White Paper 2013-18A

The Effects of a Production Level "Voice-Command" Interface on Driver Behavior: Summary Findings on Reported Workload, Physiology, Visual Attention, and Driving Performance

Bryan Reimer and Bruce Mehler

November 18, 2013 (Replaces limited release v. 2013-18 dated November 4, 2013)

Abstract - This report summarizes key results of an on-road study assessing perceived workload, physiological arousal, visual attention, and basic driving performance metrics while drivers engaged in a number of tasks with a production version, in-vehicle voice-command system. The same metrics were also evaluated while participants carried out an implementation of the manual radio tuning reference task (Driver Focus-Telematics Working Group, 2006) and three levels of an audio-presentation / verbal response delayed digit recall task (n-back) that is known to produce graded levels of cognitive demand. Extensive training on all tasks was provided prior to assessment under highway driving conditions. Results for an analysis sample of 60 drivers, equally distributed across both genders and two age groups (20-29 and 60-69), are summarized here and presented in detail in an associated technical report (Reimer, Mehler, Dobres, & Coughlin, 2013). Depending on the task assessed and measure evaluated, both positive features and concerns associated with the use of the voice interface were identified. Physiological arousal during the voice tasks was comparable or lower than that observed during the more difficult level of manual radio tuning task as measured by skin conductance and heart rate, respectively. Perhaps most notable was the identification of a high level of visual demand / engagement during selected tasks, such as the use of the voice-command interface for entering addresses into the navigation system. It also appeared that different age / gender groupings tended to interact with the voice system in different ways.

These findings highlight that implementations of voice interfaces can be highly multi-modal and are not necessarily free of visual-manual demands on attentional resources. If one were to apply the current National Highway Transportation Safety Administration (NHTSA) visual-manual distraction guidelines to the tasks assessed, a number of "voice" interactions would not meet the total off-road glance time criteria of the guidelines. While these data were not collected in full alignment with NHTSA's simulation-based guidelines, the overall structure and metrics are similar, and so this work raises a number of important questions. It is clear that visual demand needs to be considered in the design of multi-modal voice interfaces. This highlights the question of how an acceptable level of visual demand should be defined in the context of multi-step and extended task time interactions that characterize activities involving voice-command interfaces. Finally, the results illustrate the necessity for additional research assessing the generalizability of these findings to other production level and hand-held "voice" interactions, and in developing methods of quantitatively assessing the net attentional costs and benefits of providing drivers with information across different modalities. Voice interactions can play an important role in the vehicle environment. Optimizing the selection of activities in which the driver utilizes voice interaction and the appropriate design of displays will help to maximize driver attentional focus towards information necessary for vehicle operation, while allowing, where appropriate, interactions with interfaces for comfort, convenience and communication functions.

Introduction

Voice-command interfaces have been proposed, and in some cases aggressively advertised, as a means to allow drivers to engage with an expanding array of entertainment and connectivity options in the modern automobile while keeping their eyes on the road and hands on the steering wheel. This is an intuitively appealing concept, and a reasonably respectable body of computer science, psychology and human factors based laboratory, simulator, and test track studies have identified situations in which primarily experimentally created voice interfaces have shown distinct advantages over visual-manual interfaces in terms of primary task performance (driving or driving like tasks) and glance behavior (see Barón & Green, 2006; Lo & Green, 2013; and Reimer, Mehler, Dobres & Coughlin, 2013 for reviews). However, it is not clear how directly the interactions observed with these types of experimental, hand-held, or aftermarket, voice interfaces generalize to production level automotive systems (systems integrated into the vehicle directly by the manufacturer). Assessments of production level systems have been far fewer in number and generally examine a limited set of task characteristics with modest sized samples (Carter & Graham, 2000; Chiang, et al., 2005; Harbluk, Burns, Lochner, & Trbovich, 2007; Owens, McLaughlin, & Sudweeks, 2010; Shutko, et al., 2009; Shutko & Tijerina, 2011). Furthermore, only Chiang, et al. (2005) and Owens, McLaughlin, & Sudweeks (2010) assessed driver behavior with production systems under actual field driving conditions. While the findings reported in the latter studies have generally presented voice interfaces in a positive light, there is some evidence that voice-based interfaces may not always be completely free of visual-manual demand (see Reimer, Mehler, Dobres, et al., 2013 review). In addition, some questions have been raised about the extent to which "eyes on the road" necessarily equate to "mind on the road". In other words, to what extent might interaction with a voice interface or audio content from e-mail or a phone conversation result in cognitive demands or absorption that might produce another critical form of distraction, ultimately resulting in a loss of situational awareness?

The study summarized in this report was conceived and implemented with the goal of developing a comprehensive assessment of a production-level voice command interface and the demands such a system places on drivers under real-world highway driving conditions. Metrics included visual behavior, physiological arousal as a measure of cognitive demand, driving performance measures, and self-reported workload in younger (20-29 years) and relatively older (60-69 years) samples of drivers broadly representative of the general driving population. Voice control of the radio, music selection from a connected MP3 device, and voice dialing of a stored phone number were selected as basic entertainment and communication tasks. Voice

entry of a full street address into a navigation system was of particular interest, since manual entry of addresses into navigation devices while underway is generally recognized as being highly visual-manually demanding. Some OEMs have chosen to lock-out manual entry of addresses while the vehicle is underway, while others allow it. Objectively evaluating the extent to which a voice entry implementation makes this task acceptable under driving conditions is thus quite relevant. Implementations of easy and hard levels of a radio tuning task were developed to support a "side-by-side" comparison of identical tasks using the visual-manual interface and voice interface for the same functional activity in the same vehicle. In addition, three levels of a delayed digital recall task (audio presentation of stimuli with a verbal response from the driver) were included. This task, known as the "n-back", has been used extensively in research by our group (Mehler & Reimer, 2013; Mehler, Reimer, & Coughlin, 2012; Reimer & Mehler, 2011; Reimer, Mehler, Wang, & Coughlin, 2012) and is known to produce graded levels of cognitive demand as reflected in various physiological measures and self-report levels of workload. It was anticipated that the multiple cognitive demand levels represented by the n-back task could be used as a "ruler" against which various responses to the other tasks might be compared. Ranney and colleagues (Ranney et al., 2011) suggested in their exploratory work with the measure that the 2-back condition (the hardest level assessed) could "serve as a starting point for setting a limit for acceptable 'dose' of cognitive distraction" (p. 52).

Methods

An MIT owned 2010 Lincoln MKS with factory installed voice-command systems (Ford SYNC™ for voice control of the phone and media connected by USB and the "next-generation navigation system" with Sirius Travel Link) was selected as a convenient example of a widely available production level voice interface when this project was initiated in 2011. Funding for this project was initially secured from a private driving safety foundation (The Santos Family Foundation) and the United States Department of Transportation's New England University Transportation Center program. Support for significantly expanding the sample size of what was originally conceived of as a pilot study, as well as funding for follow-on research, was subsequently obtained from the Toyota Collaborative Safety Research Center (CSRC). Participants were provided with step-by-step training on how to complete each of the tasks under study in the most efficient fashion using the default system settings (see Reimer, Mehler, Dobres, et al., 2013 for complete methods and protocol). Each secondary task was then presented twice under highway driving conditions. Except for the phone tasks, all other interface tasks were presented in a counterbalanced design to control for order effects. An

analysis sample of 60 drivers was obtained, equally balanced by gender and across the two age groups (20-29 and 60-69 years).

