TITLE: Using Accident Location and interpretable Risk to fine tune Advanced Rider Assistance Systems for Motorcyclists

Andreas Hula
AIT Austrian Institute of Technology GmbH – Center for Low-Emission Transport
Giefinggasse 4, 1210 Vienna,
Austria

Klemens Schwieger
AIT Austrian Institute of Technology GmbH – Center for Low-Emission Transport
Giefinggasse 4, 1210 Vienna,
Austria

Peter Saleh
AIT Austrian Institute of Technology – Center for Low-Emission Transport
Giefinggasse 4, 1210 Vienna,
Austria

Paper Number: 23-0061

ABSTRACT

Motorcycle riding is a popular activity among riders of all ages and the number of motorcyclists is still increasing, despite safety issues being tricky to resolve for this mode of transport. Motorcycle rides constitute a type of vulnerable road user (VRU) since accidents tend to have more severe consequences for them due to the lack of physical protection for riders compared to passengers in passenger cars. Since this is a consequence of the very nature of the vehicle (being less heavy and more dynamic to move) potential safety interventions for motorcyclists need to be based on predictive indicators for unsafe situations and aim to avoid crashes altogether.

This paper presents the results of ongoing work to improve motorcycle safety by finding causally interpretable risk characteristics based on accident data and motorcycle riding dynamics collected from test rides by individual riders. Dynamics data at known accident spots and representative data for individual rider-typical motions is associated to the type of historical accident in order to produce an estimate not only of risky areas and maneuvers, but also to associate types of riding dynamics that put the driver at risk. The relation to potential causes is essential for the inclusion of the resulting risk warnings in the activation of an Advanced Rider-Assistance System (ARAS), in order to produce a tailor-made response to the individual.

INTRODUCTION

Though a popular mode of transport, the share of motorcyclist accidents and in particular fatalities is still remarkably high. In Austria in 2021 for instance, motorcyclists constituted 20.7% of all road user fatalities [1] with the absolute number of motorcyclist fatalities having remained roughly constant over several years, while the absolute numbers of fatalities for several other modes decreased.

Worldwide approaches to improve the safety of motorcyclists are being investigated. Approaches include physical methods of various sorts (see for instance [2]). Studies are also tackling the identification and use of accident hotspots for motorcycling safety research (see [3]). Historical accident data is one of the main guides to understanding motorcyclist safety (see [4]) and will play an important role in the approach presented here.

We present further developments of a method published recently (see [5],[6],[7],[8],[9]) for deriving risk maps based on gathered data during test rides. Vehicle dynamics data on 6 popular motorcycle tracks were collected by 5 experienced riders on a Motorcycle Probe Vehicle (dubbed “MoProVe”, see [10]), which was comprised of a motorcycle with accessible control area network (CAN)-Bus Data and several external sensor systems (additional geo-positioning systems, additional inertial measurement units), as well as a camera to produce video documentation on all rides.

Data was analysed by combining different approaches from machine learning/statistics: Individual riding data was clustered to determine typical motions of each driver, driving dynamics at known accident and finally, separation functions between accident prone dynamics and uncritical dynamics were derived. Going beyond earlier work, the resulting separation functions were now associated with interpretable risk warnings and potential interventions are listed.

Specific risk profiles can be associated to individual riders, by extension allowing to consider specific warnings to the ARAS and interventions during riding. Additionally, the dynamic variables contributing the most to a present warning give further indications on the cause of a risk warning and can be used in a similar manner, to adjust riding dynamics.

The use of an individual rider's profile derived from riding dynamics data and dynamics derived from known accident spots with associated causes paves the way for a much more specific response of ARAS systems, which might save lives without distracting drivers during critical moments.

MATERIALS AND METHODS

Data was collected using a KTM 1290 Super Adventure (provided by KTM [11], to support this research) equipped with several additional data collection systems (see VBOX [12]; Debus & Diebold [13]), to obtain high quality riding dynamics data (in particular angular movements: Yaw-Rates, Pitch-Rates and Roll-Rates) alongside a high-quality GPS localisation.

Measurements

We collected data on 6 different popular motorcycling tracks via 5 different test drivers. Test drivers were instructed to drive in different riding styles (conservative, comfortable, dynamic) to obtain a range of different driving behaviours from each rider. On a given track all riders rode several times (at least 3 times in either direction) so as not to fit our model to any particular ride but rather to more stable tendencies identifiable from several rides in multiple styles. Obtained data was checked for validity, excluding data errors and annotating time spans during which the motorcycle was following other vehicles.

Time based data was projected to a location based (per meter) grid by partners at TU Vienna in earlier work (see [8],[9]),

Accident data was obtained from Statistik Austria for the years 2012 to 2015 on the given tracks. We were interested exclusively in single vehicle motorcyclist accidents and collisions with oncoming traffic (as a proxy for narrow curves with potentially poor visibility).

Model

Our model is based upon per meter values of dynamics data (j yaw-, r roll- and p pitch-rates, as well as a measure of driven curvature, see [6] for details) which are used to fit a separation approach (see [14],[15]), based on known accident locations and k-means clustering (see [14],[16]) of dynamics data. We use a linear separation model (see [14],[17]) to separate cluster centers and dynamics data at known accident locations.

For a set of variables V, consisting of the yaw-rate j, roll-rate r and pitch-rate p (all in degrees per second) and a measure of driven curvature

$$\kappa = \frac{r}{|j|+0.1}, \tag{1}$$

we apply a number of transformations to allow for meaningful processing of the obtained data. Firstly, we apply a rollmean over 20 meters for each variable, to smooth the sensor data. Then we split given variables V into positive and negative parts $V_+ = max(V, 0)$ and $V_- = max(-V, 0)$ and calculate approximate derivates of the per meter values i.e., "first differences" of the obtained values $dV(m) = |V(m) - V(m-1)|$. The idea behind this preprocessing is to allow for separate weights in the statistical model on accelerations and decelerations, as well as left and right movements of the motorcycle.

$$md = (j_+, p_+, r_+, \kappa_+, j_-, p_-, r_-, \kappa_-, dj_+, dp_+, dr_+, d\kappa_+, dj_-, dp_-, dr_-, d\kappa_-) \tag{2}$$

We use k-means clustering on this data to define "standard motions" for each rider and use those as references for "non-risky" dynamics. Conversely, "risky" dynamics are simply defined as dynamics data at known historical accident locations. We use these two references to fit a separation model S based on a linear regression with target values S = 1 for risky dynamics and S = 0 for non-risky dynamics.

$$S = \sum_V a_V V_+ + \sum_V b_V V_- + \sum_V c_V dV_+ + \sum_V d_V dV_- + \varepsilon \tag{3}$$

In this model equation we have $a_V, b_V, c_V, d_V \in \mathbb{R}$ with ε denoting the error term of the regression. The coefficients in equation (3) can be used to assign values of $\mathsf{S}(md)$ for all considered driving dynamics data md

of the form in equation (2). A threshold separation between risky and non-risky dynamics data was defined for each rider and subjected to a joint optimization (see [5] for details).

We used the language R for our implementations [18].

The fit separation function can be used to assign weights dubbed "responsibilities" to all the components of a given data vector md using the contribution each component makes to the positive value of S(md) . If u denotes a single element of (a_V, b_V, c_V, d_V) and md_u denotes the corresponding dynamics variable value, then the responsibility U_{md} of md_u for the separation value S(md) is:

$$U_{md} = \frac{u\, md_u}{\sum_V (a_V V_+)_+ + \sum_V (b_V V_-)_+ + \sum_V (c_V dV_+)_+ + \sum_V (d_V dV_-)_+}. \quad (4)$$

In this sense responsibilities denote the share of a positive contribution of the component u to the value S(md) (responsibilities are nonnegative and sum up to 1).

RESULTS

Using the separation approach outlined above, the responsibilities for known accident sites were investigated. To illustrate, we show an example of the responsibilities for all riders combined into a single estimate in Fig. 1, for a right curve and a left curve accident:

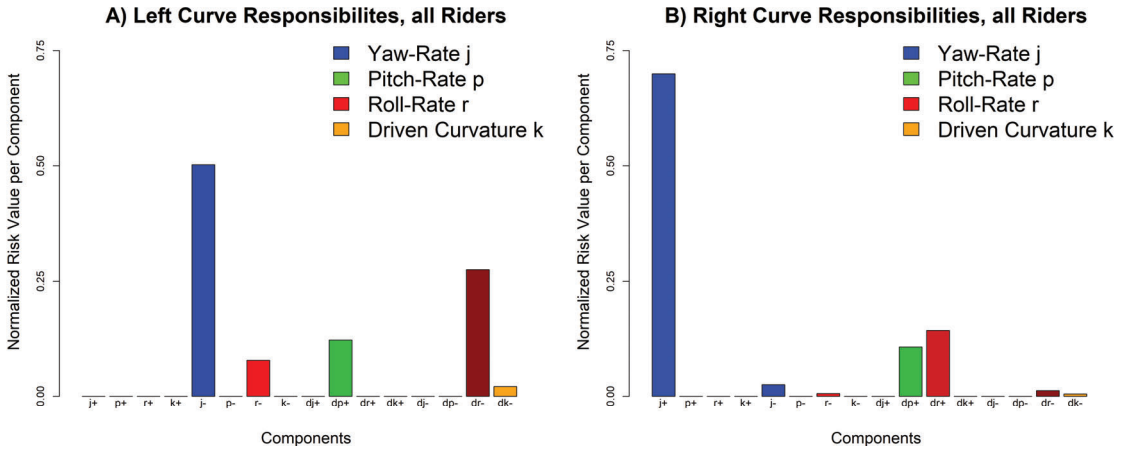

Figure 1: Responsibilities by data component for a) a Left curve with a known left curve accident and b) a Right Curve with a known right curve accident. Data while following other vehicles or having poor satellite connection has been removed.

It can be seen that the primary contribution in Fig. 1 a) stems from the negative yaw-rate j_- and in Fig. 1 b) from the positive yaw-rate j_+. Investigating the respective yaw-rate values for drivers at accident spots of the same type, we find results depicted in Fig. 2 below.

