The Impact of Repeated Exposure to a Multi-Level Working Memory Task on Physiological Arousal and Driving Performance

Daniel Belyusar
MIT AgeLab
77 Massachusetts Avenue, E40-279 Cambridge, MA 02139
Tel: 617-452-2177; Email: belyusar@mit.edu

Bruce Mehler, Corresponding Author
MIT AgeLab & New England University Transportation Center
77 Massachusetts Avenue, E40-279 Cambridge, MA 02139
Tel: 617-253-3534; Email: bmehler@mit.edu

Erin Solovey
The College of Computing & Informatics, Drexel University
3141 Chestnut St., Philadelphia, PA 19104
Tel: 215-571-4598; Email: erin.solovey@drexel.edu

Bryan Reimer
MIT AgeLab & New England University Transportation Center
77 Massachusetts Avenue, E40-279 Cambridge, MA 02139
Tel: 617-452-2177; Email: reimer@mit.edu

Word count: 6060+ (250 * 3) (1 table & 2 figures) = 6810

TRR Paper number: 15-4605

Original Submission date: August 1, 2014
Revised: November 15, 2014
Revised based on second review: March, 15, 2015

ABSTRACT
Drivers are increasingly confronted with a multitude of simultaneous secondary tasks while behind the wheel. Consequently, the ability to manage this elevated workload is a rising concern. A growing body of evidence has suggested changes in physiology such as heart rate and skin conductance levels, along with driving metrics such as velocity and steering wheel reversal rates, could allow a real-time monitoring of this workload. However, laboratory research often does not evaluate drivers' experience of secondary tasks over multiple repetitions or time. The present study examined the sensitivity of a variety of performance and physiological measures to changes in workload and assessed the impact of repeated exposure to a set of n-back calibration tasks. When averaged across repeated exposures, the results closely follow previous literature, showing that increased working memory load from the n-back tasks resulted in significant increases in heart rate and skin conductance level. Interestingly, though each repetition of the task showed similar changes as workload increased (e.g. increased heart rate during more difficult tasks), this effect was consistently higher for the initial vs. the second and third exposures to each level of the task. Furthermore, some measures (heart rate, skin conductance) were more reliable over time than others (steering wheel reversals, velocity). This may be in part related to the emotional arousal of being asked to do a task for the first time. Implications are discussed.

INTRODUCTION

Drivers are increasingly required to negotiate a multitude of simultaneous tasks that draw upon visual, manipulative and cognitive resources, particularly with the proliferation of driver vehicle interfaces (DVIs) for controlling communication, navigation, and entertainment functions. Performing even simple tasks in dynamic environments, such as those encountered while driving, necessitates the fluid management of workload and arousal to maintain adequate performance without falling into states of under- or overload *(1)*. Since physiological measures can be used to capture continuous changes in driver state, they hold the potential to be used to report graduated levels of cognitive workload and detect changes in a user's stress levels in changing environments *(2)*. Specifically, there is growing evidence that changes in heart rate as well as other physiological measures such as electrodermal activity (skin conductance / GSR), can detect changes in cognitive load when drivers are engaged in secondary tasks *(3–7)*. Thus, there has been great interest in the use of physiological measures to dynamically monitor and manage in-vehicle systems *(1, 4, 8–10)*.

A solid body of literature has shown peripheral physiology measures to be more sensitive to changes in scaled levels of workload from experimental reference or "calibration" tasks than performance measures such as vehicle metrics or task accuracy *(3, 6, 11)*; nonetheless, significant changes in vehicle performance metrics have also been reported in a number of instances and sensitivity may vary depending on the type of task *(4, 5, 12)*. Recent research by Reimer and colleagues *(13)* suggests there are a number of complexities that need to be considered if physiological measures are used as indicators of cognitive demand in a real vehicle. It was observed that measures of arousal associated with a number of real interface tasks did not exceed that imposed by a moderately demanding working memory calibration task (1-back). They suggested that some drivers actively compensate for the added DVI task demand by shedding workload associated with driving, i.e. changing lanes less often, driving slower, etc., and utilizing visual support offered by some interfaces to reduce memory demands. It was also found that repeated exposure to tasks produced a modest decrease in cardiac activity, while more mixed changes in electrodermal activity and visual demand were observed (see appendix C of *(14)*). These findings indicate potential confounds in interpreting experimental driving research, with the latter highlighting a possible issue for a limited number of task exposures during the training period as well as during the experiment.

In reality, driving, and increasingly the secondary tasks we engage in while driving, such as operating a navigation interface, can be relatively complex and may require multiple interactions for a driver to become proficient and comfortable with the particular task, i.e. have a 'learning curve'. It is unclear as to the degree in which studies that focus primarily on more "novice" use cases adequately reflect the demands of more experienced operations. Sometimes provisions to support such learning curves are built into a DVI; for example, incorporating 'novice' and 'expert' modes *(15)*, or allowing 'one-shot' entry when the user has become familiar with command sequences. However, the time course needed to fully become an expert on a given technology often far exceeds the time any given participant is able to reasonably spend in an experiment. So, although the effects of repeated exposures to complex tasks are clearly relevant to understanding how we interact with technologies, practical considerations often limit the extent to which this factor is explored.