Primary Findings

This section highlights key findings. For detailed results, statistical assessment, several alternate analyses of the data, and additional discussion and comment, please refer to the complete technical report (Reimer, Mehler, Dobres, et al., 2013).

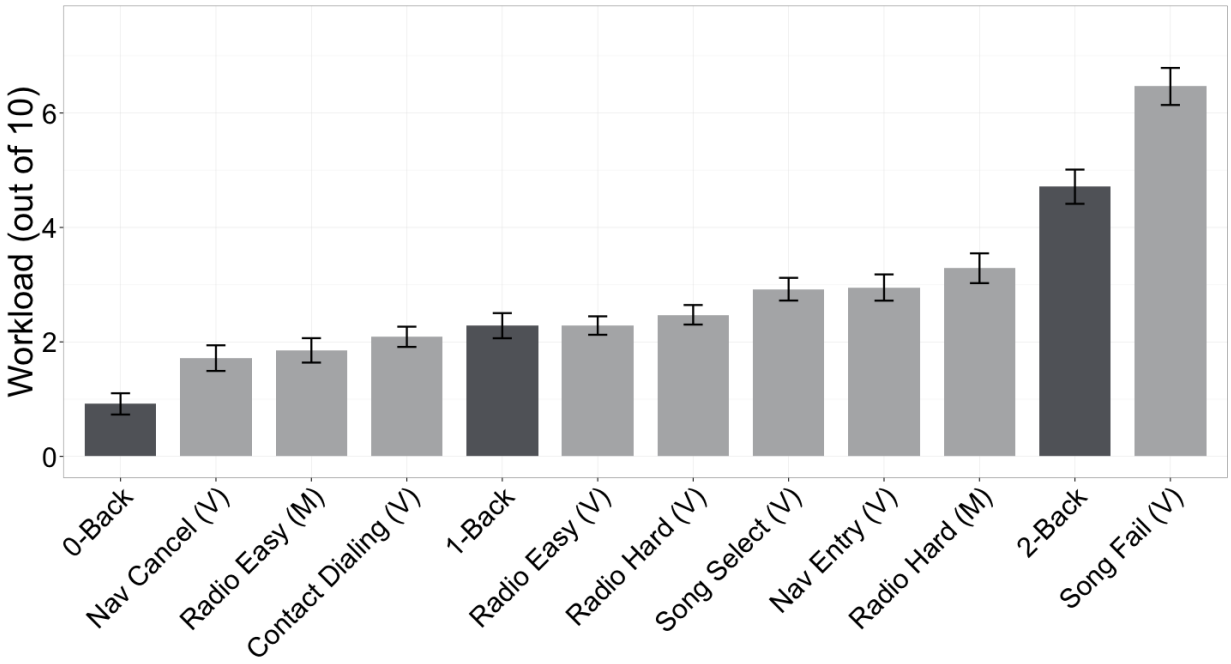

Figure 1. Tasks listed in ascending order for mean reported workload level. N-back reference tasks are denoted with darker bars. Error bars represent 1 SEM. Tasks marked (V) used the voice interface. Tasks marked (M) utilized traditional manual/tactile interactions. (Figure adapted from (Reimer, Mehler, McAnulty, et al., 2013).)

Self-Reported Workload (Figure 1) - As might be expected, a task such as manually selecting a preset on the radio (an easy task) was given a relatively low workload rating, while the hard version of the manual radio tuning task employed in the current study (press the volume button, switch the radio band, and manually rotate a tuning knob to a specified station) was given the highest workload rating of all valid interface tasks. Only the 2-back high demand surrogate task and the song selection fail task, which was deliberately made impossible to complete, were rated higher on workload than manual radio tuning. In line with the potential advantages of a voice-interface, self-reported workload ratings for the voice-control version of the radio (hard) tuning task were notably lower than ratings for the manual version. All other

voice-command tasks (again excepting the song fail task) also resulted in lower workload ratings than the manual radio tuning (hard) task. These perceived workload ratings align well with the idea that the manual radio tuning task represents a "socially accepted, reasonably-demanding reference condition" (Driver Focus-Telematics Working Group, 2006; p. 46) and that the other automotive system interfaces have been designed to keep demand at or below this level.

N-Back Scaling & Physiological Indicators of Workload - As mentioned previously, the n-back task provides a useful scalar reference for comparing certain aspects of the demand associated with other tasks, and the 2-back level has been suggested as a conceptual "dose limit" for the level of acceptable cognitive distraction. The low demand 0-back task was given a markedly lower self-reported workload rating than the other tasks, the moderately demanding 1-back task was ranked intermediately among the other tasks, and the high demand 2-back task was rated higher than all others except for the "Song Fail" task, which was deliberately designed to be impossible to complete successfully.

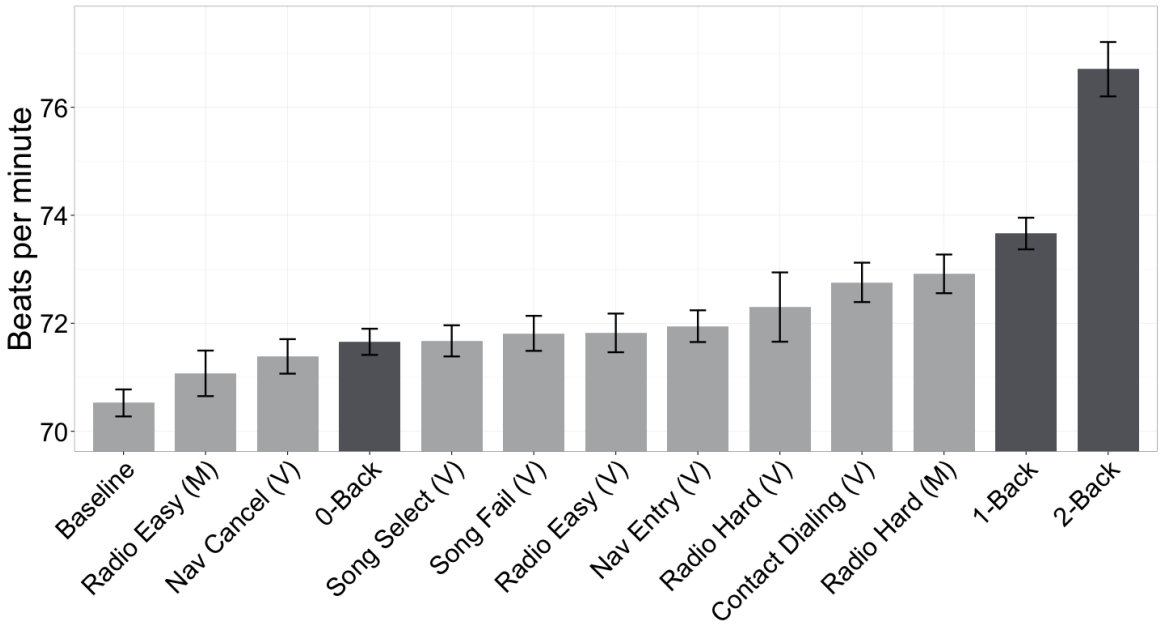

Figure 2. Tasks listed in ascending order for mean heart rate. The baseline shown represents an average of all baseline periods recorded prior to each task. Error bars represent 1 SEM. (Figure adapted from (Reimer, Mehler, McAnulty, et al., 2013).)