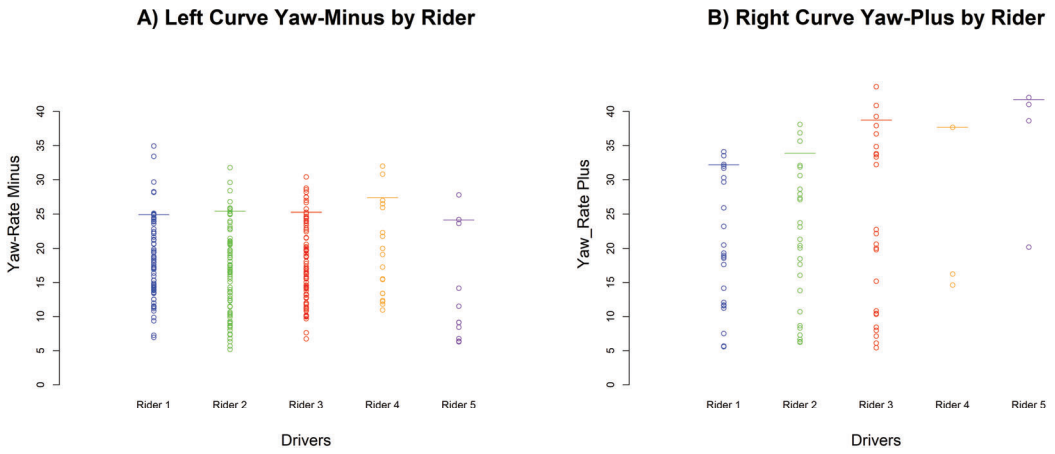

Figure 2: Values of the yaw rate at known accident locations with the 90% quantiles (signified by horizontal bars in the graph) of each rider being represented by a horizontal bar in the data. Panel a) depicts values of the yaw-rate minus at left curve accidents, while Panel b) depicts values of the yaw-rate plus at right curve accidents. Data while following other vehicles or having poor satellite connection has been removed.

This is a first opportunity to determine driver specific thresholds from the dynamics data obtained. We use the distribution shown in Fig. 2 to find a limit in the respective parameter for each rider and consider all values above to be challenging dynamics that might warrant ARAS Systems to prepare to engage. The thresholds can be seen in Fig. 2 as horizontal bars. Since we also note substantial contributions by the change of size in the roll-rate (dr_- and dr_+) and changes in the size of the pitch rate dp_+ to the risk responsibilities of left and right curve, we represent those in a similar manner in Fig. 3 and Fig. 4 below:

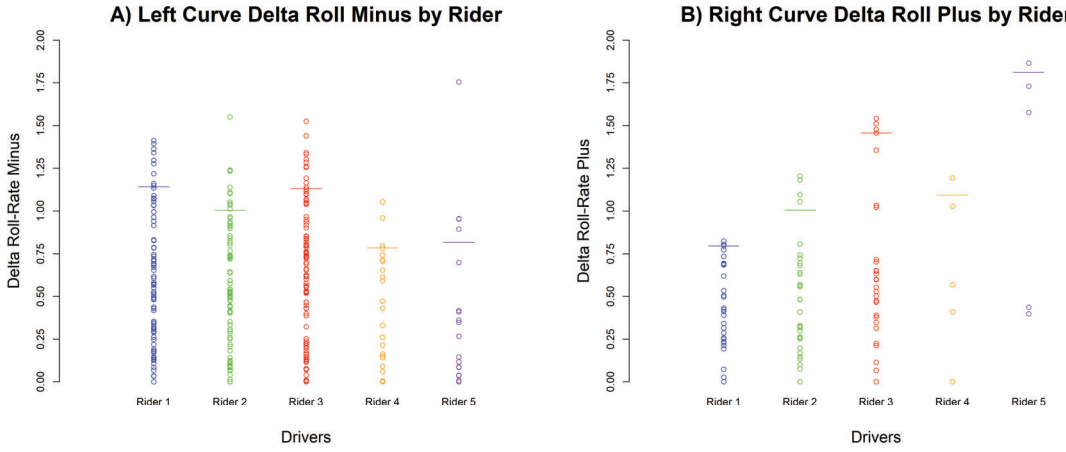

Figure 3: Values of the change in roll rate size at known accident locations with the 90% quantiles (signified by horizontal bars in the graph) of each rider being represented by a horizontal bar in the data. Panel a) depicts values of the delta roll-rate minus at left curve accidents, while Panel b) depicts values of the delta roll-rate plus at right curve accidents. Data while following other vehicles or having poor satellite connection has been removed

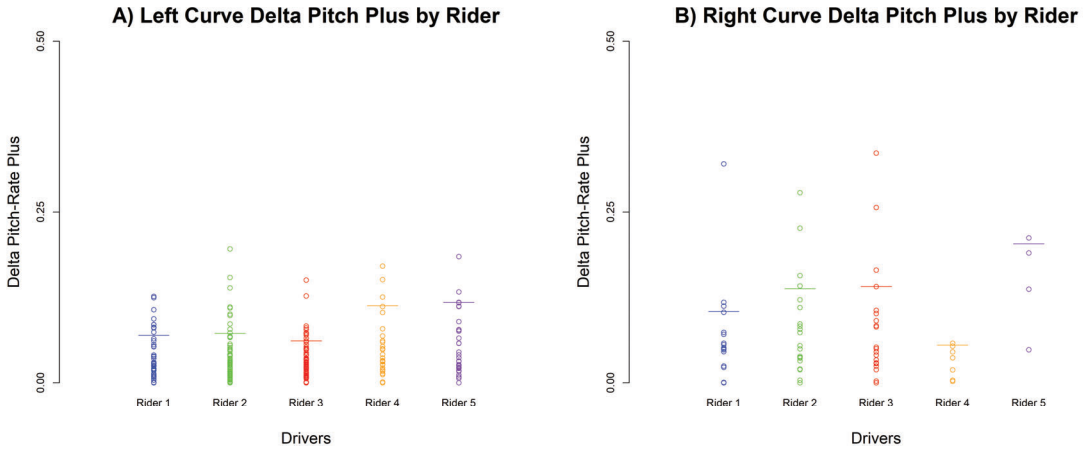

Figure 4: Values of the positive change in pitch rate size at known accident locations with the 90% quantiles (signified by horizontal bars in the graph) of each rider being represented by a horizontal bar in the data. Panel a) depicts values of the delta pitch-rate plus at left curve accidents, while Panel b) depicts values of the delta pitch-rate plus at right curve accidents. Data while following other vehicles or having poor satellite connection has been removed.

We chose thresholds according to quantiles of the respective data. A 90% Quantile is a good starting point to define the thresholds for the challenging domain in terms of driving dynamics. We see that with some variation the 90% quantiles in the yaw-rate are fairly close for our (experienced) test riders for the left curve, suggesting comparable limits to how riders would typically drive in a given curve. Interestingly estimates vary far more widely for the right curve. This may however be due in part to the lower number of accident locations for this type.

This procedure can be applied to obtain thresholds based on different accident locations and types and form specific boundaries on yaw-rate, roll-rate and pitch-rate to inform ARAS Systems such as traction control or ABS.

The thresholds derived in this reference data could then be transferred to drivers having similar driving profiles i.e., using the k-means clusters derived from the general driving data for a rider. The model which has the lowest squared distances in terms of components of the cluster centers md_i^{rc} of the reference data (denoted by rc) and components md_i^c of the not previously classified riders with dynamics data clusters c:

$$Dist = \min_{orders\ of\ c} \sum_c \sum_i (md_i^c - md_i^{rc})^2 \quad (5)$$

This can be used to transfer the threshold determined on the reference data here to riders in general. Thus, the reference high detail data set used in earlier work, can be used to find safety indicators for riders more generally, while the reference data set can continue to be expanded with more reference rider types on more tracks.

CONCLUSIONS

Building on a model of finding accident risk locations from driving dynamics data, we have investigated first approaches in associating particular accident sites with particular risk types. This allows us to derive thresholds for driving dynamics parameters for individual riders for particular accident types. These thresholds can be evaluated quickly even during the operation of a vehicle, thus making this a potential approach to guiding ARAS for motorcyclists. We have furthermore presented first consideration of how to transfer the models fit on precise reference data to more general settings, by using the cluster centers and finding the closest reference types in terms of these clusters.

The kinds of interventions that ARAS might implement based on this information is very much still up for debate. Speed recommendations or regulations could be a way forward, given the drivers experience (as evidenced by their quantile values of driving dynamics during various maneuvers) or, perhaps more challenging, stabilization features to still be developed.

Similar models could be fit to various levels of experience of the respective drivers and particular states (i.e., detecting fatigue from the driving style) and thus instantiate supportive measures in the same manner (reduce available power, recommend speed, stabilize motions).

Using the profiles of certain accident types, it appears feasible that those could also be used to classify accident spots with “unknown” accident causes and develop the methodology further towards accident reconstruction.

The most immediate limitations of these results stem from the size of the data set (5 riders on 6 tracks) and the precision of the available accident data. Quality checking available accident data will be necessary to expand the available data sets. The quality of the transfer of risk models from one rider to another will have to be demonstrated in future data. We note that a system based on these kinds of models might want to have an updating methodology and we have discussed first ideas of such a methodology in [19]. Alternatively, transferability might be sufficient to address this, as there could be clusters/profiles that encode various stages of driver experience.

Acknowledgements

We thank KTM Austria for providing the vehicle used as MoProVe in this study, as well as the Austrian Road Safety Fund (VSF [20]) for their funding of the underlying study.

We thank our partners at TU Vienna (Horst Ecker, Manfred Neumann) for annotating the data in the previous project.

References

[1] Statistik Austria 2021, https://www.statistik.at/en/statistics/tourism-and-transport/accidents/road-traffic-accidents

[2] Afquir, S., Melot, A., Ndiaye, A., Hammad, E., Martin, J.-L. and Arnoux, P.-J. 2020. Descriptive analysis of the effect of back protector on the prevention of vertebral and thoracolumbar injuries in serious motorcycle accident. Accident Analysis and Prevention, 135 , 105331.

[3] Ryder, B., Gahr, B., Egolf, P., Dahlinger, A. and Wortmann, F. 2017. Preventing traffic accidents with in-vehicle decision support systems - The impact of accident hotspot warnings on driver behaviour. Decision Support Systems, 99 , 64 – 74.

[4] Li, X., Liu, J., Zhang, Z., Parrish, A., and Jones, S. 2021. A spatiotemporal analysis of motorcyclist injury severity: Findings from 20 years of crash data from Pennsylvania. Accident Analysis & Prevention, 151 , 105952

[5] Hula, A., Schwieger, K., Saleh, P., Neumann, M. and Ecker, H. 2019. Objectifying and Predicting Motorcycle Accident Risk through Riding Dynamics. In Proceedings of the 26th International Technical Conference on the Enhanced Safety of Vehicles (ESV).

[6] Hula, A., Fürnsinn, F., Schwieger, K., Saleh, P., Neumann and M., Ecker, H. 2021 Deriving a joint risk estimate from dynamic data collected at¨ motorcycle rides, Accident Analysis and Prevention, 159, 2021, 106297, ISSN 0001-4575, URL: https://doi.org/10.1016/j.aap.2021.106297

[7] Hula, A. 2020. METHOD FOR DETERMINATION OF ACCIDENTAL HOTSPOTS FOR MOTORISED TWO-WHEELED VEHICLES. ”https: //register.epo.org/application?lng=en&number=EP19823762”.