Though limited, there is previous research suggesting that the effects of secondary tasks while driving may diminish over repeated exposures. For instance, a 2004 NHTSA study *(16)*

evaluated cell phone tasks that were either artificial (e.g. solving math problems) or conversational while driving in a simulator. The authors found the negative effects of the secondary tasks on some driving performance variables, such as increased steering wheel reversals, lessened over time. Chishom, Caird and Lockhart *(27)* examined interaction with an iPod for song selection over six experimental simulator sessions and found that driving performance while operating the device improved over the sessions. Although slowed responses to driving hazards declined somewhat during the dual task interactions, there was still a decrement relative to baseline driving. Cooper and colleagues *(28)* looked at the impact of receiving and responding to auditory messages during simulated driving. When considering the results across three driving tasks, they reported no substantive evidence of learning adjustments to the task across exposures.

Previous work in our lab has shown that some physiological variables such as changes in pupil diameter and horizontal scanning, were reduced upon a second exposure to an n-back cognitive task paradigm during simulated driving *(17)*. However, since pupil dilation can also be affected by fatigue *(18)*, and memory *(19, 20)*, it remains unclear if these effects are the result of an actual decrease in workload, indicative of learning, or a simple decrease in novelty-effects with additional exposure. In order to further examine the effects of repeated exposure on the physiological and vehicle control dynamics associated with cognitive workload *(2–4, 12)*, we extended on these previous studies by utilizing multiple presentations of the n-back calibration task.

METHODS

Participants

Thirty participants were recruited from the greater Boston area. All were considered active drivers, reporting having been licensed for a minimum of three years, driving 3 or more times per week, and having a driving record free from any police-reported accidents for the past year. Further exclusion criteria included self-reported health conditions and medication usage that could adversely affect driving. Compensation of $30 was provided for participation. The MIT institutional review board approved recruitment and experimental procedures.

Apparatus

Data collection was carried out in the MIT AgeLab driving simulator that is built around a fixed base, full cab 2001 Volkswagen New Beetle. An 8' by 6' (2.44m by 1.83m) projection screen was positioned 76" (1.93m) in front of the mid-point of the windshield and provided approximately a 40-degree view of the virtual world at a resolution of 1024 x 768 pixels. Graphical updates were generated at a minimum frame rate of 20 Hz using STISIM Drive version 2.08.02 (Systems Technology, Inc., Hawthorne, CA) based upon a driver's interaction with the steering wheel, brake and accelerator. Force feedback was provided through the steering wheel and auditory feedback consisting of engine noise, cornering, and braking sounds was provided through the vehicle's sound system. Instructions and audio tasks were pre-recorded and presented through the vehicle sound system. Vehicle dynamics such as velocity and steering dynamics were recorded at 10 Hz.

Physiological data was obtained from a MEDAC System/3 instrumentation unit (NeuroDyne Medical Corporation). A modified lead II configuration was employed for electrocardiograph (ECG) recording in which the negative lead was placed just under the right

clavicle (collar bone), ground just under the left clavicle, and the positive lead on the left side over the lower rib. The skin was cleaned with isopropyl alcohol and pre-gelled silver/silver chloride disposable electrodes (Vermed A10005, 7% chloride wet gel) were applied. Skin conductance was measured utilizing a constant current configuration and non-polarizing, low impedance gold plated electrodes that allowed electrodermal recording without the use of conductive gel. Sensors were placed on the underside of the outer flange of the middle fingers of the non-dominant hand and held in place with medical grade paper tape. Physiological data was recorded at 250 Hz.

Simulation

The simulation consisted of a divided highway with two lanes in each direction plus a 2-foot (0.61 m) shoulder on each side of the roadway. Lane width was 15 feet (3.62 m) and posted speed limit was 65 mph (104.6 km/h). Typical traffic events on the virtual highway included passing vehicles, lane changes, and slow downs. The average traffic density in the virtual scenario was set at 23 vehicles/mile (14.3/km). Average traffic speed for vehicles in the left lane was set equal to the posted speed limit of 65 mph (104.6 km/h) and 5 mph slower (96.5 km/h) for the right lane.

Procedure

After preliminary screening and consent, participants were trained to complete an audio presentation / verbal response n-back task while outside of the simulator (see *(21)* for details). Depending on the demand level of the task, participants were instructed to listen to a list of single-digit numbers and after each number either do nothing (blank-back), repeat out loud the number just presented (0-back), the previous number (1-back), or the number presented two numbers back in the sequence (2-back). See Table 1 for an example stimulus set and the way a participant would be expected to respond for each form (demand level) of the task.

Each n-back set consisted of 10 single digit numbers (0-9), with each number presented once in a random order. The interval between number presentations was 2.25 seconds. The resulting stimulus sets were approximately 30 seconds in duration. Participants practiced until they were able to complete each level with at least 50% accuracy (i.e. no more than four errors on the 2-back). Participants were given between one and four repetitions to learn each level of the n-back task, depending on performance.