Heart rate (Figure 2) and skin conductance level (SCL) values scaled in a step-wise fashion from baseline driving across the three levels of the n-back, replicating previous on-road findings (Mehler, Reimer, & Coughlin, 2012). Somewhat in contrast with our initial expectations, while all of the tasks were associated with increases in group mean heart rate and SCL values relative

to baseline driving, the magnitude of increase for the most challenging tasks all fell near or clearly below the 1-back level. We had anticipated that what we initially perceived to be highly cognitively engaging tasks might approach a 2-back level of arousal. Based on these data, the 1-back's mean heart rate and SCL responses may represent good reference points for defining the high end of an acceptable range of physiological arousal when performing secondary tasks in the vehicle. Alternately, an 85th percentile criterion might be considered, which would take into consideration more of the variability of individual response patterns, and argue for a higher reference point, such as the 2-back. The question of how physiological arousal reference points might most appropriately be applied is an area worthy of further investigation.

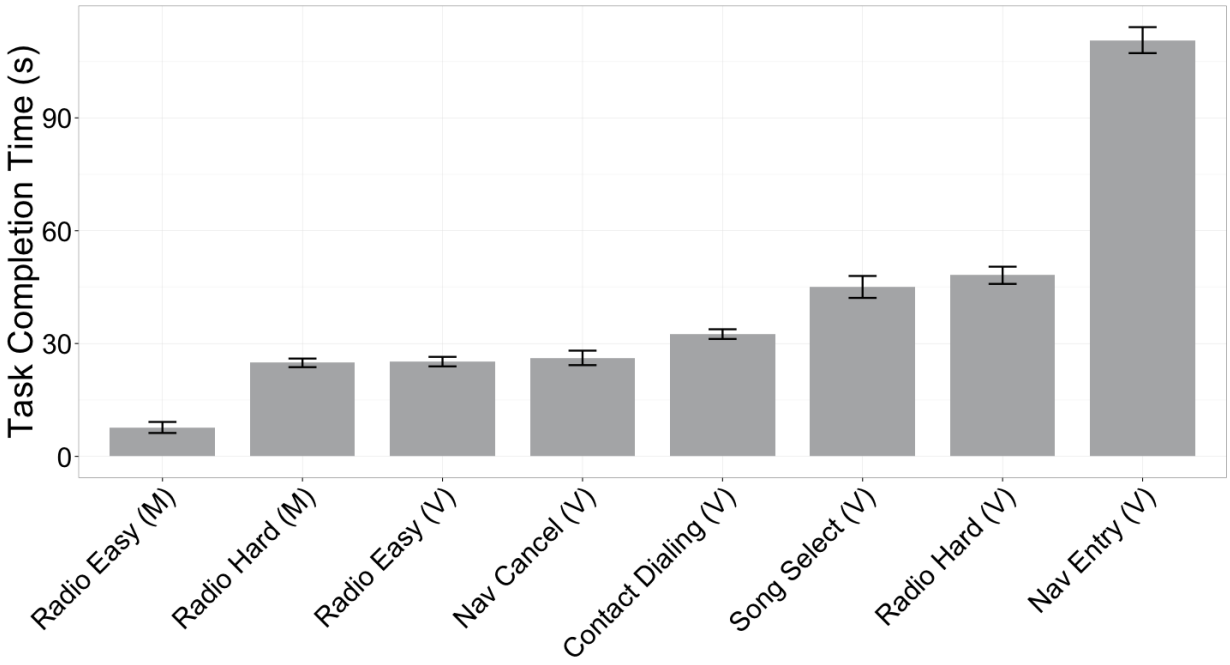

Figure 3. Tasks listed in ascending order for the amount of time needed to complete each task. Error bars represent 1 SEM. Tasks marked (V) used the voice interface. Tasks marked (M) utilized traditional manual/tactile interactions. (Note: the n-back tasks and song fail task are of fixed duration and therefore are not represented in the plot.)

Task Completion Time (Figure 3) - Manually selecting a radio preset (radio easy) took the least amount of time to complete (mean of less than 8 seconds from the prompt "begin" to completion of the task). In contrast, verbally requesting a radio preset took as long as the manual tuning (hard) task (25 seconds). The verbal version of the radio tuning (hard) task took nearly twice as long as the manual version (48 seconds). Thus, while the overt workload (self-report and physiological arousal) of using the voice interface was lower than what was observed for the manual tasks, the total time that attention was divided between a secondary task and driving was much greater. Notably, mean task completion time for voice-command entry of a street address into the navigation system was almost two minutes. This questions as

to whether assessments of voice interactions need to include a consideration of overall task time. Is there a point at which an overtly low to moderately demanding task becomes problematic due to length of engagement? Does the ability to self-pace a task effectively fully compensate for extending the time required to complete the task? Burns, Harbluk, Foley and Angell (2010) provide a very useful discussion on total task time as an important metric in considering designs intended to limit distraction.