[8] Schwieger K., Saleh P., Hula A.; Ecker H. and Neumann M.,2018. “viaMotorrad – Can motorcycle safety be measured?” ,12th International Motorcycle Conference Cologne

[9] Schwieger, K., Hula, A., Saleh, P., Ecker, H. and Neumann, M. 2020. Avoiding Motorcycle Accidents by Motorcycle Risk Mapping. In Proceedings of the 13th International Motorcycle Conference Cologne

[10] Ecker, H. and Saleh, P. 2016. "MoProVe - A Probe Vehicle for Traffic Accident Research", EVU Proceedings 25th Annual Congress, EDIS - Editing Centre of University of Zilina, ISBN: 978-80-554-1260-3; S. 29 – 41

[11] KTM Sportmotorcycle GmbH

[12] VBOX automotive, VBOX 3i Dual Antenna | 100Hz Vehicle Dynamics Measurement

[13] 2d-datarecording, 2D Debus & Diebold Messsysteme

[14] Bishop, C. 2006. Pattern Recognition and Machine Learning. Springer Berlin Heidelberg-New York

[15] Duda, R.O., Hart, P.E. and Stork, D.G. 2001. "Pattern Classification", Wiley, New York

[16] McLachlan and G.J. 1988, "Mixture Models: inference and applications to clustering", Statistics: Textbooks and Monographs, Dekker

[17] Draper, N.R. and Smith, H. 1998“Applied Regression Analysis”, Wiley, New York

[18] R: A language and environment for statistical computing. URL: https://www.R-project.org.

[19] Andreas Hula , Klemens Schwieger and Peter Saleh, Individual Motorcycling Safety: Creating a safety profile from riding data, To appear in: Proceedings of the Transport Research Arena TRA 2022, November 14-17, 2022, Lisbon, Portugal

[20] VSF Austrian Road Safety Fund

SPATIAL SOUND ASSISTANCE SYSTEM FOR 360 DEGREE HAZARD AWARENESS AND SAFE DRIVING

Masaki MARUYAMA
Junichi SAKAMOTO
Honda R&D Co. Ltd.
Japan

Paper Number 23-0111

ABSTRACT

In everyday driving situations, potential sources of collisions can appear from any direction around the driver. Driver assistance systems have been highly desired to assist driver's hazard awareness from all directions in order to eliminate any kinds of traffic accident fatalities. The current study addressed whether simulated spatial sounds providing directional and hazard attribute cues for potential collisions can facilitate drivers' identification of traffic hazards and reduce collision incidence in the front and rear spaces.

Forty-eight participants took part in our simulator experiment. We used a driving simulator (Honda Driving Simulator Type-DB Model S) to present them various traffic scenes with respect to the hazardous direction and recorded their driving operations. Participants' gaze directions were also recorded with an eye tracker implemented on the simulator. To provide a directional cue of hazardous traffic participants, we presented spatial sounds on the directions of hazard participants, using two speakers implemented in a driver's seat. To provide an attribute cue for hazardous objects, we classified the traffic participants into four categories (vehicles, motorcycles, bicycles, pedestrians) and presented a corresponding imitative sound for each hazard object. Presentations of monaural sound without directional cue and signal sound without attribute cue were also used as a comparison basis.

The current study observed a decrease in collision frequency and a significant reduction of onset time for pushing down the brake pedal for frontal hazard when spatialized signal sounds were presented compared with no HMI condition. A decrease in collision frequency with gazed hazards in the rear space was also observed when spatialized imitative and signal sound were presented relative to no HMI condition. The results lend to support our hypothesis that the directional cue can be effective for safer driving behaviors. On the other hand, improvements were not obtained when attribute cues were presented for both behavioral responses or the collision frequency. Significant facilitations were found in gaze responses and decelerate operations especially for rear hazards, but they did not result in a reduction of collision frequency.

Although the well-known front-rear ambiguity was confirmed in stationary sound localization, the current study observed the effectiveness of directional cue in reducing the collision frequency. It is possible that movements of spatial sound sources with hazard traffic participants could improve the resolution of front-rear sound localization. The influence of front-rear ambiguity might have also been reduced by extended spatial attention from the rear to

the front under the auditory directional cue towards the rear space. The attribute cue did not provide any effective improvements in the current study. However, we believe that in certain traffic situations where the type of hazards involved could represent more important information to the driver, the effects of attribute cue could reveal a potentially larger impact.

Our observations of the effective assistance of directional cue in spatial sound provide important references in terms of human factors for considering informative HMI that facilitates hazard awareness from all directions and help safer driving behaviors.

INTRODUCTION

In everyday driving, collision possibilities with hazardous traffic participants (e.g., pedestrians, cyclists, other vehicles) can occur from any direction around the driver. Hazardous traffic participants can thus appear outside the visual field of drivers, including behind them in the rear space. Recently, many of modern vehicles are equipped with sensor systems that possess capabilities to detect hazardous traffic participants all around the vehicle. There is also a growing interest for human machine interfaces (HMI) in driving assistant systems to provide hazard information in all direction around the drivers in order to support safer driving behaviors.

While simple alarms have often been used to inform hazard states or events, previous studies have addressed more informative cues in order to identify hazard sources. For example, a directional cue provided by a spatialized sound source location has been shown to facilitate our responses to visual targets within front space [1-4] and across front and rear spaces [5, 6]. These findings suggest that drivers can rely on the auditory spatial cue to direct their attention effectively towards the space where hazardous traffic participants appear, facilitating identifications of such potential hazards. However, such spatial auditory HMI has not been widely implemented in driver assistance systems in vehicles to date. The current study revisited the hypothesis on the effectivity of auditory directional cue, using a recent spatial sound technology which can produce sound images from any direction around drivers.

It is widely acknowledged that in everyday experiences, sound-producing events due to material interaction establish auditory informative cues for humans [7, 8]. Through common driving experiences, drivers are also able to use traffic environment sounds as auditory cues directly to distinguish between different types of potential hazards. For visual cues, selective attention based on object features such as size, color and shape have been well understood [9], however object-based auditory attention has not been well addressed, because of the lack of precise definition of object formation [10, 11] and, to our best knowledge, because of the lack of available realistic sound in experimental conditions. Using a recent interactive sound simulation technology, the current study created sounds that realistically imitated sounds emanating from actual traffic participants. We examined whether the auditory cues for selective object-based attention are effective in facilitating behavioral responses for safer driving behaviors and reducing collision possibility.

The objective of the present work is to reveal whether recent sound technologies can provide effective cues about location of potential hazard objects to drivers and enable safer driving behaviors. Considering that our findings on audiovisual attention directed in the front space cannot always be generalized to the rear space [5, 12], we

evaluated the effectiveness of the cues regarding to front and rear hazards separately. Our findings provide important references in terms of human factors for considering informative HMI to assist raising awareness about potential hazards from all directions.

METHODS

Participants

Forty-eight participants (including 11 female participants) took part in this experiment. Their mean age was 53.4 years (range: 23 – 65 years). They had normal or corrected-to-normal eye sight. All participants had a valid Japanese driver's license. Self-reported frequency of driving was three days a week or less for every participant, except for two, who reported driving daily. This research received ethical approval from the Bioethics Committee of Honda R&D Co., Ltd. All participants gave their written informed consent prior to the start of the experiment. They received 14,500 JPY for taking part in this experiment, including travelling expenses.

Apparatus

An overview of the driving simulator used in this experiment (Honda Driving Simulator Type-DB Model S) is illustrated in Figure 1. The participants drove an automatic transmission vehicle using accelerator and brake pedals and a steering wheel. A gear shift and turn indicators were also available. Three flat-screen displays provided the outside view of the driver's vehicle from the viewpoint of a typical driver's position, i.e. at a 104.8 cm distance from the central display. The scope of view was 150 deg horizontal and 30 deg vertical. Images of side mirrors (50 deg horizontal visual field) and the rear view mirrors (30 deg) were shown on the displays.

To output spatialized sounds, we used a driver's seat with two speakers installed to the left and right side of the head at shoulders height location. Using interactive audio middleware software, the Vector Based Amplitude Panning sound spatialization algorithm [13] was used to control in real-time the relative level of sounds for the left and right speakers. A filter for non-individualized head related transfer function (HRTF) was also applied to implement binaural processing to generate a sense of sound source location that can distinguish the front and rear spaces, as well as distinguish the leftward and rightward directions. Monaural sounds were provided from a single speaker (JM10 pro., Conisis) placed at the center of the front panel.

To create imitation sounds, we first classified traffic participants that could cause hazards into four categories: (1) vehicles, (2) motorcycles, (3) bicycles and (4) pedestrians. Samples recorded sounds were acquired for each category. Using a layered sound approach mixed with interactive audio middleware, the samples of vehicles were simulated by combining engine sound, road noise, and wind noise. The simulated vehicle sound playback precisely reflected the effects that engine rotation speed and engine load have on the vehicle engine sounds, and the effects of the vehicle's speed on the road and wind noises, which made it possible to create imitation sounds that realistically represent the sounds of actual vehicle in the simulation. This was also the case with motorcycles. Additionally to the sound spatialization, a level attenuation effect as a function of relative distance was also applied to the simulated vehicle sounds.

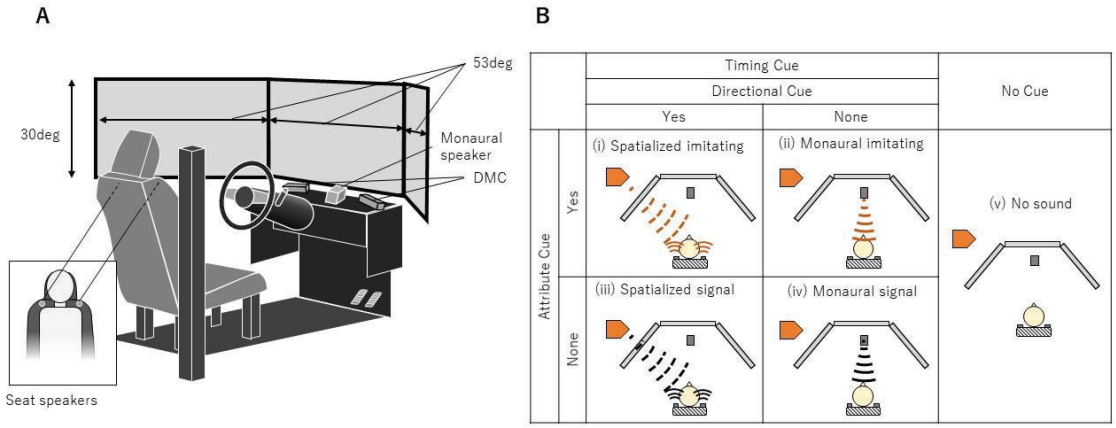

Figure. 1 A. An overview of the driving simulator used for the sound HMI experiment. The experiment participants drove an automatic transmission vehicle viewing the outside visual scenes presented on three flat screen displays. Spatial and monaural sounds were output with the seat speakers and monaural speaker respectively. Driver monitoring cameras (DMCs) were used to record participants' gaze direction. B. Summary of sound HMI conditions and cue factors considered in the present study. Presentation of spatialized sound from hazard object provided a directional cue, whereas monaural sound presented from front regardless of direction of hazard object gave no directional cue. To provide an attribute cue, traffic participants were classified into four categories (vehicles, motorcycles, bicycles and pedestrians), and an imitative sound for the hazard object category was presented. In contrast, a common signal sound was presented for all four hazard objects categories in order to evaluate sound presentation condition without any attribute cue. While the sound presentation conditions (i, ii, iii and iv) provide a timing cue of realized hazard event, the no sound HMI condition did not give any cue.