After training, participants were offered a short break and were then fitted with ECG and skin conductance sensors. They then moved to the simulator, adjusted the seat and steering wheel as needed, and were exposed to a brief 2.5 mile (4.26km) drive (approximately 3.5 to 4.5 minutes) to develop basic familiarity with the simulator. After ‘stopping the vehicle’, participants practiced one additional set of each level of the n-back while seated in the simulator. Participants then resumed driving in a simulated highway environment, accelerating to highway speed at their own pace. Upon reaching 55 mph, a timer was started and an additional 3.5 minutes of driving without a secondary task took place to allow for further acclimation to the simulation and controls. Participants then were exposed to three assessment blocks where they completed the secondary cognitive tasks while driving. Each block contained one complete set of each level of n-back task (blank, 0-, 1-, and 2-back) presented in random order and also included one floating reference in which the participant was asked to simply continue driving for a period of time equivalent to an n-back set. Each subtask level was separated by a 90-second recovery period to

allow physiological signals to recover. Thus, each complete block was approximately 10 minutes in duration and the total post-adaptation assessment period was approximately 30 minutes.

Data processing / analysis

For purposes of obtaining inter-beat intervals and calculation of instantaneous heart rate, the locations of R-wave peaks (a prominent feature that occurs once per normal cardiac cycle) were identified in the raw ECG signal using analysis software developed at the MIT AgeLab. Using this software, trained research staff executed automated QRS complex detection algorithms (using either a modified version of the Pan-Tompkins algorithm, Wavelet decomposition, or filter banks algorithm) on each participant's raw ECG. Detection results were visually reviewed and misidentified or irregular intervals manually corrected. Using another in-house developed processing package, high frequency noise was removed from the skin conductance signal following *(22)* and clearly identified motion artifacts manually edited where practical. Corrupted or otherwise invalid ECG and skin conductance signal segments that could not be recovered were identified as such and excluded from subsequent analyses.

For both skin conductance level (SCL) and heart rate, we analyzed and report here the mean physiological unit values, as opposed to normalized percent change scores. While both are valid approaches, the study utilizes a within-subject design and repeated measures tests which consider changes by condition relative to each participant's own values, hence the data is essentially normalized to each participant. Reporting data in physiological unit values allows for comparison of values across studies.

In addition to physiological measures, we also computed several vehicle metrics. Standard deviation of lane position (SDLP) is a measure of variation in lateral vehicle positioning. It is calculated as the standard deviation of the distance in feet from the center line. Minor steering wheel reversals (SWR) were calculated using the values established in the European AIDE project as the sum of the number of reversals larger than 0.1 degree of angle occurring over each segment of interest *(23)*. Since every segment is of identical length (i.e. 10 trials, or 30 seconds) the calculation of raw values as opposed to rates is also valid. (Minor SWRs were considered as they have been observed to increase in response to auditory-cognitive tasks while major SWRs may increase during visual-manual tasks *(5)*). We also evaluated velocity in miles-per-hour and the standard deviation of velocity (SDV).

As physiological data tends to be non-normally distributed, preliminary tests of sphericity were run and found significant violations in most measures (HR- block and task, SCL- block and task, SDLP- block and task, SDV- task only; all Mauchly's Tests $p<.001$). For consistency we thus utilized Friedman's test which is a non-parametric repeated-measures test, for omnibus comparisons, and Wilcoxan signed-rank test, a non-parametric version of the t-test to explore any potential differences between specific task levels where appropriate. The use of ranks to avoid the assumption of normality is described in *(24)*.

RESULTS

Analysis sample

As our primary interest centered on examining the impact of the different levels of the secondary demand task across repetitions, only participants who were clearly engaged at the defined level of each n-back task were included in the final analysis. Six participants were dropped for having one or more task periods where they gave an incorrect response on all items;

this is consistent with a participant misinterpreting the task, e.g. responding to a 1-back trial as if it was a 0-back. Two additional participants were excluded due to unavailability of usable heart rate and SCL data respectively. The maximum number of errors for any single n-back task period of the remaining participants ranged from 0 to 4. These exclusion criteria resulted in a final set of 22 participants; eight of whom were female. Age ranged from 20 to 33 years with a mean of 24.9 (SD 3.8).

Analysis by demand level (task)

As noted above, participants were excluded from the analysis if they responded to one or more task sets at the 100% error level. The remaining participants were 100% accurate on the 0-back task, 96.9% accurate on the 1-back and 95.2% on the 2-back.

Figure 1 shows the mean and standard error for each dependent measure, for each level of the n-back averaged across all three repetitions. Consistent with previous literature, statistical tests found that there was a significant effect of task level for heart rate (Freidman's Chi^2= 72.18, $p<.001$). As expected, these differences were driven by increases from blank-back (M=73.99, SE=2.6) to 0-back (M=78.15, SE=2.64) to 1-back (M=79.58, SE=7.64), to 2-back (M=84.02, SE=2.64). All planned pairwise comparisons by levels were significant ($p<.01$). Similarly, there were significant differences in SCL (Freidman's Chi^2=27.20, $p<.001$) that showed a general increase with demand level (blank-back (M=10.27, SE=.81) to 0-back (M=10.70, SE=.82) to 1-back (M=10.60, SE=.83) to 2-back (M= 11.10, SE=.83)), although the 0-back and 1-back levels did not differ significantly from each other ($p>.05$); all other signed-ranks tests by level did differ significantly ($p <.05$).