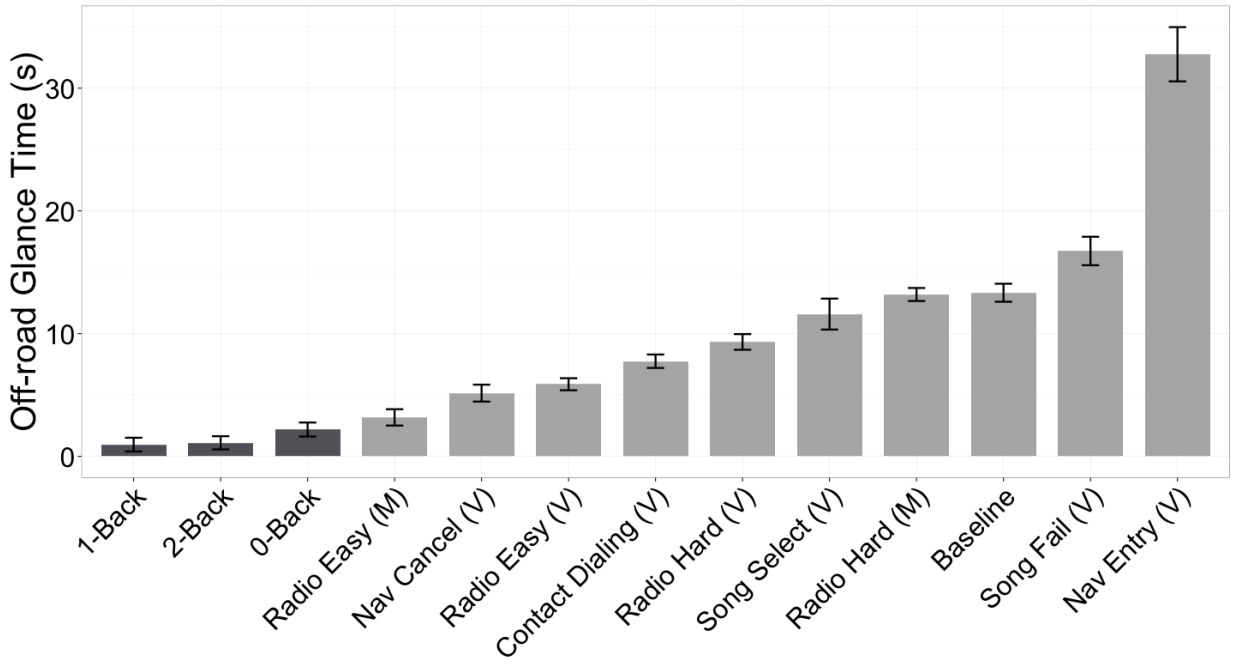

Figure 4. Tasks listed in ascending order for the amount of off-road glance time that occurred during the completion of each task. Error bars represent 1 SEM. Tasks marked (V) used the voice interface. Tasks marked (M) utilized traditional manual/tactile interactions. Baseline represents the mean off-road glance time for 2 minute periods averaged across all 7 baselines collected.

Visual Demand - Perhaps the most notable findings appear in the glance metrics. While previous work demonstrates that some voice-interface tasks are not completely free of visual-manipulative demand (Chiang, Brooks, & Weir, 2005; Maciej & Vollrath, 2009; Neurauter, Hankey, Schalk, & Wallace, 2009; Shutko, Mayer, Laansoo, & Tijerina, 2009) (see review of these studies in Reimer, Mehler, Dobres, et al., 2013), findings for total glance time when using the voice-command entry of addresses into the navigation device in this study are particularly of interest. Mean total off-road glance time during the address entry task was 32.8 seconds for the sample as a whole (25.9 seconds for younger adults and 41.7 for older adults) (see Figure 4). As detailed in the technical report, if one were to extend the current visual-manual distraction guidelines to the "voice" interactions, sample, and methods employed in this study, the voice-based address entry task would fail to meet NHTSA's (2013) new 12-second maximum eyes off-

road threshold. The same result holds if the Alliance of Automobile Manufacturers' (Driver Focus-Telematics Working Group, 2006) time to device metric if a 20 second threshold (criterion 2.1 A) is employed.

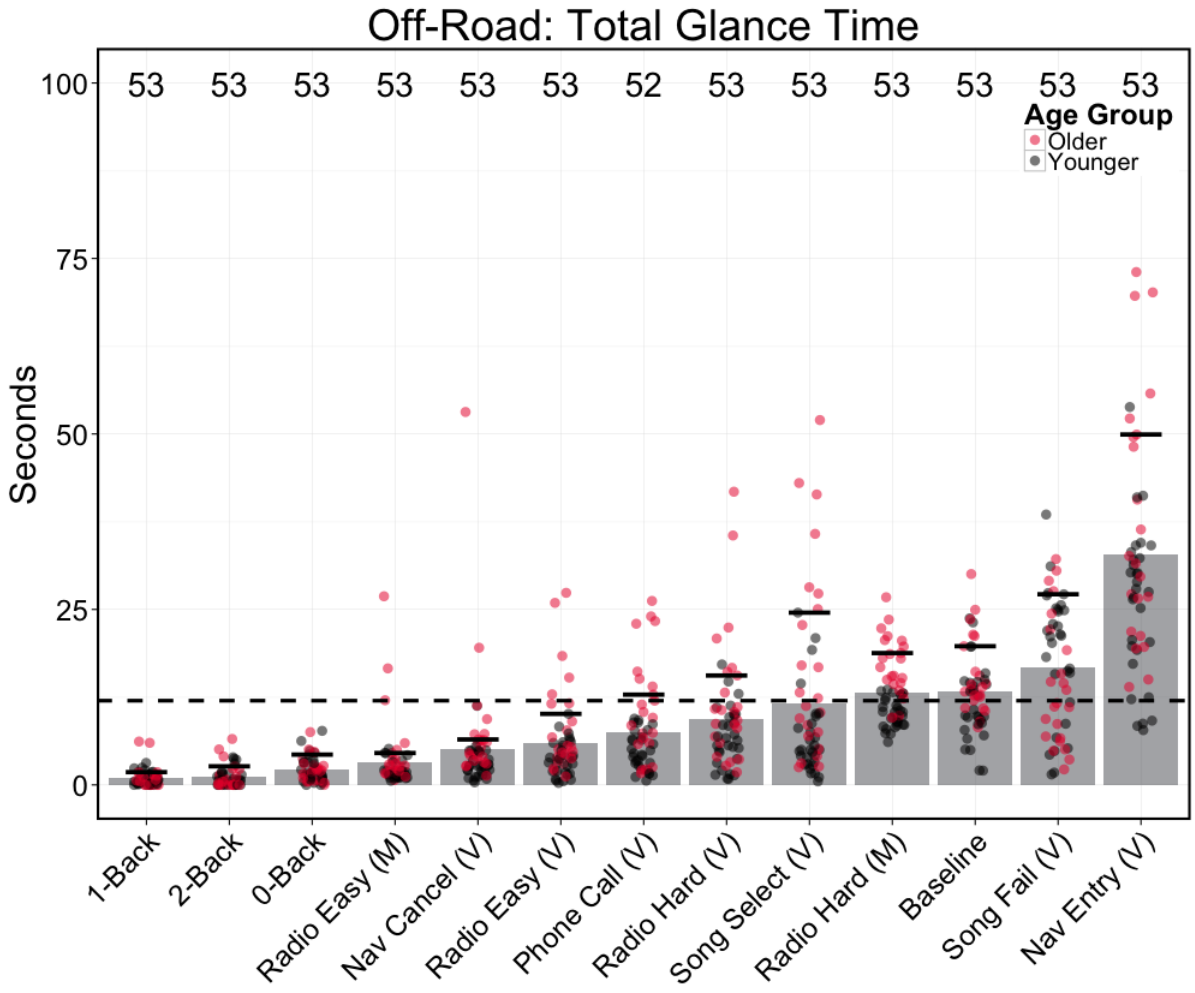

Figure 5. Total off-road glance time for each task with the NHTSA (2013) 12 second threshold shown as a dashed line. The individual line segments above each bar represent the 85% point in the sample distribution for each task. One outlier data point in the Nav Entry task is excluded from view to improve the readability of the plot. Note that the NHTSA threshold values are shown here for discussion purposes only since, among other considerations, the sample does not conform to the NHTSA recommended age distribution and the data was collected under real driving conditions as opposed to the specified simulation conditions.

