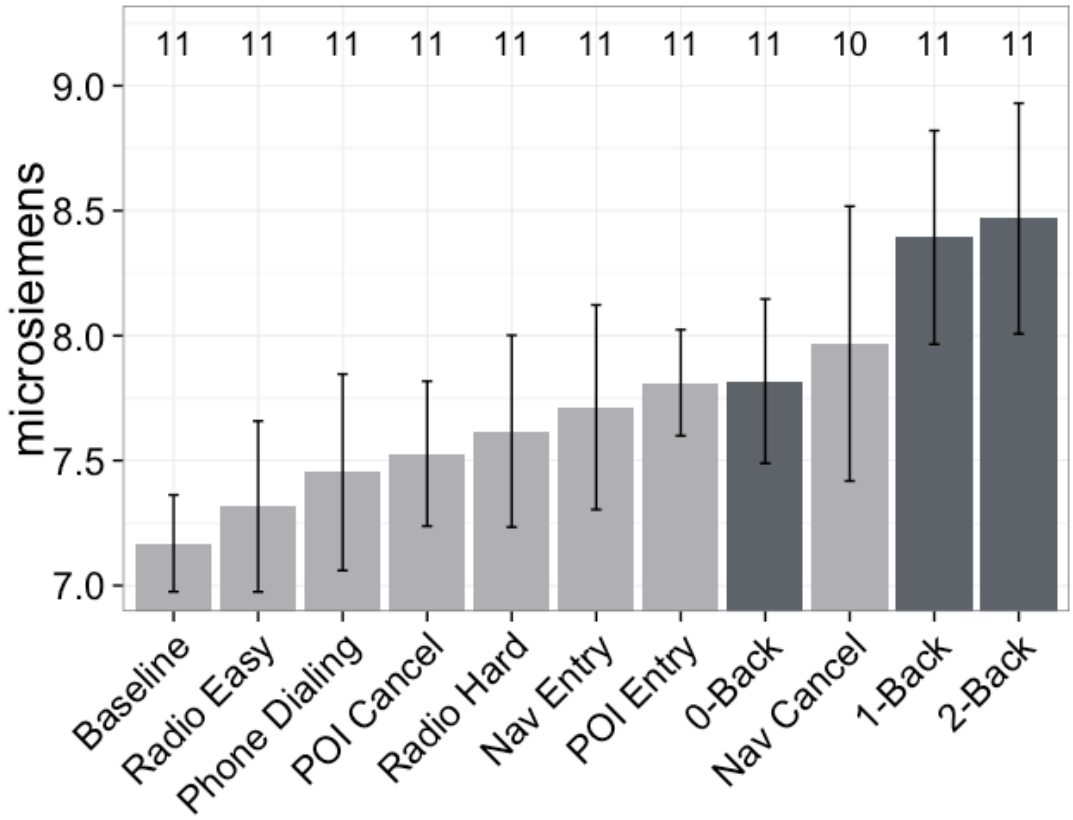

Figure 27: Mean skin conductance level for sample in microsiemens (micromhos) for each task under study. Labeling as in Figure 8.