The signal sound consisted of impulsive synthetic tones of individual duration that were repeated at a rate of 0.3 seconds. The same sound was used for the different categories of traffic participants, and thus it did not provide any cues as to the attributions of hazard traffic participants. Attenuation as a function of relative distance was not applied to the signal sound.

HMI presentation

There were five variations of HMI presentation (Fig. 1B): (i) spatialized imitating sound, (ii) spatialized signal sound, (iii) monaural imitating sound, (iv) monaural signal sound and (v) no sound. A directional cue was presented in (i) and (ii) but not in (iii) and (iv). Therefore, differences in driving behaviors between (i) and (iii) and between (ii) and (iv) indicate an effect of multiple factors including directional cue. On the other hand, since an attributional cue was given in (i) and (iii) and not in (ii) and (iv), differences between (i) and (ii) and between (iii) and (iv) reveal an effect of multiple factors including the attribute cue. Whereas the conditions (i) to (iv) gave a timing cue of realized hazard to the experiment participants which was not the case in (v). Therefore, comparing the driving behaviors seen in (i) to (iv) with those in (v), we examined the effect of hazard presence cue on driving

safety.

In order to assist attention allocation to hazardous traffic participants, HMI requires a distinction between hazardous and non-hazardous objects. While the development of discrimination rules in real traffic have been difficult issues, the predefined scenarios of the simulator make it possible to identify traffic participants that could become hazardous, based on the driving behavior of own vehicle. We identified the potential hazardous participants in the scenarios before conducting the experiment. To detect a transition from potential hazard to realized hazard, we used the centripetal time to collision (centripetal TTC; Fig. 2A), which extends the well-known TTC from one dimension to two dimensions. This was obtained by dividing a relative distance from the own vehicle to the object by a relative velocity component in the direction from the own vehicle to the object. To define the threshold of centripetal TTC, we conducted a preliminary test of subjective discrimination while driving the hazard prediction courses. We found that the hazard sense was usually evoked when centripetal TTC became less than 5 sec. Here, the TTC is known not to be an effective indicator when the hazard is evoked by uncertainty in behaviors of objects that are at a close distance, at a low relative velocity. It is also the case with centripetal TTC. Indeed, we found the transition often occurred when the relative distances of potential hazard objects are shorter than 5 m. Putting our preliminary findings together, we deemed hazard as realized when the centripetal TTC is shorter than 5 sec or the relative distance is shorter than 5 m. The simulator calculated the relative distance and centripetal TTC at a frequency of 100 Hz. The sound HMI was presented in the conditions (i) to (iv) when this criterion was satisfied.

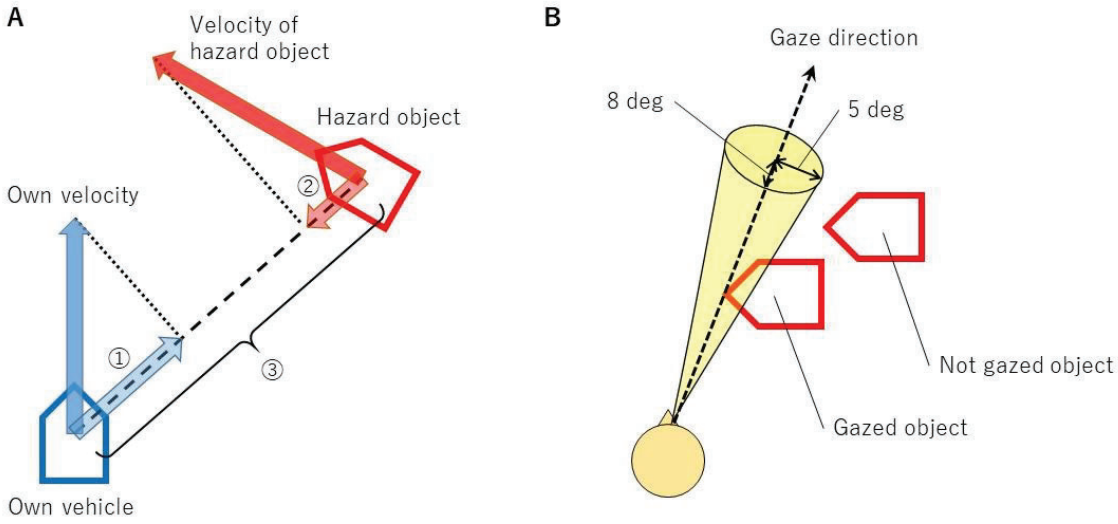

Figure. 2 A. Illustration of simulated parameters used for the estimation of time to collision (TTC) in a centripetal form, that is, relative distance to a traffic participant (①) was divided by approaching speed toward the traffic participant (②+③). B. Diagram explaining the estimation of effective visual field for spatial perception in the current study. The effective visual field is represented with a virtual cone oriented towards the gaze direction. We estimated that the experiment participants obtained spatial perception of realized hazard objects when the objects and cone overlapped for longer than 200 ms within a time window of 250 ms, which we term "gaze" in this study.

Procedure

The experiment started with a measurement of the directional accuracy of sound images. We presented the spatialized sounds of the vehicle, bicycle, pedestrian and signal from the directions of front, back, left, right and four diagonals, sequentially and in a random order. Note that the location of spatialized sound was fixed for each presentation. The time duration of each presentation was 1 sec. After every presentation, the participants indicated the direction of the sound image by marking the direction on an egocentric coordinate figure illustrated on their response sheet.

Next, the participants drove a driving course preinstalled in the simulator to familiarize themselves and verify the sensitivities of steering and pedaling and the size of their own vehicle. They also drove the 4th of six hazard prediction training courses preinstalled on the simulator, in order to familiarize themselves with inner-city driving on the simulator.

We then recorded the driving behaviors (gaze directions, accel and brake pedal operations, vehicle movements) as they drove the remaining five courses of hazard prediction training. The driving of five courses were separated by short breaks. The order of the courses (1st, 2nd, 3rd, 5th, 6th) was same for all participants, while the five Human-Machine Interface (HMI) conditions (spatialized imitating, spatialized signal, monaural imitating, monaural signal, no HMI) were given in a random order. We stopped the experiment if participants showed signs of simulator sickness

Recordings and analyses

To evaluate the effectiveness of HMI presentation on driving operations in the experiment participants, we recorded their operations of vehicle accelerator and brake pedals. We also recorded simulations of positions and velocities of their own vehicle and the twenty closest traffic participants, as well as the centripetal TTC and events of collision. The recording frequency was 100 Hz. To indicate a degree of safe driving with respect to the realized hazard traffic participants, the current study used latencies of deceleration operations from the time of hazard realization. Here, an onset of brake-pedal and offset of accel pedal after the beginning of hazard realization were respectively extracted as the latencies of deceleration operation.

The effectiveness of additional sound HMI on their attention allocation to realized hazard participants was analyzed, using the gaze direction recorded at a frequency of 20 Hz with a camera-based driver monitoring system developed by Seeingmachines company. To represent an effective visual field, we set a virtual cone towards the gaze direction (Fig. 2B). We estimated that they became aware of realized hazard objects that were overlapped with this cone. A previous study indicated that an effective visual field size of spatial perception is larger than 5 deg, when a visual target stimulus was presented for 250 ms on a dynamical background simulating a driving situation [14]. Hence, we set the horizontal radius of cone at 5 deg. Taking into account a large error of gaze recording in vertical direction, we set the vertical radius of cone at 8 deg. We deemed that they became aware of realized hazard objects when the objects and cone overlapped for longer than 200 ms, within a time window of 250 ms. In addition to the latencies of vehicle operations and gazing, the frequencies of collision were compared among the HMI conditions to show the assistance effect of HMI presentation.

In statistical analyses, we tested significance of differences in the collision frequency amongst the HMI conditions

using a z-test. For the difference in the latencies of gazes and deceleration operations, we conducted a Wilcoxon signed-rank test. The current study reports uncorrected p-values in multiple comparisons.

RESULTS

Collision and hazard gaze frequencies

Collision frequency with traffic participants to the front and rear of own vehicle at the time of hazard realization have been summarized for each HMI condition in Fig. 3A and B. In the case of the front hazard, a decrease in collision frequency was observed in the conditions of spatial signal sound HMI (6.9 %) relative to the no HMI condition (11.1 %). On the other hand, in the case of the rear hazard, an increase in collision frequency was observed in the presentation of monaural signal sound (10.5 %) relative to the no HMI condition (4.9 %), whereas such increase was not found for any other conditions of sound HMI presentation (5.6 – 6.8 %).