Only 13.3% of the younger participants met NHTSA's off-road glance time criterion, while 0% of the older adult sample met it (see Figure 5, which presents the data showing individual participant performance). As is clearly visible in the detailed figure, a majority of the longer glance times were associated with older adults. Voice-based phone contact dialing and song selection also appear problematic, depending upon what metric and threshold is used (see

technical report). On the other hand, values for visual demand associated with the voice version of the radio hard tuning task appear nominally lower than with manual tuning, thus suggesting the benefit commonly expected of a voice interface.

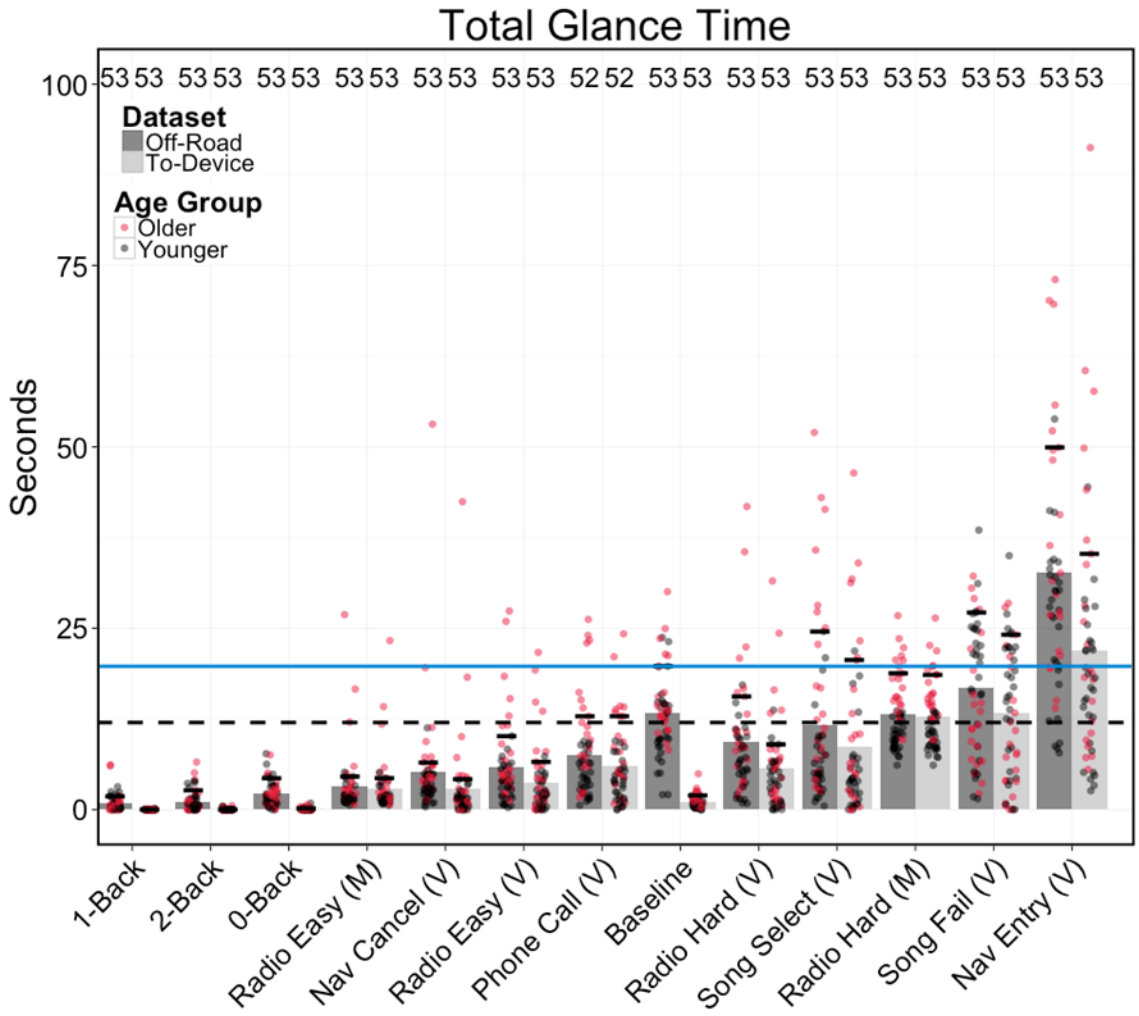

Figure 6. For each task, dark bars represent the NHTSA total Off-Road Glance Time metric and Light bars The Alliance Total Glance Time to Device (criterion 2.1 A) metric. NHTSA 12s threshold shown as a dashed line and the Alliance 20s threshold in blue. The longer individual line above each bar represents the 85% point in the sample distribution for each task.

Much of the glance behavior observed during voice tasks was associated with looking at a console display screen to view options presented by the system, such as available commands or to make a selection from a list if the system identified multiple options for street names during address entry. The number of support and confirmatory steps was much greater for the navigation entry task than for any of the other tasks. Consequently, the total glance time metrics were directly impacted by the number of glances in the task (see Figure 7).