Total Glance Time to Device

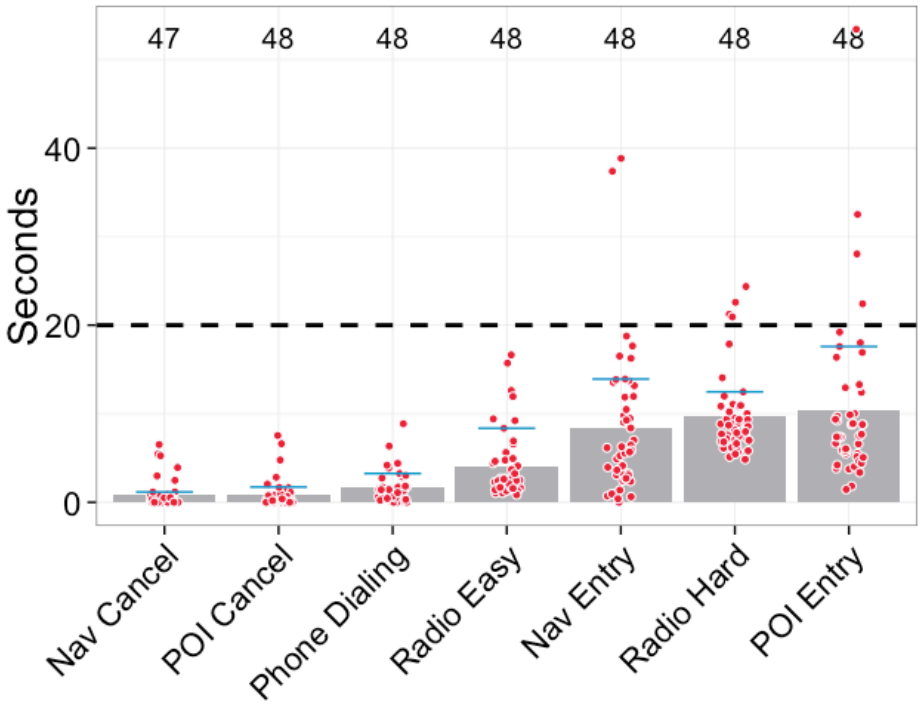

Figure 28: Cumulative glance time to the center cluster (device) for each task period under study. Labeling as in Figure 12.

While NHTSA's guidelines (2013) assess glance behavior in terms of the total time a driver's eyes are directed away from the forward roadway (TEORT metric), the earlier Alliance (2006) guidelines consider the total time during a task that a driver's eyes are directed to a device, including glances associated with any aspect of the operation of that device, and specify a 20 second criterion. Initiating a voice-based task in each of the vehicles studied in the Phase I & II projects required pressing a push-to-talk button located on the steering wheel. In this regard glances to the push-to-talk button should conceptually be included in a glance to device calculation. However, separating glances to a button on the steering wheel from glances through the steering wheel to the instrument cluster is difficult if not impossible in many instances. Thus, for pragmatic purposes, the Phase I studies in the Lincoln MKS and the Phase II studies 2 (CLA) and 3 (Corolla) consider only glances toward the center stack in the glance to device metric. In contrast, Phase II Study 1 (Impala) included glances to the steering wheel / instrument cluster in addition to the glances toward the center stack in the glance to device metric since the Impala actively used a small display located in the center of the instrument cluster as an active component of voice-command tasks studied. This should be kept in mind in comparing glance to device data across the studies.

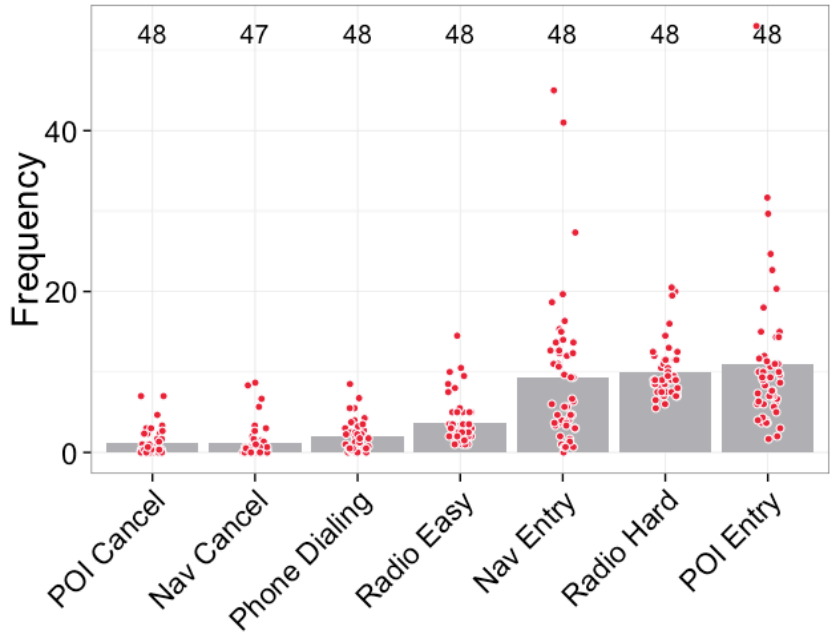

Figure 29: Number of glances to the center stack (device) for each task period under study. Labeling as in Figure 12.