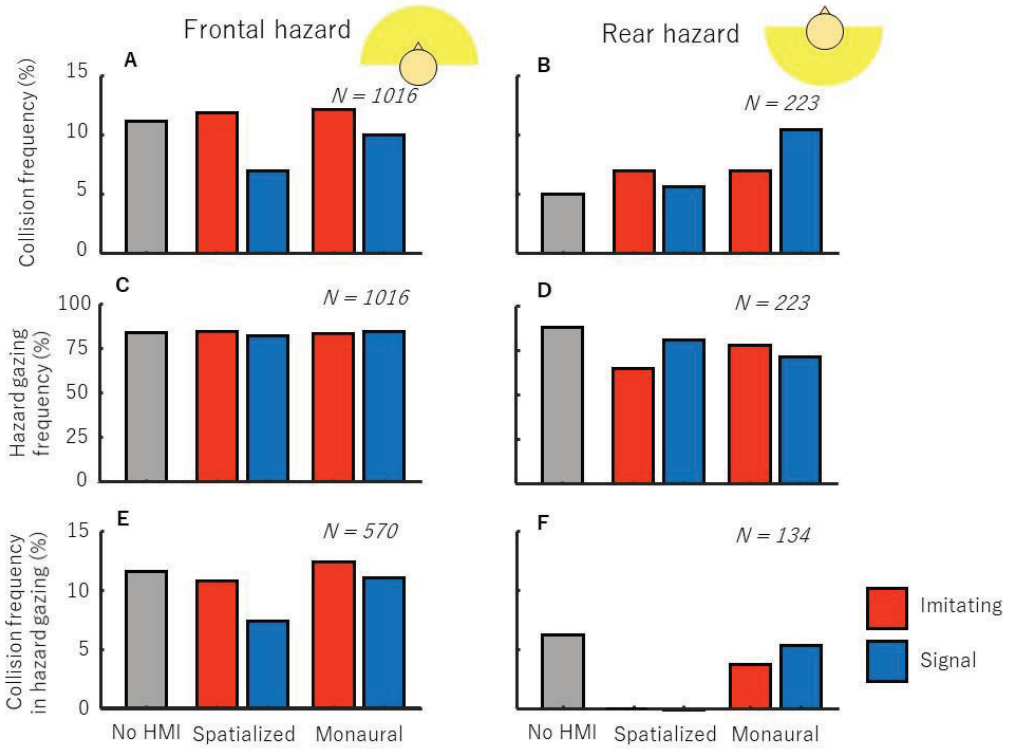

Figure. 3 A. Collision frequency with traffic participants to the front of own vehicle at the time of hazard realization. B. Collision frequency with traffic participants in the rear space. C, D. Gaze frequency of hazard traffic participants in the front and rear spaces. E, F. Collision frequency with gazed hazardous participants in the front and rear spaces, respectively. The results in each scene and participant were merged, and then the frequency was obtained in each HMI condition. The sample number is indicated on the top right in each panel. Note the differences in the frequency among the HMI conditions did not reach to a significance level in our statistical test ($p > 0.1$).

To find whether there was an assistance effect that helped prevent drivers from missing hazard objects, the gaze frequencies of hazard object in the front and rear spaces is illustrated in Fig. 3C and D. The frequency of front hazard gaze ranged from 82 % to 85 % in every condition of sound HMI presentation, and no significant difference was obtained relative to the no HMI condition (85 %). In the case of rear hazard, the gaze frequency in the condition of spatialized imitating sound HMI (64 %) was lower relative to the condition of no HMI (89 %) and the other conditions of sound HMI presentations (Spatialized signal: 81 %; Monaural imitating: 77 %; Monaural signal: 71 %), though the differences were not significant ($p > 0.1$).

To indicate whether the spatial perception of hazard objects with gazing was effective, the collision frequency in the condition of hazard gaze was compared under different HMI conditions. In the case of front hazard, the collision frequency obtained with spatialized signal sound (7.3%) was lower relative to the no HMI condition (11 %), whereas the other sound HMIs (11 – 12 %) did not show any significant improvement with respect to the no HMI condition (Fig. 3E). In the rear hazard scenario, no collision was observed in the spatialized sound HMIs, whereas it appeared at a frequency of 6.3 % in the condition of no HMI (Fig. 3F).

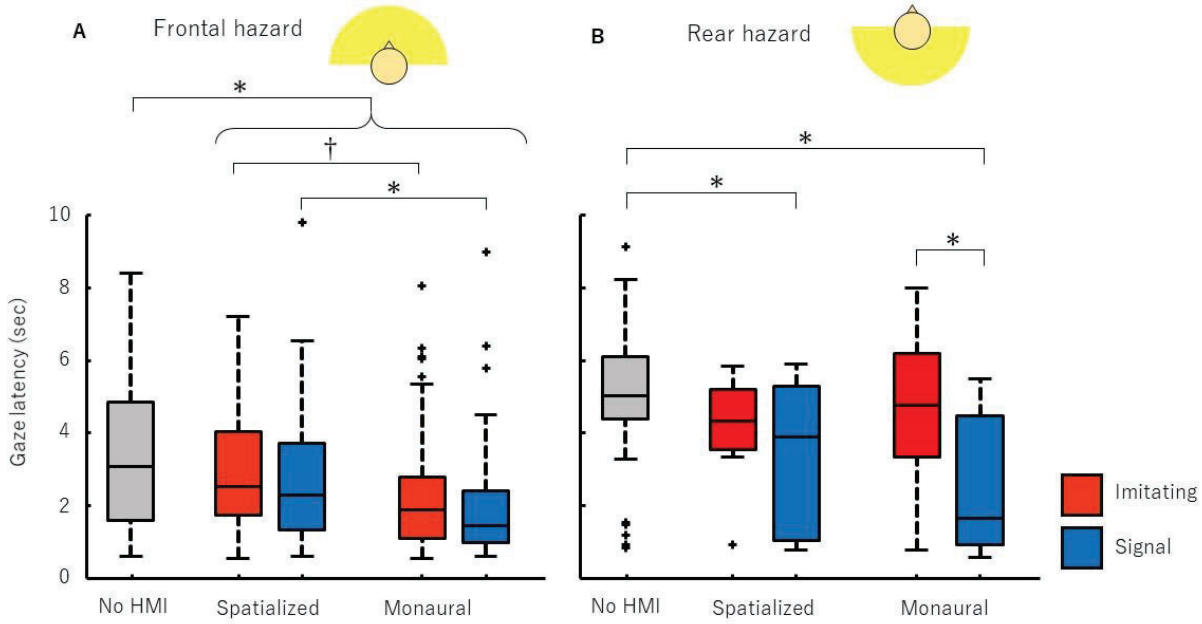

Figure. 4 A. Latencies of front hazard gaze from the hazard realization (i.e., the time when the centripetal TTC falls below 5 sec or the relative distance is shorter than 5 m.) B. Latencies of rear hazard gaze from the hazard realization. Box plots are used to indicate medians, quartiles and range of latency distributions in each HMI condition. The results in each scene and participant were merged in each distribution. Significant and marginal difference in the median between HMI conditions are marked by an asterisk (*, $p < 0.05$) and dagger (†, $p < 0.1$), respectively.

Hazard gaze latencies

Figure 4A summarizes the latencies of frontal hazard gaze in the form of boxplots. We found significantly shorter medians of latencies in every condition of sound presentation (Spatialized imitating: 2.5 s; Spatialized signal: 2.3

s; Monaural imitating: 1.9 s; Monaural signal: 1.5 s), relative to the condition of no sound presentation (3.5 s). Significantly shorter medians of latencies were also found with the signal sound presentation compared with the imitating sound presentation in both spatialized and monaural conditions. In the cases of rear hazard, the spatialized and monaural signal sounds significantly decreased the gaze latency relative to the condition of no sound presentation (Fig. 4B; Spatialized signal: 3.9 s; Monaural signal: 1.6 s; No HMI: 5.0 s; $p < 0.05$). We also found a significantly shorter median of latency with the signal sound relative to the imitating sound in the monaural condition ($p < 0.05$). No significant difference was found between the conditions of no sound presentation and 3D and monaural imitating sound presentations ($p > 0.1$).

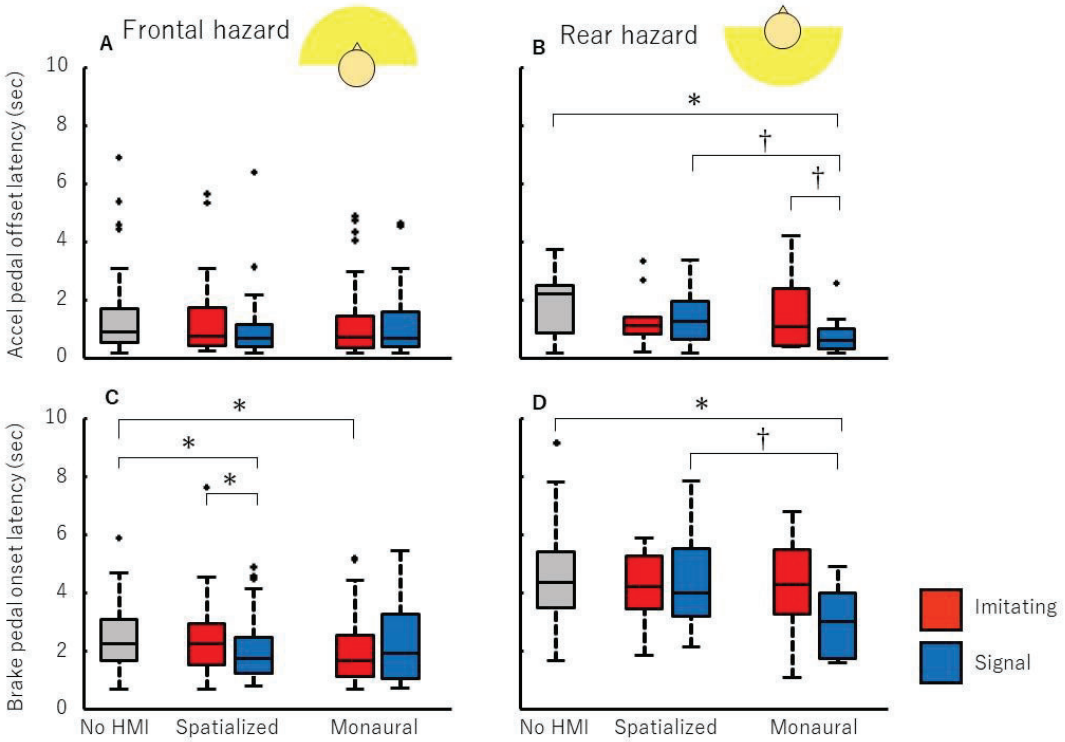

Figure. 5 A, B. Latencies of accel pedal offset from the hazard realization in the front and rear spaces, respectively. C, D. Latencies of brake pedal onset from the hazard realization in the front and rear spaces, respectively. Box plots are used to indicate medians, quartiles and range of latency distributions in each HMI condition. The results in each scene and participant were merged in each distribution. Significant and marginal difference for the median between HMI conditions are marked by an asterisk (*, $p < 0.05$) and dagger (†, $p < 0.1$), respectively.

Decelerate operation latencies

The offset latencies of acceleration pedaling relative to the front hazard realizations were summarized in Fig. 5A. No significant difference in their median was found among the HMI conditions ($p > 0.1$). When the hazard traffic participants were in the rear space (Fig. 5B), a significant decrease in the latency of acceleration pedaling was obtained with the presentation of monaural signal sound (0.6 s), relative to the no HMI condition (2.2 s; $p < 0.01$). A marginal decrease was also observed relative to the conditions of spatialized signal HMI (1.3 s; $p < 0.1$) and

monaural imitating HMI (1.1 s; $p < 0.1$).