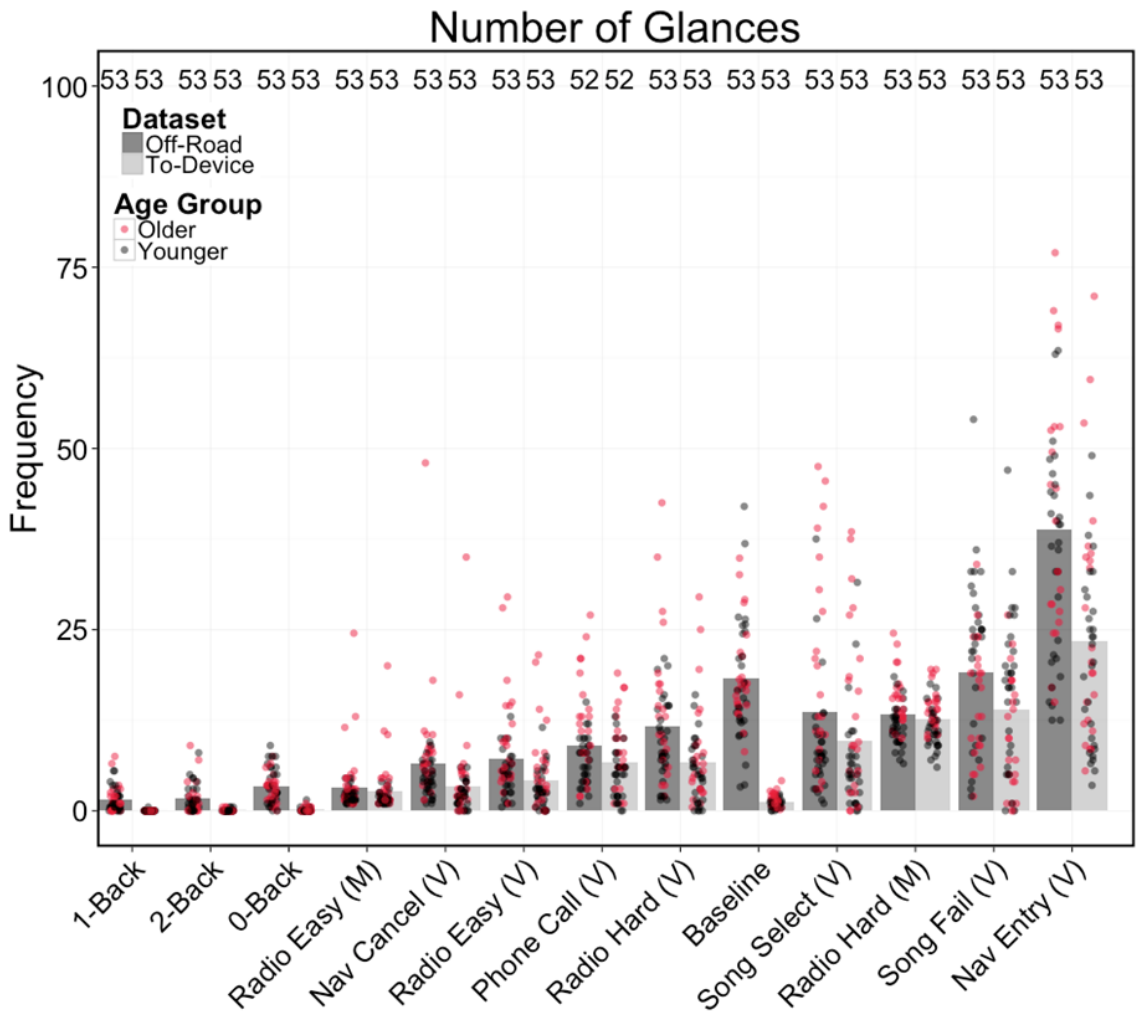

Figure 7. Tasks listed in ascending order for number of off-road glances. The baseline shown represents the combined baseline periods recorded prior to each task and represent mean behavior over 2 minute long intervals.

A consideration of mean individual glance duration and the percentage of glances longer than 2 seconds for the navigation entry task fell fully in line with guidelines issued both by The Alliance (criterion 2.1 A) and NHTSA for visual-manual interface design (please refer to the complete technical report, Reimer, Mehler, Dobres, et al., 2013). Thus, there are no issues from a current guideline perspective around the individual glance characteristics displayed during the navigation task; it is the number of glances involved that drive the total glance time metrics. The support displays presented during the tasks are generally employed to reduce the amount of cognitive load that would be placed on the driver by having to remember specific command phrases or needing to listen to an extended list of destination options. There clearly is a human

factors design challenge to find a balance across visual, manual, auditory, and associated cognitive demands to optimally support drivers of differing capabilities and preferences.

Vehicle Control & Safety-Critical Events - Across the tasks, driving performance data suggest self-regulatory patterns associated with both voice and visual-manual interface use, e.g. reductions in driving speed. Standard deviation of velocity generally was lower during task periods relative to baseline driving, with the exception of the address entry which showed a simple mean value similar to baseline driving. It should be noted, however, that lower standard deviation of velocity values for most of the tasks could be an artifact of the shorter durations of these periods.

The highest major steering wheel reversal rate values (number of steering wheel inputs exceeding an angular reversal gap of 3° (Östlund et al., 2005)) appeared during the manual Radio Hard tuning task, followed by the manual Radio Easy (preset selection) task. This pattern makes intuitive sense, since the driver must remove one hand from the steering wheel to complete the task and must devote visual-motor control attention to manually manipulating the radio interface. In terms of clear safety-critical activities, such as high g acceleration incidents, no events were recorded that exceeded thresholds typically utilized in current research. Only one event greater than 0.30g appeared during voice interface activities (a lateral acceleration of 0.31 g during a song selection task). Three additional high g events occurred (between 0.25g and 0.30g) during baseline driving. No events at the 0.25g or greater level were observed during the demanding address entry navigation system task. There was a total of 26 0.20g or greater acceleration events, 11 of which occurred during baseline driving and 15 during task periods (5 during the navigation entry, 1 during navigation cancel, 3 during song selection, 2 for voice radio tasks, 3 for manual radio tasks, and 1 during a 2-back task). Since the total analysis period durations were quite similar for both baseline and task intervals (baseline periods totaled 14 minutes and the mean total task time across the sample was 14.6 minutes), the data suggest that engaging in the secondary voice control tasks resulted in only a nominal incidence of moderate 0.20g or greater acceleration events relative to single task (baseline) driving behavior. In as much as acceleration events are a reasonable measure of an adverse driving event, and that typical thresholds used to define a near crash include braking at greater than 0.50g or lateral acceleration greater than 0.40g (Fitch et al., 2013), this suggests that use of the voice interface studied, although visually and cognitively demanding, did not result in an overt increase in adverse driving events or control corrections during the period studied.

Discussion

Taken in aggregate, both positive features and concerns associated with the use of the voice interface appear in the data. Of particular note was the identification of a high level of visual demand / engagement during selected tasks, such as the use of the voice-command interface for entering addresses into the navigation system. When present, this engagement can often be logically linked to overt requirements in the interface design that require the driver to look multiple times at information presented on a visual display, such as a listing of street or city names from which a selection is required. In addition, a number of drivers were observed to engage in what could be characterized as Orienting Responses (ORs). These are instances in which drivers spoke directly to the graphical user interface, oriented their bodies towards it, or acted in a way that suggested the voice system was perceived to be "in" the device screen. This phenomenon is described in more detail in the technical report. These findings highlight that implementations of voice interfaces can be highly multi-modal, not necessarily free of visual-manual distraction, and may not meet NHTSA's visual-manual distraction guidelines if they were applied to such interactions.

NHTSA has explicitly stated that the guidelines are voluntary recommendations, and that there is significant need for ongoing research to determine if the guidelines should change as the science in this area evolves. NHTSA (2013) further stated that the guidelines "are currently not applicable to the auditory-vocal portions of human-machine interfaces of electronic devices." The 2006 Alliance guideline document (Driver Focus-Telematics Working Group) makes it clear that this industry-developed 20 second total glance time to device criterion (2.1 A), was based on research data on interactions with a traditional radio interface, and it is noted that the influence of control type (i.e. then emerging automotive implementations of technologies such as touch screens and voice control) on the proposed criteria should be addressed in future research. As is apparent in the results of the present study, this does present the designer with some open questions of what visual-manual demand criteria to use when audio-vocal components are present. In this spirit, we consider in the full technical report a number of alternative analyses of the data to assess the impact on the overall pattern of findings. These include:

- looking at both NHTSA's eyes off-the-forward-roadway metric and the glance-to-device metric originally adopted by the Alliance;
- consideration of NHTSA's 12 second total glance time criterion and the Alliance's 20 second criterion (2.1 A);

- use of mean values across two repetitions of each task during the on-road drive as well as the first and second trials alone;
- consideration of only error-free trials as defined by NHTSA's guidelines (2013);
- consideration of values for the entire sample as well as younger and older drivers as distinct groups.