Considering vehicle performance, there was a significant main effect of demand level (n-back task) for standard deviation of lane position (SDLP) (Friedman's Chi^2= 29.27, $p<.001$), pairwise comparisons did reveal differences between blank-back (M=0.67, SE=0.08) and all other levels (0-back (M=.77, SE=.21); 1-back (M=.55, SE=.12); (2-back (M=.62, SE=.11)). To the extent that SDLP may decrease with increased cognitive load, this does not appear as a clear linear trend. Minor steering wheel reversals (SWR) were also showed a significant main effect of demand (Freidman's Chi^2 = 13.53, $p<.001$). Signed ranks tests determined these differences to be driven by general increases in the number of reversals as the load increased from blank-back (M=11.14, SE=.65) to 0-back (M=13.24, SE=.71) to 1-back (M=13.73, SE=.72) to 2-back (M=13.77, SE=1.00); all conditions compared to blank-back were significantly different ($p<.05$). Velocity and standard deviation of velocity (SDV), did not differ significantly across the task levels (both Friedman Chi^2 test, $p>.05$).

Analysis across blocks

Our primary concern in this report was in the comparison of dependent measures over repeated presentations of the secondary task. We thus tested each variable in a similar non-parametric repeated measures test to detect differences across blocks (repetitions).

As can be seen in Figure 2, there was an overall main effect of block for heart rate (Friedman's Chi2 = 15.64, $p<.001$), most clearly characterized by a drop in heart rate between the first and second repetitions in the 0-, 1- and 2-back conditions. Differences in vehicle based measures between blocks, across demand levels, appear generally quite variable, and no other comparisons for main effects by block were statistically significant for these metrics or SCL (all comparisons $p>.05$). Nonetheless, it can be observed that there is a similarity in patterning in both heart rate and SCL across all active conditions (i.e. 0-, 1- and 2-back), where there were

either significant or nominal drops from the first to the second repetitions. In HR, the effects of levels remain statistically significant across blocks for all but 0-back to 1-back. That is, each level of the n-back is significantly different than nearly every other, for every block (all $p<.05$). Characterizing the block wise shifts in heart rate in an alternate manner, across the demand levels mean heart rate for block 1 dropped significantly from 81.32 (SE=2.75) to 78.08 (SE=2.6) ($p<.01$). While block 1 and block 3 (M=77.40, SE=2.5, $p<.01$) differ, the decline from block 2 to block 3 was moderate and not statistically significant.

While SCL does show some significant differences in each block, they are somewhat inconsistent. In all three blocks, blank-back compared to 2-back was significantly different, and 0-back to 2-back was significant in blocks 2 and 3. Blank-back compared to 1-back was significant in the third block, and marginal in blocks 1 ($p=.07$) and block 2 ($p=.06$). Subsequently, the decline in heart rate by repetition appears to largely stabilize, while SCL increases at least nominally across all levels. This apparent divergence in the SCL data will be considered further in the Discussion. The vehicle metrics describe even less consistent patterns. SDLP can distinguish the blank-back from the 1-back across all three blocks; however, in observing the patterns in Figure 2, this is not necessarily in the same direction. That is, in Block 1, the blank-back has a lower value of SDLP than the 1-back, but in the second repetition, 1-back is lower than blank-back.

DISCUSSION

The present study examined the sensitivity of a variety of performance and physiological measures to changes in workload and assessed the impact of repeated exposure to the task on outcomes. When averaged across repeated exposures, the results closely follow previous literature, showing that increased working memory load from an n-back task resulted in significant increases in heart rate and skin conductance level *(2, 4, 13, 22)*. The present data shows the most consistent linear relationship for heart rate, where each increasing workload level is associated with a statistically significant increase from the level below it. This driving simulation data is consistent with results obtained under actual highway driving conditions *(4)*. SCL shows a significant increase between the "just listening to numbers" task (blank-back) and the 0-back and between the 1- and 2-back levels. The lack of a significant SCL difference between the 0- and 1-back in this sample is somewhat anomalous, as differentiation between these levels has more typically been observed in larger data sets *(2, 4, 13)*.

While vehicle control metrics sometimes show differences between "baseline" driving (driving without a secondary task) and driving with added secondary cognitive load *(5, 12)*, primary task metrics like minor steering wheel reversals do not appear as sensitive as heart rate and SCL in discriminating fine gradations in demand in response to the working memory task. In this dataset, minor steering wheel reversals during the 0-, 1-, and 2-back levels were significantly different from the blank-back level but not from each other. A similar result for this metric was observed under real driving conditions *(12)*. Other metrics, as seen in the standard deviation of lane position, show changes across the demand range, but not in a unitary direction. In this dataset, SDLP can be seen to increase between the blank-back and 0-back conditions but then decrease during the 1-back. Such non-linear variation in SDLP has been observed before in this simulator across multiple levels of the n-back task *(2)*. Velocity and standard deviation of velocity show no discrimination across the cognitive task demand levels in this dataset.

In discussing the relationship between physiological measures of workload and driving performance metrics, Mehler and colleagues *(2, 4)* and Lenneman and Backs *(25)* have

suggested previously that there are sound theoretical reasons to expect that these classes of measures may show sensitivity at different points along a demand profile. To the extent that an operator has spare resources to invest in a secondary task, primary operational control may not be noticeably impacted as demand increases, but physiological arousal can provide an indicator of increased investment of effort. Consistent decrements in overt vehicle control may not become apparent until cognitive or other relevant resources begin to become saturated. Wilson *(11)* has made this same point in discussing the measurement of operator workload in other domains, such as aviation.