As to the onset latencies of pushing down the brake pedal, a significantly shorter median of latency was obtained in the condition of spatialized signal sound HMI (1.9 s) compared with the conditions of no HMI (2.2 s; $p < 0.01$) and spatialized imitating sound HMI (2.2 s; $p < 0.05$) in the front hazard (Fig. 5C). We also found a significantly shorter median of latency in the condition of monaural sound HMI relative (1.6 s) to the condition of no HMI ($p < 0.05$). When the hazard traffic participants were in the rear space (Fig. 5D), a decrease observed in the presentation of monaural signal sound (3.0 s) relative to the condition of no HMI was significant (4.4 s; $p < 0.05$), and that the effect relative to the condition of spatialized signal sound was marginal (4.0 s, $p < 0.1$).

Directional accuracy of 3D sound image

To estimate the sound localization accuracy of spatialized sound HMI, we used the directional responses acquired for every participant at the beginning of the experiment. We took an absolute value of angular difference between the presented spatial sound and the directional response reported by the participants. The errors of directional responses obtained from each participant were summarized for vehicle, bicycle, pedestrian, and signal sounds in Fig. 6. The median errors obtained for all participants were below the chance level of 90 deg, except for the directly in front and rear directions, where the sound image was often perceived as located in the opposite direction. The overall median across all the participants, indicated a localization accuracy of 30 deg.

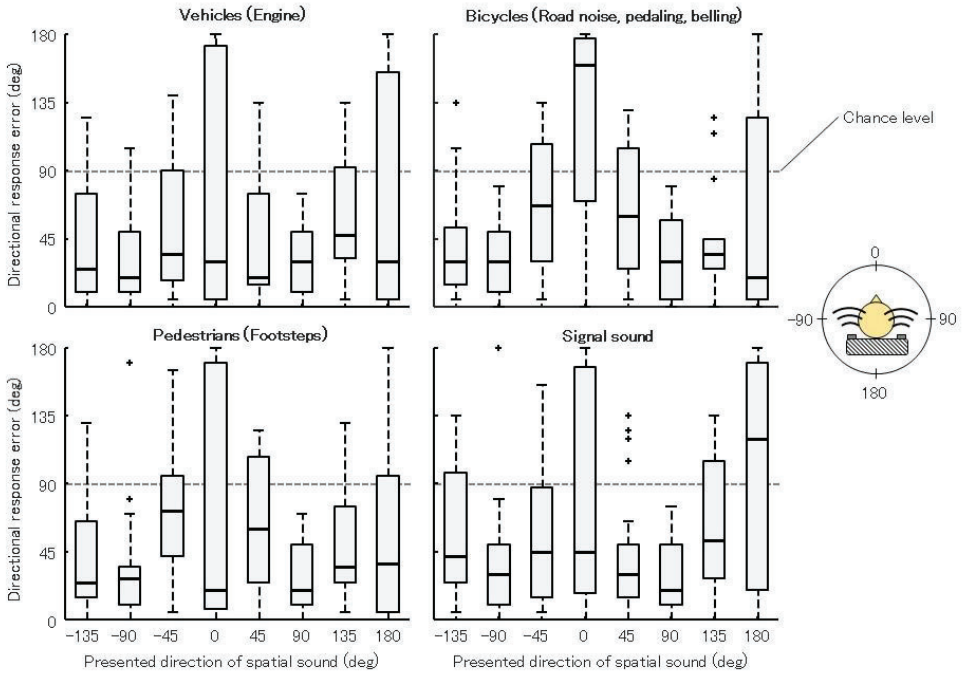

Figure. 6 Distributions of localization errors of spatial sounds presented on the directions of front, back, left, right and four diagonals. Imitative sounds of vehicle, bicycle and pedestrian and signal sound were used in the measurement of localization error. An absolute value of angular difference was taken between the presented spatial sound and the directional response reported by the participants. The chance level indicates the error size that we would obtain for random responses.

DISCUSSIONS

Timing cue assists early hazard gaze

Significantly earlier gazes of front hazard were found with every condition of sound HMI presentation compared with the no sound HMI condition (Fig. 4A). Our results confirmed that the presentation of an auditory cue on the occurrence of a visual event can facilitate earlier behavioral responses without providing any spatial or attribute cue, as indicated in the previous studies [15-17]. One might think that a cue of hazard presence, rather than a timing cue of transition event from potential hazard to realized hazard, decreased the gaze latency. If it were true, the experiment participants could have been aware of hazard presence more frequently, which could have been observed as an increase of hazard gaze frequency under the conditions of sound HMI presentation relative to the condition of no sound HMI. However, the little increase we observed (Fig. 3C) does not support this hypothesis. Our results suggest that the early cue of transition from potential hazard to realized hazard decreased the latency of gaze under the conditions of sound HMI presentations.

In the rear hazard situation, a significant decrease of gaze latency was obtained with the presentation of signal sound HMI compared with the no HMI condition (Fig, 4B), which could be related to the effect of the timing cue. On the other hand, no significant decrease of gaze latency was found with the presentation of imitation sound. In this case, because of the attenuation effect on the volume of imitation sound as a function of relative distance, an auditory awareness of the gradual onset of imitation sound might be delayed compared with the signal sound that was free from the attenuation effect. In addition, a visual awareness of the transitions from potential hazard to realized hazard in the rear space would tend to be delayed with respect to those in the front visual field while driving forward. For these reasons, no significant decrease was found in the current study with the presentation of imitation sounds for rear hazard situations.

Directional cue assists a decrease of collision frequency

A decrease of collision frequency with the frontal hazard was observed in the presentation of spatialized signal sound compared with the no HMI condition (Fig. 3A and E), which could be related to the significantly early onset of pushing down the brake pedal in the spatialized signal sound presentation relative to the no HMI condition (Fig. 5C). It was not the case where the monaural signal sound was presented. Taken together, the current study indicated that a directional cue is effective in the situation of front hazard to facilitate the earlier onset of pushing down the brake pedal, resulting in the decrease of collision frequency. The results are consistent with the previous studies in which a presence of spatial auditory cue reduced the time for visual search [2] and acceleration or braking response in driving hazard avoidance [5].

The current study also found that the latency of brake pedaling onset was significantly shorter under the condition of spatialized signal sound compared with the spatialized imitation sound. Given that the directional cue was already enough to identify a hazard in the front space, an intense sense of hazard under the signal sound would be more effective to decrease the latency of brake pedaling relative to the spatialized imitation sound attribute cue. The assumption was supported by our finding of significantly early onset of pushing down the brake pedal for the monaural imitation sound (Fig. 5C), in which the attribute cue would assist an early identification of hazard for

participants.

No collisions were observed when the rear hazard was gazed for the presentations of spatialized sound HMI, whereas it was not the case with either the presentation of monaural sound HMI or no sound HMI presentation (Fig. 3F). This tendency was not observed when the collision frequency includes the cases where the realized hazard object was within the effective visual field of spatial perception [14] (Fig. 3B). In other words, neither presence perception of realized hazard object, which would be available within an effective field extending 15deg [14, 18], nor spatial perception of potential hazard object would be effective to use the directional cue. Taken together, the auditory directional cue was informative when the experiment participants obtained the visual spatial perception of hazardous object in the rear space, which was effective to reduce the collision frequency.

The previous psychophysical studies obtained a facilitation effect on earlier behavioral responses as visual targets moved towards the peripheral visual field [19] and even towards the rear space [20]. It should be also noteworthy that rapid responses were previously obtained with a presentation of close rear auditory warning signal relative to a far front auditory warning signal [21]. Taking into account that the seat speakers were close to the ears and were in the rear space, the presentation of spatial sound from the seat speakers could result in a better performance relative to the monaural sound from the front speaker. Although we did not obtain the rapid responses in both of eye movements and pedaling, a better performance was found in the form of low frequency of rear hazard collision under the presentation of spatial sound compared with the monaural sound.

Regarding the sound localization accuracy of the directional cue, front-rear ambiguity is known to occur in localization of sound source in humans. Although we can use monaural spectral cues, originating from the spectral filtering of sound by the pinnae, head, and torso in order to identify whether sound sources are in front or rear, the ambiguity remains for both actual and spatialized sound sources [22-24]. While the spectral filtering is different among individuals because of the individual anatomical variance of pinnae, head and torso, the current study used the same spectral filtering across the experiment participants (i.e., non-individualized HRTF). For these reasons, the relatively large front-rear confusion in localizing sound image occurred in our presentation of spatial sound (Fig. 6). Nevertheless, the current study observed the tendencies of effectiveness for the directional cue.

Recall that the spatial sound source from the hazardous traffic participants moved relative to the own vehicle. Considering that the resolution of front-back ambiguity in the localization of a sound source can be improved by the movement of sound source [25], the perceived directional cue in the experiment participants might have been more accurate than those we measured for the stationary directional sounds.

It should be also noteworthy that there might be neurophysiological mechanism that makes the front-rear ambiguity less serious. Receptive field (RF) of audiovisual neurons in the superior colliculus (SC), known to be involved in allocating spatial attention [26-28], can show different spatial properties in the RF extent depending on their locations. That is, the neurons in the rostral portion of the SC, responsive to stimuli presented from frontal space, have often showed visual and auditory RFs less than 10 deg and 20 deg respectively in diameter, whereas those in the caudal SC that respond to stimuli presented from the periphery have visual RFs ranging from 40 to 100 deg and auditory RFs from 60 to 135 deg in diameter [12, 19, 29]. Thus, an auditory directional cue presented in the rear-left spatial area, for instance, is possibly captured within the RF of an audiovisual neuron that has a visual RF responsive to visual stimuli in the front-left spatial area. Although the current study found the relatively

large front-rear confusion in localizing sound image, the auditory directional cue towards a direction in the rear might be effective to allocate spatial attention extending from the direction to the front left, which would assist the hazard awareness not only in the lateral rear but also in the lateral front on the same side.

Early gaze and deceleration with monaural signal sound presentation for rear hazard

A significantly shorter latency of rear hazard gaze was found with the presentation of monaural signal sound compared with the no HMI condition (Fig. 4B). It was also true for the onsets to release acceleration pedal and to press brake pedal (Fig. 5B and D). While the early latencies could give a little extra time to collision, the experiment participants were required by themselves to identify the traffic participant causing the hazard status, because the monaural signal sound conveyed no spatial nor object cue. Therefore, it would be possible that the early gaze did not assist to decrease the collision frequency. The early responses of deceleration would not always help drivers from preventing the collision, since the hazard participants were in the rear space. Indeed, the current study did not observe any significant decrease of collision frequency when the rear hazard was gazed under the presentation of monaural signal sound HMI (Fig. 3F). Rather, an increase of the collision frequency was observed under the presentation of monaural signal sound compared with the other HMI conditions (Fig. 3B). Our results suggest that the signal sound could decrease the latency of gaze and deceleration, but it could not always assist collision avoidance because of the lack of directional information on hazard participant.