Not surprisingly, the distribution of the total number of glances in the direction of the center stack closely follows the total glance time to the center cluster (device).

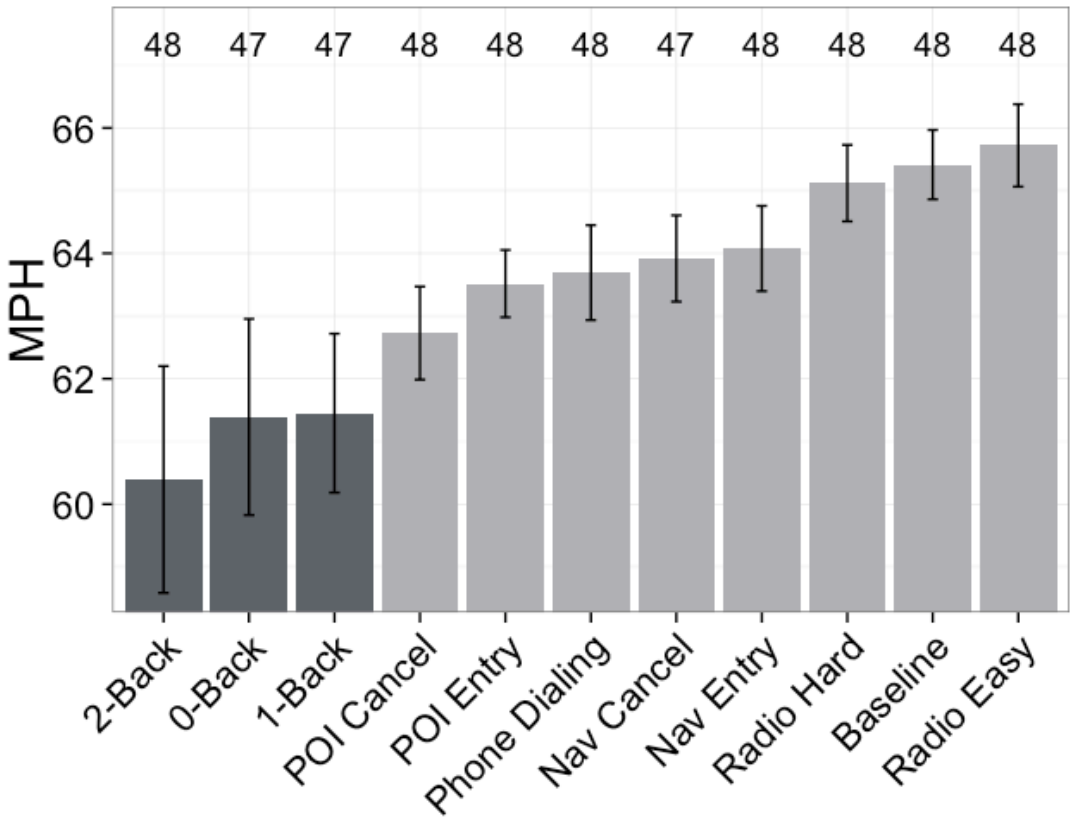

Figure 30: Mean velocity measures in miles per hour.