Perhaps most interesting in this dataset is an examination of the extent to which different measures show consistency across multiple exposures to the same task. Mean heart rate, SCL and SDLP are highly consistent for the blank-back across the 3 blocks, suggesting that the state of driving the simulator and just listening to numbers was in fact relatively consistent across the repetitions. In contrast, the variability seen in mean velocity, standard deviation of velocity, and minor steering wheel reversals across the blank-back periods indicates that these may be less stable reference points for a "base" driving state, at least under these simulated driving conditions. Heart rate shows a highly consistent pattern across the 3 repetitions of the 0-, 1-, and 2-back demand levels. In each, mean heart rate drops between the first and second repetitions and then remains nominally in the same range during the third repetition of the 0- and 1-back, with a modest further drop at the 2-back level. We will return to the apparent higher level of arousal indicated in heart rate during participants' first exposure to each level of the active memory task in a moment.

SCL shows some similar characteristics across the repetitions and levels, but without the same degree of consistency. There is the same trend for mean SCL to drop modestly between the first and second repetitions at each level; however, there is also a clear increase in mean SCL during the third repetition of the 1-back task and a nominal trend suggesting a modest increase during the blank-back and 0- and 2-back. This may simply be reflective of greater variability in SCL as a measure, or might alternately indicate a degree of fatigue and frustration during the third block as participants begin to find continuation of the same set of tasks somewhat monotonous and begin to look forward to the completion of the experiment. One of the strengths of SCL is its sensitivity to these types of emotional states. Some consistency is suggested in SDLP across repetitions for the 0- and 1-back levels, but this clearly not present for the 2-back. As might logically be expected given the control connection between steering wheel positioning and lane positioning, SWR shows somewhat similar consistency in the patterning across these levels and repetitions.

Several points can be made about the clear decline in heart rate between the first and second repetitions of each of the working memory task demand levels (0-, 1- and 2-back). The first concerns the consistency of the patterning across levels by repetition that has already been summarized. This indicates that each of these samples is providing a highly repeatable measure of the participant experience such that the relative difference between active task levels is comparable over repetitions. Second, arousal or effort as indicated by heart rate is consistently higher for the first vs. the second and third exposures to each level of the task. This may be in part related to the emotional arousal of being asked to do a task for the first time while also driving the simulator. The trend for SCL to drop nominally between the first and second repetitions provides some concurrent support for this interpretation. It is also possible that the effective workload for each task becomes somewhat less as the participant becomes familiar with and potentially develops a more refined strategy for managing both tasks simultaneously. The

apparent leveling out, or near leveling out of heart rate, by the third repetition suggests that this settling of apparent workload may occur fairly rapidly. Though the rate of change from the second to the third repetition does decline, it is not established in this dataset if this would ever completely plateau, or how many exposures would be needed to reduce the change to insignificant levels. In this regard, data developed concurrent with this study under actual on-road driving conditions *(9)* showed that drivers presented with 24 2-back task periods continued to show significant heart rate elevations relative to baseline driving across the 24 repetitions.

The data presented here continue to support a role for both driver behavior based metrics (vehicle metrics) and physiology in the study of the impact of cognitive workload, and provide further evidence for potentially greater sensitivity of physiological measure for fine grain discrimination of levels of demand up to the 2-back level for relatively non-challenging simulated highway driving conditions studied. Moreover, heart rate proved the most reliable discriminator across gradations over repeated measures. In addition, these data point to some methodological and interpretive issues to be considered in assessing task associated workload.

Not surprisingly, the findings highlight the role of practice / experience in evaluating the workload associated with a task. As described in the Methods section, participants in this study were trained on the n-back both outside and inside the simulator to ensure a high degree of task competency. They also had an acclimation drive in the simulator prior to secondary task evaluation. Furthermore, participants who exhibited apparent confusion as to what task level they were to perform were removed as outliers for the purpose of this analysis. Nevertheless, heart rate reactivity to the task declined, most notably for the relatively difficult 2-back task, over the course of three exposures while driving the simulator. This would suggest that the experience of doing the task for the first time under driving conditions was more demanding than subsequent repetitions of the task. This strongly argues for presenting a participant with more than one presentation of any given task type, which many study designs already do, unless the goal is specifically to evaluate the workload associated with undertaking a novel task. In the newest guidelines for research on driver distraction, NHTSA has recommended that research participants 'practice as many times as needed until they think that they have become comfortable in performing the task' and that they practice each 'testable task' while driving a simulator (pg. 24887) *(26)*.

It can be noted that participants in this study were relatively young (20-33 years). We have found errors rates to be somewhat higher at the 1-back and 2-back levels in participants in their 40s and 60s *(2)*, thus the tasks might be experienced as somewhat more challenging on average as a function of age and the impact on driving does, in some instances, appear somewhat greater (13-14). Nonetheless, the overall pattern of physiological response in relatively healthy older drivers is similar to that of younger samples.

The results presented here suggest that while a participant may be comfortable and proficient at performing a task during training, measures of demand may continue to decline with repeated exposure during dual task (driving) conditions. While the results do not provide an answer to how much exposure to an activity may be optimal prior to demand assessment (and perhaps this may be task specific), these findings suggest a need for additional research and clarity on the topic of repeated exposure to interface demands. Furthermore, future research in this area should aim to provide a clear presentation of how participants are trained, and what was done to ensure competency. This information would appear necessary so that results can be adequately evaluated with respect to the level of novelty.