Early gaze of frontal hazard with monaural sound from the front

The source of monaural sound in the current study was always at the front, regardless of the directions of hazard participants. Thus, our monaural sound could not exactly provide a cue on the direction of hazard traffic participants. However, under the condition of high frequency of hazard events in the front hemifield compared with the rear hemifield, the monaural sound from the front location would be beneficial to allocate their attention on the front direction. If it was true, an effect of the valid cue should have been found as early responses, as shown in the previous studies [5, 15, 30]. Indeed, we found a significantly early gaze of hazard traffic participant under the condition of monaural sound presentation relative to the no HMI condition, when the hazard traffic participants were in the front (Fig. 4A). Overall, although the monaural sound was not informative on the hazard direction, its source location in the front space could be effective to facilitate the front hazard gaze.

CONCLUSIONS

Our observations of the effective assistance of directional cue in spatial sound provide important references in terms of human factors for considering informative HMI that facilitates hazard awareness from all directions and help safer driving behaviors.

Acknowledgements

We thank Mr. N Fujimoto for fruitful advices on the present works, Mr. K Nagura and H Ono and Managebusiness Co.,Ltd for functional extensions of driving simulator. We are grateful to Mr. T Ezaki, Y Shimura and Auto Technic Japan for their helps in acquiring data, Mr Nakayama for his helps on data processing.

References

[1] Perrott, D. R., Sadralodabai, T., Saberi, K., & Strybel, T. Z. (1991). Aurally aided visual search in the central visual field: Effects of visual load and visual enhancement of the target. Human Factors, 33, 389-400.

[2] Begault, D. (1993). Head-Up Auditory Displays for Traffic Collision Avoidance System Advisories: A Preliminary Investigation. Human factors.

[3] Corneil, B. D., van Wanrooij, M., Munoz, D. P., & van Opstal, A. J. (2002). Auditory-visual interactions subserving goal-directed saccades in a complex scene. Journal of Neurophysiology, 88, 438-454.

[4] Hancock, P. A., Mercado, J. E., Merlo, J., & Van Erp, J. B F. (2013). Improving target detection in visual search through the augmenting multi-sensory cues. Ergonomics, 56, 729-738.

[5] Ho, C., & Spence, C. (2005). Assessing the effectiveness of various auditory cues in capturing a driver's visual attention. Journal of Experimental Psychology: Applied, 11, 157-174.

[6] Lee, J., & Spence, C. (2015). Audiovisual crossmodal cuing effects in front and rear space. Frontiers in Psychology, 6:1-10.

[7] Gaver, W. W. (1993a). What in the world we hear? An ecological approach to auditory event perception. Ecological Psychology, 5, 1-29.

[8] Gaver, W. W. (1993b). How do we hear in the world? Explorations in ecological acoustics. Ecological Psychology, 5, 285-313.

[9] Desimone, R. and Duncan, J. (1995) Neural mechanisms of selective visual attention. Annu. Rev. Neurosci. 18, 193–222.

[10] Bizley JK, Maddox RK, Lee AK (2016) Defining auditory-visual objects: behavioral tests and physiological mechanisms. Trends Neurosci 39:74–85.

[11] Shinn-Cunningham, B. G. (2008) "Object-based auditory and visual attention," Trends Cogn. Sci., 12, 182–186.

[12] Spence C., Lee J., Van der Stoep N. (2020). Responding to sounds from unseen locations: Crossmodal attentional orienting in response to sounds presented from the rear. Eur. J. Neurosci. 51:1137–1150.

[13] Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. Journal of the Audio Engineering Society. 45. 456-466.

[14] Ayama, M., Mekada, Y., Kodama, D., & Kasuga, M. (2005). Effects of visual and conversation loads on functional visual field on a dynamic background. J. Illum. Engng. Inst. Jpn. 89, 453-471.

[15] Diederich, A., & Colonius, H. (2008). Crossmodal interaction in saccadic reaction time: Separating multisensory from warning effects in the time window of integration model. Experimental Brain Research, 186, 1-22.

[16] Los, S. A., & Van der Burg, E. (2013). Sound speeds vision through preparation, not integration. Journal of Experimental Psychology: Human Perception and Performance, 39, 1612-1624.

[17] Spence, C., & Driver, J. (1997). Audiovisual links in exogenous covert spatial orienting. Perception & Psychophysics, 59, 1-22.

[18] Kodama, D., Mekada, Y., Ayama, M., Kazuga, M. (2000). Consideration on a functional visual field in visual

and conversation tasks while driving. ITE Technical Report, 24, 7-12.
[19] Lee, J., & Spence, C. (2017). On the spatial specificity of audiovisual crossmodal exogenous cuing effects. Acta Psychologica, 177, 78-88.
[20] Perrott, D. R., Saberi, K., Brown, K., & Strybel, T. Z. (1990). Auditory psychomotor coordination and visual search performance. Perception & Psychophysics, 48, 214-226.
[21] Ho, C., & Spence, C. (2009). Using peripersonal warning signals to orient a driver's gaze. Human Factors, 51, 539-556.
[22] Asano, F., Suzuki, Y., & Sone, T. (1990). Role of spectral cues in median plane localization. Journal of the Acoustical Society of America, 88, 159-168.
[23] Parise, C. V., Knorre, K., & Ernst, M. O. (2014). Natural auditory scene statistics shapes human spatial hearing. Proceedings of the National Academy of Sciences of the USA, 111, 6104-6108.
[24] Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. Journal of the Acoustical Society of America, 94, 111-123.
[25] Wightman, F. L., & Kistler, D. J. (1999). Resolution of front–back ambiguity in spatial hearing by listener and source movement. Journal of the Acoustical Society of America, 105, 2841-2853.
[26] Ignashchenkova, A., Dicke, P. W., Haarmeier, T., & Their, P. (2004). Neuron-specific contribution of the superior colliculus to overt and covert shifts of attention. Nature Neuroscience, 7, 56-64.
[27] Kustov, A. A., & Robinson, D. L. (1996). Shared neural control of attentional shifts and eye movements. Nature, 384, 74-77.
[28] Moore, T., Armstrong, K. M., & Fallah, M. (2003). Visuomotor origins of covert spatial attention. Neuron, 40, 671-683.
[29] Kadunce, D. C., Vaughan, W. J., Wallace, M. T., & Stein, B. E. (2001). The influence of visual and auditory receptive field organization on multisensory integration in the superior colliculus. Experimental Brain Research, 139, 303-310.
[30] Roberts, K. L., Summerfield, A. Q. and Hall, D. A. (2009) "Covert auditory spatial orienting: An evaluation of the spatial relevance hypothesis," J. Exp. Psychol. Hum. Percept. Perform., 35, 1178–1191.

NOVEL INTERFACES THAT ENHANCE A DRIVER'S ABILITY TO PERCEIVE FORWARD COLLISION RISKS

Takahiro Matsuoka
Tsuyoshi Nojiri
Honda R&D Co., Ltd.
Japan

Vanessa Krüger
Honda R&D Europe GmbH
Germany

Matti Krüger
Honda Research Institute Europe GmbH
Germany

Paper Number 23-0196

ABSTRACT

Forward Collision Warning (FCW) systems that alert a driver about the risk of rear-end collisions can contribute to a reduction of traffic accidents caused by human errors. Typically, FCWs create alerts that appear late when the risk is already high and are of binary nature, i.e., either in an alerting state during high risk or not producing any alert at lower risks. The choice at what risk level to start alerting in a binary manner is subject to a tradeoff between how much time an alert gives the driver to react and how necessary the alert appears to the driver. Our goal is to circumvent this limitation of classical binary FCWs to allow drivers to perceive developing risks early and in an intuitive manner and, accordingly, better avert developing risks with foresight. To that end, we propose a new system that assesses potentially hazardous situations in real time and continuously outputs a signal that alters its strength depending on the risk level. Here we report a study on the effect of variations of the proposed system on driving behavior and user acceptance.

The experiment was carried out in a driving simulator equipped with prototypes of visual, auditory, and tactile human-machine interfaces (HMIs). The participants performed driving tasks in two different driving scenarios. The subjective ratings of system acceptance were assessed with questionnaires on two dimensions, a usefulness scale and an affective satisfaction scale. The results indicate that, compared to an existing FCW system, all HMIs reduced driver reaction times and the visual HMI showed positive average scores of both usefulness and satisfaction in the driving scenario with high and medium collision risk. On the other hand, there was no HMI that achieved a good balance between the effect on driving safety and system acceptance in the scenario with lower criticality. These results suggest that the proposed notification system can improve driving safety and be perceived as subjectively acceptable in situations with high and medium collision risk despite the early signal. This makes

it a promising approach for circumventing the tradeoff between notification timing and risk perception. To address system effects on driving safety in situations with lower risk, further development iterations and long-term evaluations in a variety of traffic situations may be required.

INTRODUCTION

Forward Collision Warning (FCW) systems that alert a driver about the risk of rear-end collisions can contribute to a reduction of traffic accidents caused by human errors [1]. However, it is not guaranteed that a driver will successfully avert an accident after such an alert as it is typically triggered rather late when the risk is already high [2]. One approach to address this issue is to provide alerts at an earlier point in time to give drivers more time to react. One issue of such an approach lies in the risk of causing annoyance to the driver. Early warnings are more frequent than warnings that only appear in critical conditions and are thus at higher risk of appearing in situations regarded as uncritical by the driver [see 3, p. 31]. The driver could then consider system alerts as irrelevant or even as false alarms. This can lead to a *cry wolf effect* [4, 5], which is characterized by the ignoring of alarms that were "wrong" previously - even in critical cases. Another approach consists of adapting the timing of an alert to the capabilities of the driver. For instance, Jamson et al. [6] have proposed an adaptive FCW system that adjusts the timing of its auditory alarms according to each individual driver's brake reaction time. However, it is difficult to collect such individual reaction data in real-world driving environments because the driver's response to a hazardous event including risk cognition, judgment and averting action can vary depending on not only individual drivers but also driving situations.

To nevertheless convey an increasing collision risk early and successfully in various driving situations, we are proposing a system that assesses potentially hazardous situations in real time and continuously outputs a signal that alters its strength depending on the risk level to intuitively convey increased forward collision risk to drivers. As such it may be considered to represent an instance of so-called *likelihood alarm systems* [7]. The potential benefit of this approach is a circumvention of the tradeoff between notification timing and risk perception. Thus, a driver may perceive the signal as less annoying despite its early onset. Our goal is to encourage drivers' early voluntary risk averting action before there is a need for a more salient alert such as those used by present FCW systems. Such early and gradual risk communication may further be combined with existing salient FCWs as additional "guarantee" (see Figure 1).