Appendix E: Experimental Task Details

"Voice" Destination Address Entry (Nav Entry)

The voice-command interface was used to enter full street addresses. The first two addresses were the same as those used in the Phase I studies to allow direct comparison across vehicle systems. Two additional addresses were added in Phase II Study 1: the participant's own home address (or other address that they know well) plus an additional fixed address. The fourth address was added to further assess the extent to which improvement in performance / reduction in demand is observable as the participant gains additional experience using the interface while underway. After experience with Experiment 1, the fourth address was dropped as part of several modifications to reduce overall study time and demand on participants.

Note: In the Phase I studies, addresses were presented to the participant auditorially and also displayed on a cue card located on the center of the steering wheel. In Phase II, it was decided to eliminate the cue card as looking at the card during the task would artificially inflate the off-road glance time metrics to some degree. It was the impression of the lead RA that few participants had difficulty recalling the addresses and it was decided to eliminate the use of the cue card.

The three addresses to be entered were:

- 177 Massachusetts Avenue, Cambridge
- 293 Beacon Street, Boston
- Participant's home address

The steps involved in using the interface are presented below. The notation **[xxxx]** indicates a hard button such as the push-to-talk button on the steering wheel, soft button on a touch screen, etc. Quotation marks ("") are used to indicate what the participant should speak. The **bold** text indicates the auditory prompts delivered by the system.

<u>Entune Premium Audio with navigation implantation in 2015 Toyota Corolla:</u>

[touch push-to-talk button]

How may I help you? → "Address" → **For direction to an address, please say the full address including the city and the state.** → "177 Massachusetts Avenue, Cambridge, Massachusetts" → **I heard 177 Massachusetts Avenue, Cambridge, Massachusetts, is that correct?** → "Yes" → **Starting guidance for a new route**

"Voice" Point-of-Interest (POI) Selection (POI Entry)

The voice-command interface was used to request POIs for specific locations (i.e. locations in a specific city as opposed to "nearest" which was the default mode in the Impala and some other systems). In developing this task in Phase II Study 1 (Impala), this structuring made the task longer than the "nearest mode" approach, but ensured that the system gave the same set of options no matter where (in terms of physical location) the request was made during the drive.

As with several other tasks, an "easy" and a "hard" level were developed. In the Impala, the two "easy" POIs were selected such that the final target address selection was shown on the top level display screen. This meant that the participant could select the target POI without any need to search the list by paging or scrolling. The "hard" level required paging / scrolling down 1 page level. Adjustments were made in the target POIs in the subsequent studies so that the amount of scrolling / list searching was comparable across the vehicles. This approach supports comparison of the relative ease-of-use, layout, and physical characteristics of each of the interfaces rather than introducing bias that might have occurred from how POIs originally selected for Study 1 happened to fall within the menu / list structure of the infotainment systems tested later. As noted previously, the number of POIs was dropped from 4 in Study 1 to 3 in Studies 2 & 3 to reduce overall study time. The POIs in Study 3 did need to be changed from those used in Study 2 to preserve a comparable level of search difficulty.

The three POIs to be entered in Study 3 were:

- Tavern in the Square, Cambridge (easy)
- Liberty High School, Boston (easy)
- Lamont Library, Cambridge (hard)

The steps involved in using the interface in the current vehicle are presented below. The notation **[xxxx]** indicates a hard button such as the push-to-talk button on the steering wheel, soft button on a touch screen, etc. Quotation marks ("") are used to indicate what the participant should speak. The **bold** text indicates the auditory prompts delivered by the system.

<u>Entune Premium Audio with navigation implantation in 2015 Toyota Corolla:</u>

[touch push-to-talk button]

How may I help you? → **"**Find dining in a city" → **Say the name of the city and state where you want to find dining** → "Cambridge, Massachusetts" → **Showing dining in Cambridge, select the one you want by name or number** → "line 4" → **To navigate to this point of interest say go there, or you can say call them, or mark location** → "Go there" → **Navigating to Tavern in the Square** (for the hard level, you must say next on the "select a line number" menu, to move forward 1 full page)

Canceling Navigation Tasks

As detailed in the reports covering Study 2 in the Phase I work, canceling a route entered into the navigation system can be a challenging task. This can particularly be the case depending on the specific command syntax required to complete the task in a given voice-command implementation. The same command structure for canceling is used for both destination entry and POI, meaning that a minimum of 6 instances of engaging with this task were potentially available for each participant.