ACKNOWLEDGEMENTS
Support for this work was provided in part by the US DOT's Region I New England University Transportation Center at MIT and the Toyota Class Action Settlement Safety Research and Education Program. The views and conclusions being expressed are those of the authors, and have not been sponsored, approved, or endorsed by Toyota or plaintiffs' class counsel.

REFERENCES

1. J. F. Coughlin, B. Reimer, and B. Mehler. Monitoring, Managing, and Motivating Driver Safety and Well-Being. *IEEE Pervasive Computing*, vol. 10, no. 3, 2011, pp. 14–21.
2. B. Mehler, B. Reimer, J. F. Coughlin, and J. A. Dusek. Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers. In *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2138, no. -1, 2009, pp. 6–12.
3. B. Reimer, B. Mehler, J. F. Coughlin, K. M. Godfrey, and C. Tan. An On-Road Assessment of the Impact of Cognitive Workload on Physiological Arousal in Young Adult Drivers. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '09)*, Sept. 2009, pp. 115-118.
4. B. Mehler, B. Reimer, and J. F. Coughlin. 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*, vol. 54, no. 3, 2012, pp. 396–412.
5. J. Engström, E. Johansson, and J. Östlund. Effects of Visual and Cognitive Load in Real and Simulated Motorway Driving. *Transportation Research Part F Traffic Psychology & Behaviour*, vol. 8, no. 2, 2005, pp. 97–120.
6. J. K. Lenneman, J. R. Shelley, and R. W. Backs. Deciphering Psychological-Physiological Mappings While Driving and Performing a Secondary Memory Task. In *Proceedings of the 3rd International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 1995, pp. 493–498.
7. K. A. Brookhuis and D. de Waard. Assessment of Drivers' Workload: Performance And Subjective And Physiological Indexes. In *Stress, Workload, and Fatigue*, P. A. Hancock and P. A. Desmond, Eds. Mahwah, NJ: Larence Erlbaum, 2001, pp. 321–333.
8. J. A. Healey and R. W. Picard. Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, 2005, pp. 156–166.
9. E. T. Solovey, M. Zec, E. Abdon, G. Perez, B. Reimer, and B. Mehler. Classifying Driver Workload Using Physiological and Driving Performance Data : Two Field Studies. In *Proceedings of the 32nd Annual Conference on Human Factors in Computing Systems (CHI 2014),* Toronto, Canada, April 26 – May 1, 2014, pp. 4057-4066.
10. W. Hajek, I. Gaponova, K. H. Fleischer, and J. Krems. Workload-Adaptive Cruise Control – A New Generation of Advanced Driver Assistance Systems. *Transportation Research Part F Traffic Psychology & Behaviour* vol. 20, 2013, pp. 108–120.

11. G. F. Wilson,. An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures. *International Journal of Aviation Psycholology*, vol. 12, no. 1, 2002, pp. 3–18.
12. B. Reimer, B. Mehler, Y. Wang, and J. F. Coughlin. A Field Study on the Impact of Variations in Short-Term Memory Demands on Drivers' Visual Attention and Driving Performance Across Three Age Groups. *Human Factors*, vol. 54, no. 3, 2012, pp. 454–468.
13. B. Reimer, B. Mehler, J. Dobres, and J. F. Coughlin. The Effects of a Production Level 'Voice-Command' Interface on Driver Behavior: Reported Workload, Physiology, Visual Attention, and Driving Performance. MIT AgeLab Technical Report 2013-17A, Massachusetts Institute of Technology, Cambridge, MA,, 2013.
14. B. Reimer, B. Mehler, J. Dobres, and J. F. Coughlin. The Effects of a Production Level 'Voice-Command' Interface on Driver Behavior : Reported Workload, Physiology, Visual Attention, and Driving Performance (Appendix). MIT AgeLab Technical Report 2013-17A, Massachusetts Institute of Technology, Cambridge, MA, 2013.
15. B. Reimer, B. Mehler, J. Dobres, H. McAnulty, A. Mehler, D. Munger, and A. Rumpold. Effects of an 'Expert Mode' Voice Command System on Task Performance, Glance Behavior & Driver Physiology. In *Proceedings of the 2014 International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2014.
16. D. Shinar and N. Tractinsky. Effects of Practice on Interference From an Auditory Task While Driving : A Simulation Study. NHTSA Report No. HS-809 826. Washington D.C., 2004.
17. Y. Wang, B. Reimer, B. Mehler, J. Zhang, A. Mehler, and J. F. Coughlin. The Impact of Repeated Cognitive Tasks on Driving Performance and Visual Attention. In *Proceedings of the 3rd International Conference on Applied Human Factors and Ergonomics*, July 17-20, 2010, Miami, Florida.
18. R. E. Yoss, N. J. Moyer, and R. W. Hollenhorst. Pupil Size and Spontaneous Pupillary Waves Associated with Alertness, Drowsiness, and Sleep. *Neurology*, vol. 20, no. 6, 1970, pp. 545–54.
19. B. Heaver and S. B. Hutton. Keeping an Eye on the Truth? Pupil Size Changes Associated with Recognition Memory. *Memory*, vol. 19, no. 4, pp. 398–405, May 2011.
20. J. Beatty and B. Lucero-Wagoner. The Pupillary System. In *Handbook of Psychophysiology*, 2nd ed., J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, Eds. Cambridge Univ Press, 2000, pp. 142–162.
21. B. Mehler, B. Reimer, and J. A. Dusek. MIT AgeLab Delayed Digit Recall Task. MIT AgeLab Technical Report 2011-3B, Massachusetts Institute of Technology, Cambridge, MA, 2011.
22. B. Reimer and B. Mehler. The Impact of Cognitive Workload on Physiological Arousal in Young Adult Drivers: A Field Study and Simulation Validation. *Ergonomics*, vol. 54, no. 10, 2011, pp. 932-942.
23. J. Ostlund, B. Peters, B. Thorslund, J. Engström, G. Markkula, A. Keinath, D. Horst, S. Juch, S. Mattes, and U. Foehl, U. 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, 2005.