While some of the alternate analyses influence whether certain tasks fall immediately above or below a particular cut-point in our sample, the overall pattern is quite consistent. Readers interested in the impact of the various alternate ways of evaluating the data are encouraged to review the technical report.

It should be emphasized that there is no *a priori* reason to assume that the concerns with visual demand observed here are unique to the specific voice-command interface tested in this study. Other systems employing similar design characteristics are likely to demonstrate demands. Future work will need to replicate these findings and assess the degree to which the results generalize to other production systems.

Further, the concerns raised in this work should be viewed within the context of our also finding positive aspects of a number of the voice interactions studied. Voice-command "tuning" of a specific radio station by selecting an appropriately located "press to talk" button and saying a command such as "Radio 89.7" is less physically and visually demanding that pressing a RADIO mode button, pressing FM, and then manually rotating a tuning knob multiple turns to find the desired station. Thus, the constructive challenge is to better understand how to balance various features of multi-modal interface design to optimally support driver attention.

Limitations & Next Steps – A number of limitations exist in this work; an extended discussion of possible limitations is presented in the full technical report and should be considered as part of a serious evaluation of this work. Perhaps most importantly, it is unknown if, or to what extent, exceeding the total glance time criterion (applying either the Alliance or NHTSA visual-manual distraction guidelines) represents a safety risk. Assessing the tasks undertaken in this study with these established metrics is seen as informative regarding attentional demand characteristics, rather than predictive of risk to drivers operating their own vehicles. Future naturalistic and/or epidemiological research will be necessary to gauge the degree to which interaction with systems such as those studied here present any significant elevation in actual risk. No crash or near-crashes were observed during data collection, nor were there aberrant vehicle kinematics (e.g., accelerations > 0.35g) recorded.

It should be noted that while participants were trained to use the most efficient shortcut steps identified for completing each task, they did not have extensive experience with voice command interfaces and interacted with the default mode of the system, which utilizes extensive guidance and confirmatory feedback steps. Thus, the response characteristics observed here and in other experimental assessments may, or may not, fully reflect the behavior of owners of vehicles equipped with voice-command interfaces.

A follow-on study is nearing completion that is intended to address several possible critiques of the original design. These include the recruitment of a sample that specifically corresponds to NHTSA's recommendations for age distribution, and evaluation of a slightly less demanding version of the manual radio tuning reference task that begins with the radio in the on position, per NHTSA's specification in their visual-manual distraction guidelines. Comprehensive analyses from this second study should be available shortly; however, initial results indicate that the basic findings from both studies are highly consistent. In addition, funding has been obtained from the Santos Family Foundation to add an arm to the follow-on study that explores the impact of using the "expert" modes of the voice system to reduce the amount of auditory prompting and the number of confirmatory responses required of the driver. It is conceivable that this optional mode of operating the system might result in a reduction in task time and some of the visual orienting to the display screen.

We also anticipate being able to extend this work to formally look at voice-command systems released by other manufacturers to address the question of generalizability. Preliminary reviews of other current production voice-command systems strongly suggest to the authors that the substantial level of visual demand observed in the voice-command system interactions in this study are not unique to any single manufacturer's implementation. Across all systems considered so far, there appear to be both strengths and weaknesses with each interface. Identifying the characteristics of each interface design that reduce draws on drivers' attention should help inform system designers regarding strategies for optimizing the role for voice interaction in the vehicle environment.

In Summary - It is clear that developing a firm understanding of how cognitive workload, visual, and manipulative demands come together to impact driver attention is a complex and multi-faced topic of study, requiring considerably more specific research than has been devoted to the topic to date. Data from this study clearly suggest that assessments of interface demands may need to more broadly consider compensatory activates that may play a role in drivers' overall effective workload and the resulting shift in the distribution of attention between "a task" and the roadway. While a broad literature base has developed around the utilization of

experimental "Wizard-of-Oz" voice systems, to date, only a limited number of reports have addressed production level embedded vehicle systems. Reports that do exist have limited sample sizes and are, in part, conducted in driving simulators. Studies of note do not provide a broad representation of the range of engagements central to many current production-level voice systems. Perhaps most importantly, no study was identified that provided an identical side-by-side comparison of the radio task completed in a classic visual manual context and through voice interaction. In this study, we aimed to overcome several limitations to develop a dataset that can help answer some of the unknown questions about the complexities of how drivers utilize advanced vehicle interfaces in an experimental setting that mimicked real life as closely as possible.

The findings from this project should be useful in informing the development of more effective multi-modal driver-vehicle interfaces that incorporate voice-command interaction. Better implementation of interface options available to system manufacturers may allow consumers to benefit from the potential advantages that voice interactions offer. Moreover, it is clear that a broader consideration of how age and gender differences impact interaction style with advanced vehicle interfaces, such as the one used here, is needed to more ideally provide all drivers with safe, convenient, and easy to use entertainment and communication systems.

Version Notes

The initial version of this white paper (2013-18) dated November 4, 2013 was given limited release for background briefings on this work. The current version (2012-18A) includes a refined description of the voice systems considered in this study based on feedback from representatives of the vehicle manufacturer. In addition, Figures 6 and 7 have been added.

Acknowledgements

Acknowledgement is extended to The Santos Family Foundation and the U.S. Department of Transportation's Region I New England University Transportation Center at MIT for providing the support for the initiation of this project. This funding provided support for the project's conceptual development, instrumentation of the research vehicle, as well as initially planned data collection and preliminary reporting on the project (Reimer, et al., 2013). The vehicle itself was purchased through funding from Ford Motor Company for an earlier project assessing the Ford Active Park Assist™ feature (Reimer, Mehler, & Coughlin, 2010). The current project was conducted without consultation or involvement of Ford Motor Company.

We are grateful to James Foley's initiative that lead to Toyota's CSRC funding to expand the original planned investigation to support a larger participant sample, the inclusion of additional tasks (longer drive), and detailed manual eye glance data review and reporting. We are especially thankful for the valuable, constructive comments James Foley and Kazutoshi Ebe of CSRC provided during the development of the study. Finally, this work would not have been

possible without the support of AgeLab staff and visiting scholars including: Jonathan Dobres, Hale McAnulty, Daniel Munger, Alea Mehler, Erin McKissick, Enrique Abdon Garcia Perez, Adrian Rumpold, Thomas Manhardt, Yutao Ba, Yan Yang, Ying Wang, Brahmi Pugh, Martin Lavalliere and Brendan Drischler in the development of the protocol, collection of data, 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. This document is disseminated under the sponsorship noted above.

References

Angell, L., Auflick, J., Austria, P. A., Kochhar, D., Tijerina, L., Biever, W., et al. (2006). Driver Workload Metrics Task 2 Final Report. Washington, DC: U.S. Department of Transportation National Highway Traffic Safety Administration.

Barón, A., & Green, P. (2006). Safety and usability of speech interfaces for in-vehicle tasks while driving: a brief literature review. Ann Arbor, MI: The University of Michigan Transportation Research Institute (UMTRI).

Burns, P., Harbluuk, J., Foley, J., & Angell, L. (2010). The importance of task duration and related measures in assessing the distraction potential of in-vehicle tasks. Proceedings of the Second International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2010), November 11-12, 2010, Pittsburgh, PA, USA.