Entune Premium Audio with navigation implantation in 2015 Toyota Corolla:

[touch push-to-talk button]

How may I help you? → "Cancel Route"→ **Your route has been cancelled.**

"Voice" Contact Phone Calling

The voice-command interface was used to call 4 contacts. The first two were single name entry cases and the second two involved calling a specific phone for contacts having multiple phone numbers. These same tasks were used in an MIT/IIHS study (Mehler, Kidd, et al., 2015) that was in part stimulated by our CSRC Phase I work. The study compared visual-manual and voice-involved interfaces in two vehicles: a 2013 Chevrolet Equinox and 2013 Volvo XC60. Doing the tasks in the same manner in the Phase II studies provides a useful comparison point and can be seen as an important contribution to the development of a broader dataset. The first two tasks were categorized as "easy" and the second two as "hard" in the context of a multi-layered menu design used in many visual-manual and some voice-based interfaces. (Please see Mehler, Reimer, and McAnulty (2014) for a discussion of the background on the meaning of "easy" and "hard" (p. 11).) In the "one-shot" voice interface design used in the Impala, CLA, and Corolla, the "easy" and "hard" tasks are virtually equivalent in demand in terms of the number of steps involved as detailed below.

The four contacts are:

- Mary Sanders
- Carol Harris
- Pat Griffin on mobile
- Frank Scott at work

The steps involved in using the interface in the current vehicle are presented below. The notation [xxxx] indicates a hard button such as the push-to-talk button on the steering wheel, soft button on a touch screen, etc. Quotation marks ("") are used to indicate what the participant should speak. The bold text indicates the auditory prompts delivered by the system. The (xx) symbols indicate a confirmation step that was not always presented by a system. It is presumed that this implementation employed a set of logic rules concerning whether confirmation was warranted. (See comments above on "easy" and "hard".)

Entune Premium Audio with navigation implantation in 2015 Toyota Corolla:

[touch push-to-talk button]

Easy:

How may I help you? → "Call Mary Sanders." → **(Call Mary Sanders, correct? → "**Yes"**)** → //system begins calling//

Hard:

How may I help you? → "Call Pat Griffin on mobile." → **(Call Pat Griffin on mobile, correct? → "**Yes."**)** → //system begins calling//

Task is cancelled by pressing [hang-up phone button] on steering wheel.

Visual-Manual Radio Reference Tasks

The same task structure for visual-manual radio "easy" and "hard" tuning tasks employed in Phase I Study 2 were employed across Phase II Studies 1-3. The "easy" task required manual engagement with a single touch screen preset button selection (or hard button if available). The 2015 Corolla radio interface provided 6 touch screen located preset buttons ordered vertically on the left side of the display screen. The "hard" task required 2 button engagements and manual rotation of the fine tuning knob to obtain a specified frequency. In the Corolla, a hard button activated the audio mode and touch screen buttons supported selecting AM or FM frequency bands (or XM radio, CD, USB, etc.). A rotational fine tuning knob was located to the right of the display screen. The goal was to have the driver interact with the radio in a manner as close as possible to the NHTSA (2013) specified manual radio reference task in terms of the number and type of manual engagements required. Conceptually this was used to establish a reasonably "standard" visual-manual load reference point for comparing against the other tasks. In Study 2 in the CLA, the target radio station was changed from 1470 AM used in earlier studies due to an inability of the CLA's radio to lock into this station. AM station 1030 was substituted and the set-up adjusted so that a comparable number of manual turns of the fine adjustment knob were required for tuning. This same station selection was carried over to Study 3 in the Corolla.