24. M. Friedman. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of American Statistical Assocociation*, vol. 32, no. 200, 1937, pp. 675–701.
25. J. K. Lenneman and R. W. Backs. Cardiac Autonomic Control During Simulated Driving With a Concurrent Verbal Working Memory Task. *Human Factors,* vol. 51, no. 3, 2009, pp. 404–418.
26. NHTSA. Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices. National Highway Transportation Safety Administration, Washington, D.C., 2013.
27. S. L. Chisholm, J. K. Caird, and J. Lockhart. The Effects of Practice with MP3 Players on Driving Performance. *Accident Analysis & Prevention*, vol. 40, no. 2, 2008, 704-713.
28. P. J. Cooper, Y. Zheng, C. Richard, J. Vavrik, B. Heinrichs,and G. Siegmun. The Impact of Hands-Free Message Reception/Response on Driving Task Performance. *Accident Analysis & Prevention*, vol. 35, no. 1, 2003, pp. 23-35.

LIST OF TABLES

LIST OF FIGURES

TABLE 1 Example N-Back Set with Expected Response for Each Task Type

Stimulus	5	7	0	9	8	4	3	1	2	6
Blank-back Response	-	-	-	-	-	-	-	-	-	-
0-back Response	5	7	0	9	8	4	3	1	2	6
1-back Response	-	5	7	0	9	8	4	3	1	2
2-back Response	-	-	5	7	0	9	8	4	3	1

TABLE 1: Example N-Back Set with Expected Response for Each Task Type

Stimulus	5	7	0	9	8	4	3	1	2	6
Blank-back Response	-	-	-	-	-	-	-	-	-	-
0-back Response	5	7	0	9	8	4	3	1	2	6
1-back Response	-	5	7	0	9	8	4	3	1	2
2-back Response	-	-	5	7	0	9	8	4	3	1