Carter, C., & Graham, R. (2000). Experimental comparison of manual and voice controls for the operation of in-vehicle systems. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 3-286-283-289.

Chiang, D. P., Brooks, A. M., & Weir, D. H. (2005). Comparison of visual-manual and voice interaction with contemporary navigation system HMIs. SAE Technical Paper 2005-01-0433.

Driver Focus-Telematics Working Group (2006). Statement of principles, criteria and verification procedures on driver interactions with advanced in-vehicle information and communication systems, Version 2.0: Alliance of Automotive Manufacturers.

Fitch, G. A. Soccolich, S.A., Guo, F., McClafferty, J., Fang, Y., Olson, R. L., Perez, M. A., et al. (2013). The impact of hand-held and hands-free cell phone use on driving performance and safety-critical event risk (Report No. DOT HS 811 757). Washington, DC: National Highway Traffic Safety Administration (NHTSA).

Harbluk, J., Burns, P. C., Lochner, M., & Trbovich, P. L. (2007). Using the lane-change test (LCT) to assess distraction: Tests of visual-manual and speech-based operation of navigation system interfaces. Proceedings of the 4th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Stevenson, WA.

Lo, V.E-W., & Green, P.A. (2013). Development and evaluation of automotive speech interfaces: useful information from the human factors and the related literature. *International Journal of Vehicular Technology, 2013*, ID 924170. http://dx.doi.org/10.1155/2013/924170

Maciej, J., & Vollrath, M. (2009). Comparison of manual vs. speech-based interaction with in-vehicle information systems. *Accident Analysis & Prevention, 41(5),* 924-930.

Mehler, B., & Reimer, B. (2013). An initial assessment of the significance of task pacing on self-report and physiological measures of workload while driving. *Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, Bolton Landing, New York, June 18-19, 2013, pp. 170-176.

Mehler, B., Reimer, B., & Coughlin, J. F. (2012). Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: an on-road study across three age groups. *Human Factors, 54(3),* 396-412.

National Highway Traffic Safety Administration (2013). 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), .

Neurauter, M. L., Hankey, J. M., Schalk, T. B., & Wallace, G. (2009). Outbound texting: comparison of speech-based approach and handheld touch-screen equivalent. *Transportation Research Record: Journal of the Transportation Research Board, 2321,* 23-30.

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

Owens, J. M., McLaughlin, S. B., & Sudweeks, J. (2010). On-road comparison of driving performance measures when using handheld and voice-control interfaces for mobile phones and portable music players. *SAE International Journal of Passenger Cars – Mechanical Systems, 3(1),* 734-743.

Ranney, T. A., Baldwin, G. H., Parmer, E., Domeyer, J., Martin, J., & Mazzae, E. N. (2011). Developing a test to measure distraction potential of in-vehicle information system tasks in production vehicles (No. HS-811 463). National Highway Traffic Safety Administration.

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., & Coughlin, J. F. (2010). An evaluation of driver reactions to new vehicle parking assist technologies developed to reduce driver stress (MIT AgeLab White Paper). Cambridge, MA: Massachusetts Institute of Technology.

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. *(Note: due to size considerations, .pdf versions of the report may appear as two files, a main report and an appendix.)*

Reimer, B., Mehler, B., McAnulty, H., Munger, D., Mehler, A., Perez, E. A. G., et al. (2013). A preliminary assessment of perceived and objectively scaled workload of a voice-based driver interface. Proceedings of the Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Bolton Landing, NY, 537-543.

Shutko, J., Mayer, K., Laansoo, E., & Tijerina, L. (2009). Driver workload effects of cell phone, music player, and text messaging tasks with the Ford SYNC voice interface versus handheld visual-manual interfaces SAE Technical Paper 2009-01-0786. doi: 10.4271/2009-01-0786

Shutko, J., & Tijerina, L. (2011). Ford's approach to managing driver attention: SYNC and MyFord Touch. *Ergonomics in Design, 19(4),* 13-16.

ABOUT THE AUTHORS

Bryan Reimer, Ph.D.

Bryan Reimer is a Research Engineer in the Massachusetts Institute of Technology AgeLab and the Associate Director of the New England University Transportation Center. His research seeks to develop new models and methodologies to measure and understand human behavior in dynamic environments utilizing physiological signals, visual behavior monitoring, and overall performance measures. Dr. Reimer leads a multidisciplinary team of researchers and students focused on understanding how drivers respond to the increasing complexity of the operating environment and on finding solutions to the next generation of human factors challenges associated with distracted driving, automation and other in-vehicle technologies. He directs work focused on how drivers across the lifespan are affected by in-vehicle interfaces, safety systems, portable technologies, different types and levels of cognitive load. Dr. Reimer is an author on over 70 peer reviewed journal and conference papers in transportation. Dr. Reimer is a graduate of the University of Rhode Island with a Ph.D. in Industrial and Manufacturing Engineering.

reimer@mit.edu
(617) 452-2177
http://web.mit.edu/reimer/www/

Bruce Mehler, M.A.

Bruce Mehler is a Research Scientist in the Massachusetts Institute of Technology AgeLab and the New England University Transportation Center, and is the former Director of Applications & Development at NeuroDyne Medical Corporation. He has an extensive background in the development and application of non-invasive physiological monitoring technologies and research interests in workload assessment, individual differences in response to cognitive demand and stress in applied environments, and in how individuals adapt to new technologies. Mr. Mehler is an author of numerous peer reviewed journal and conference papers in the biobehavioral and transportation literature. He continues to maintain an interest in health status and behavior from his early work in behavioral medicine. He received an MA in Psychology from Boston University and a BS degree from the University of Washington.

bmehler@mit.edu
(617) 253-3534
http://agelab.mit.edu/bruce-mehler

About the New England University Transportation Center & MIT Center for Transportation & Logistics

The New England University Transportation Center is a research, education and technology transfer program sponsored by the US Department of Transportation. Together the faculty, researchers and students sponsored by the New England Center conduct work in partnership with industry, state & local governments, foundations and other stakeholders to address the future transportation challenges of aging, new technologies and environmental change on the nation's transportation system. For more information about the New England University Transportation Center, visit utc.mit.edu. For more information about the US Department of Transportation's University Transportation Centers Program, please visit www.rita.dot.gov/utc/. The New England Center is based within MIT's Center for Transportation & Logistics, a world leader in supply chain management education and research. CTL has made significant contributions to transportation and supply chain logistics and helped numerous companies gain competitive advantage from its cutting edge research. For more information on CTL, visit ctl.mit.edu.

About the AgeLab

The Massachusetts Institute of Technology AgeLab conducts research in human behavior and technology to develop new ideas to improve the quality of life of older people. Based within MIT's Engineering Systems Division and Center for Transportation & Logistics, the AgeLab has assembled a multidisciplinary team of researchers, as well as government and industry partners, to develop innovations that will invent how we will live, work and play tomorrow. For more information about AgeLab, visit agelab.mit.edu.