The four task goal states presented in order were:

Preset 1

Preset 5

1030 AM (Replaced 1470 AM used in earlier studies as noted above.)

100.7 FM

Steps:

Easy: Radio interface is active, radio is ON

[Preset 1] on touch screen

Hard: Neutral Menu Screen, radio is ON

[Audio] hard button to activate bandwidth menu

[FM] on touch screen to switch from AM to FM

(Tuning knob from 107.9 to 100.7)

N-back Auditory-Cognitive-Vocal Calibration Reference Task

This is the standard MIT audio presentation / verbal response task presented at 0-back, 1-back, and 2-back levels (Mehler, Reimer, & Dusek, 2011). The three levels were presented as 30 second task blocks as was done in the Phase I Study 2. The order of presentation of the three difficulty levels is randomly distributed across the sample.

Appendix F: Full Inclusion / Exclusion Criteria

Potential participants were initially screened by e-mail or phone interview on the points below, and were assessed again by direct in-person questioning by a research associate when they arrived at the research site.

For **inclusion** in the study, participants needed to:

- Be between the ages of 20 and 69 years of age
- Report having held a valid driver's license for more than three years and show a current driver's license upon arrival
- Report driving on average three or more times per week
- Clearly understand and speak English as confirmed by a research staff member.
- Indicate willingness to drive one of the MIT AgeLab's research vehicles "such as a Volvo XC60, Chevy Equinox, or Chevy Impala, from MIT onto an interstate highway and back". (Prior to answering this question, potential participants reviewed and signed an informed consent form that provided extensive detail on the nature, content, and duration of the study.)

On the basis of self-report, individuals were **excluded** from the study for:

- Answering negatively the statement "Are you in reasonably good health for your age?"
- Answering affirmatively the statement "Have you been the driver in a police reported accident in the past year?"
- A major medical illness resulting in hospitalization within the past 6 months
- A diagnosis of Parkinson's, Alzheimer's disease, dementia, mild cognitive impairment (MCI), or any other neurological problems
- Currently being treated for a psychological or psychiatric disorder
- Ever had any of the following:
    - Heart failure
    - Angioplasty or coronary artery bypass grafting (CABG)
    - A pacemaker
    - A stroke or transient ischemic attack
    - A diagnosis of diabetes (This factor was from earlier studies and was not otherwise considered necessary for the current work.)
- Using any of the following medications in the past 12 months:
    - Anti-convulsant medication

    - Immunosuppressive drugs or cytotoxic drugs. (The intent was to screen for serious medical conditions – if an immunosuppressive medication was for treatment of mild to moderate arthritis that does not markedly impact the individual's ability to drive, they could be included.)
    - Anti-psychotic medication
    - Medications to treat a major medical condition such as cancer
- Potential participants were asked if, in the past two days, they had used any medications that made them drowsy. If they regularly used medications that cause drowsiness, they were to be excluded. If medication was used on a limited basis such as a cold medicine, individual could be rescheduled for a time slot when they had been off medication for 48 hours or more. Individuals regularly using sleep medications such Ambient were only to be scheduled for afternoon appointments.

Additional Notes:

- Eye glasses and contacts were acceptable.
- Participants were asked not to wear sunglasses that blocked the view of the pupils for eye gaze assessment purposes. Individuals who were not willing or able to drive under these conditions were withdrawn from the study.
- Participants were asked to turn-off their cell phones prior to starting the vehicle training and driving portion of the study so that they would not be interrupted during the study. Individuals who were not willing or able to drive under these conditions were withdrawn from the study.
- The research associate who was to ride in the vehicle with the participant had full authority to withdraw a participant if they had concerns about the participant's wiliness or ability to operate the vehicle in a safe manner, particularly while engaging in the secondary tasks. As detailed in the analysis sample section of the results, a limited number of participants were withdrawn during the parking lot portion and during the on-road portion due to such concerns.