Note: dash = no response needed or possible

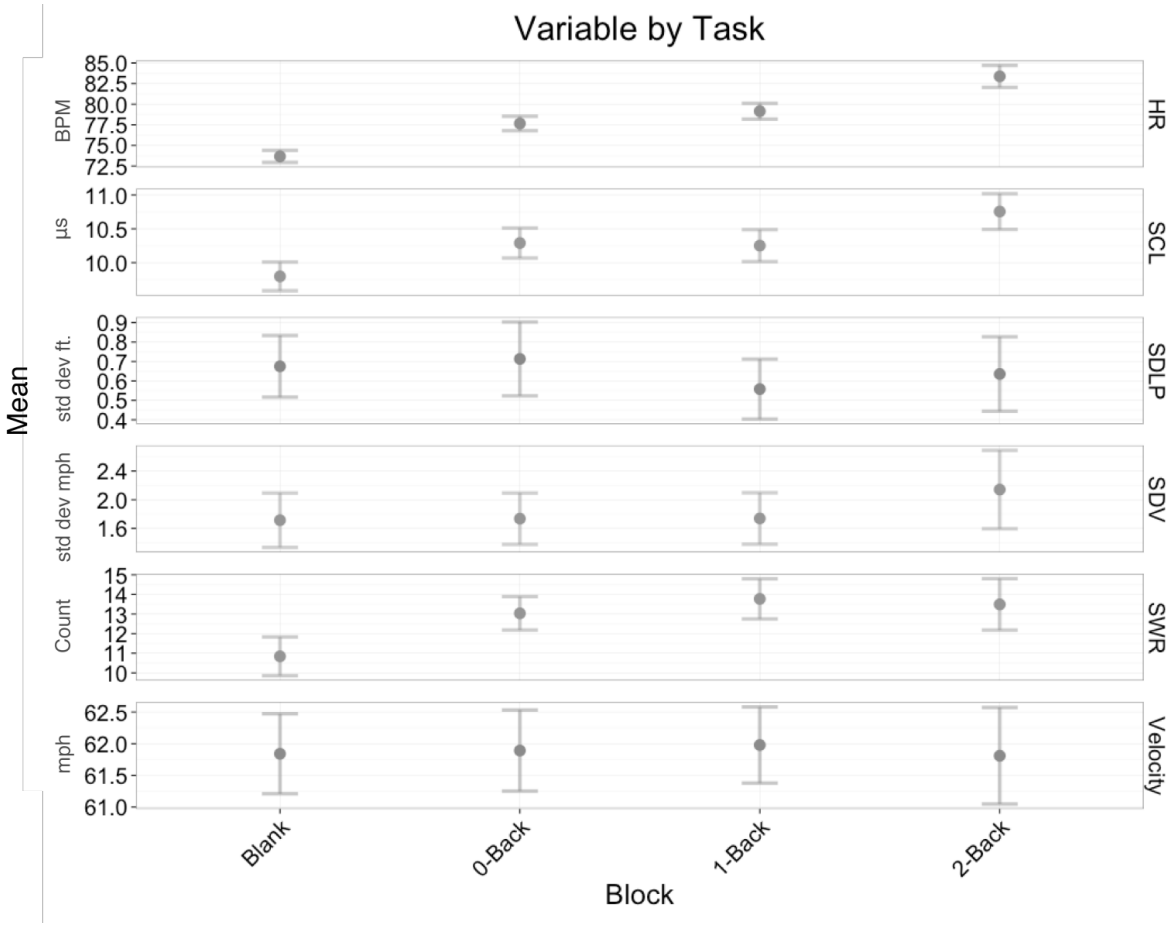

FIGURE 1 Composite means for each dependent measure across the three replications of each level of the n-back task. Error bars represent SEM. HR=Mean Heart rate; SCL=Skin Conductance Level, SDLP=Standard Deviation of Lane Position, SDV= Standard Deviation of Velocity, SWR=Steering Wheel Reversals.

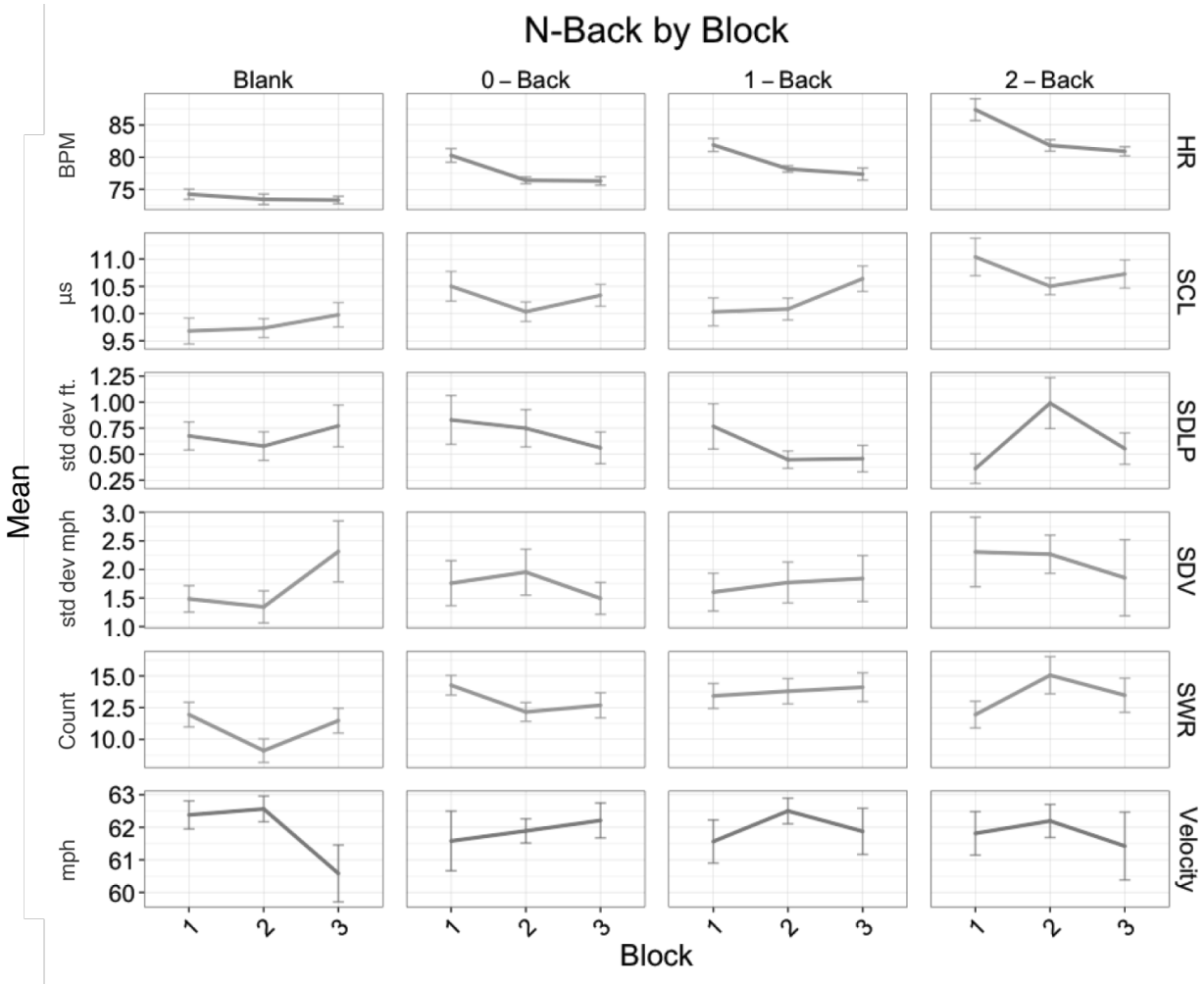

FIGURE 2 Means for each dependent measure by n-back difficulty level and repeated task blocks. Error bars represent SEM. HR=Mean Heart rate; SCL=Skin Conductance Level, SDLP=Standard Deviation of Lane Position, SDV= Standard Deviation of Velocity, SWR=Steering Wheel Reversals.