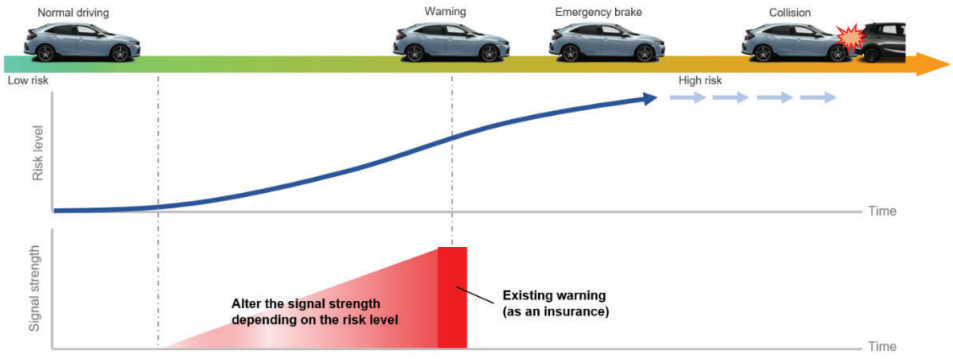

Figure 1. Illustration of the new notification that we propose.

Here we present an investigation of the effect and system acceptance of the proposed method in situations with low or medium risk levels. In particular, we try to answer the following research questions for a selection of continuous risk level communication methods:

1. Does the proposed method reduce driver reaction times to developing front collision risks compared to classical FCW?
2. Does driving safety with the proposed method increase compared to classical FCW?
3. How subjectively acceptable is the proposed method?

To address these questions, a driving simulator experiment was carried out. Because each signal modality may have a different effect on driver behavior, such as reaction times [8], four variations of Human-Machine Interfaces (HMIs) for risk level communication that utilize visual, auditory and tactile sensation were implemented into the driving simulator. The stimulus changing rate can vary depending on the risk increasing rate in our HMI concept, and thus the driver can differently perceive each HMI according to the driving situation. In this study, each HMI was tested in two different driving scenarios, which had in common that the ego vehicle eventually approached a leading vehicle, resulting in varying degrees of front risk and HMI activation. In one scenario the driver was distracted by a secondary task at the moment a sudden front risk appeared. In the second scenario the driver was indirectly motivated to produce tailgating behavior and thus become the primary source of front collision risk him or herself. Tailgating can produce an insufficient inter-vehicular distance and is one of the most severe driver-related causes of traffic accidents [9], which makes techniques that reduce such behavior particularly desirable.

METHODS

Participants

The experiment involved 17 participants (13 males and 4 females), whose ages ranged from 23 to 51. All participants had a valid Japanese driving license and reported normal, or corrected-to-normal, vision. Prior to the start of the experiment, all participants received an explanation of the contents and risks of the experiment as well as their rights and voluntarily signed a participation agreement. This study was approved by the Ethical Committee of the Honda Motor Co., Ltd.

Materials and Apparatus

The driving simulation used in this study was connected to Logicool G29 vehicle controls (Logitech Inc., CA, U.S.) for steering, accelerating, and braking. The steering wheel was mounted on a cockpit frame situated directly in front of the participant. The foot pedals were placed in a comfortable position on the floor in front of the participant. A curved monitor (effective display area: 88.0 x 36.7 cm, LG Electronics, Korea) showing an image of the driving scene was positioned 1 m from the participant. A secondary task was displayed on a touch screen (29.2 x 20.1 cm, Surface Pro, Microsoft Corporation, U.S.) positioned on the right-side of the steering wheel.

For analyzing not only the data of driving behavior from the simulator but also the gaze direction of each participant, a non-contact driver monitoring system (sampling frequency: 25 Hz, Advanced Driver Monitoring System, Seeing Machines, Australia) was mounted on the cockpit frame toward the participant's face.

For visual stimuli, a meter display (29.2 x 11.0 cm, LG Display, LG Electronics, Korea) and head-up display (18.0 x 13.6 cm, HUD622, Maxwin, Japan) were placed between the monitor and steering wheel. The velocity of the subject vehicle was also displayed on the head-up display during a driving task. Environmental sounds of the driving simulator and an auditory stimulus generated by one of the HMIs were delivered through cordless headphones (WH 1000X M3, Sony, Japan). A seatbelt component, which included a webbing, tongue plate, buckle, and retractor with a motor for generating force sensation, was installed to a pillar joined to the seat. Tactors (Vp6, Acouve Laboratory Inc., Japan) were attached inside the seat to present vibrotactile signals. Figure 2 illustrates the simulator setup.

Scenarios of the driving simulator were designed with Unity (Unity Software Inc., U.S.) and the program for HMI control was written in MATLAB / Simulink (The MathWorks Inc., U.S.).

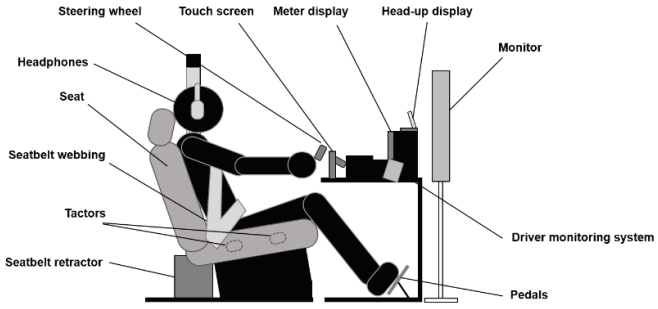

Figure 2. Schematic illustration of the driving simulator setup.

HMI Design

To convey a forward collision risk, five different HMIs, including an existing FCW as a baseline system and the other HMIs as candidates for a new notification system, were implemented into the driving simulator. For representation of continuous collision risk change, the risk estimation method is an important factor to alter the strength of stimuli. Typically, Time-to-Collision (TTC) is used for collision risk estimation. However, its value has a large variation and can quickly jump between a few seconds and infinity, especially when the subject vehicle is far away from the target vehicle or the velocities of the two vehicles are similar. To avoid sudden and extreme variations, in this study the Time to Closest Point of Approach (TCPA) was used as a risk estimation method for stimulus control. The TCPA extends the concept of the TTC by addition of a term that represents the potential deceleration of the leading vehicle at any time [10]. Effectively this makes it not just sensitive to the temporal but also the spatial distance between two cars, resulting in a less erratic variation. For two vehicles driving on the same trajectory one behind the other, the stop time t_L of the leading vehicle is given as follows:

$$t_L = -\frac{v_L}{a_L} \qquad Equation\ (1)$$

where a_L is the potential acceleration or deceleration of the leading vehicle and v_L is the velocity of the leading vehicle. When the stop time of the leading vehicle is larger than the TTC (the leading vehicle is assumed not to stop before the collision), the TCPA is given as follows:

$$TCPA = \frac{-\Delta v - \sqrt{\Delta v^2 - 2a_L d}}{a_L} \qquad Equation\ (2)$$

Where Δv is the relative velocity of the leading vehicle to the subject vehicle and d is the inter-vehicular distance between the two vehicles. In other cases, the TCPA is given as follows:

$$TCPA = \frac{d - \frac{{v_L}^2}{2a_L}}{v_S} \qquad Equation\ (3)$$

where v_S is the velocity of the subject vehicle. The outcome of this calculation, up until an upper bound, determines the stimulus strength of each HMI. The following paragraphs describe each investigated HMI.

(a) Head-up warning (HUW): This represents an existing forward collision warning system in the form of an amber ellipse (FOV: 4 x 1.5 degrees) that flashes two times on the head-up display. It is triggered when the Time-to-Collision (TTC) between the participant vehicle and another simulated vehicle falls below a 1.8 second threshold. The TTC describes the time that remains before the two vehicles collide based on their current locations and velocities.

(b) Display color: This HMI conveys the approach of the leading vehicle visually. When the TCPA falls below 4 seconds, a 9 x 10 cm red rectangle is displayed on the meter display. The color brightness continuously changes according to the TCPA value such that it increases when the TCPA becomes smaller and decreases when it becomes larger.

(c) Road sound: This HMI conveys the approach of the leading vehicle aurally. When the TCPA falls below 4 seconds, a pre-recorded sound consisting of road noise and engine sounds of the leading vehicle is played back through the headphones. Both pressure and playback speed of the sound are modulated depending on the TCPA value such that these increase when the TCPA falls (risk increase).

(d) Seatbelt tightening: This HMI conveys the approach of the leading vehicle via touch. When the TCPA falls below 4 seconds, the seatbelt webbing is retracted by the motor, resulting in seatbelt tightening. The current of the motor used for seatbelt tightening is set to depend on the TCPA value such that it increases when the TCPA falls (risk increase).

(e) Seat vibration: This HMI exemplifies another form of approach communication through touch. Even before the risk increases, vibrations are always generated by transducers inside the seat in a steady rhythm during a drive. The stimulus is designed to imitate the vibration that arises when the subject vehicle crosses a hump. When TCPA falls below 4 seconds, the interval between vibrations falls with decreasing TCPA (risk increase).

Driving Simulator Scenarios and Tasks

To evaluate the effects of each HMI on driving behavior and system acceptance, two different driving scenarios were designed on a single roadway. The roadway was made up of two lanes for each direction without stops or intersections. In both scenarios, the participants could always see a leading vehicle in front of them and a vehicle in the right-side lane. In addition, oncoming vehicles sometimes appeared (see Figure 3).

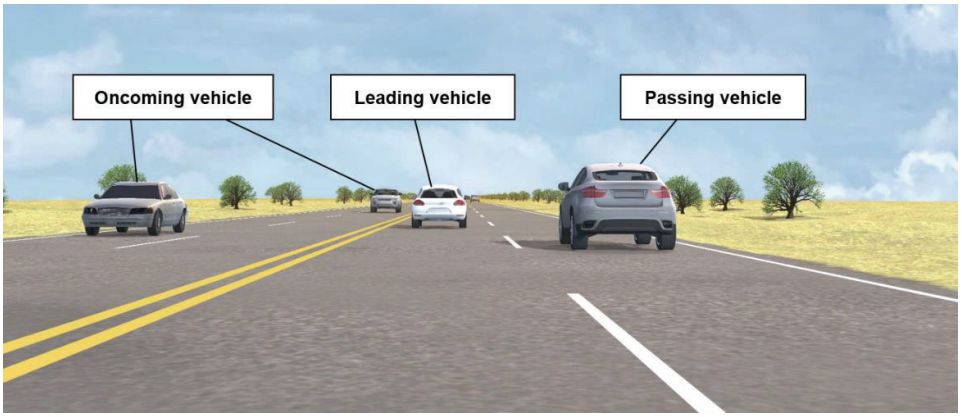

Figure 3. Driving scene of the driving simulator.

(i) Distracted driving scenario: The participants were instructed to drive at approximately 72 km/h (45 mph) without changing lanes. In addition to carrying out the driving task, the participants were instructed to perform a Surrogate Reference Task (SuRT), which required the participants to find and select the one stimulus that differed from others surrounding it [11] on the touch screen, as a secondary task (see Figure 4). A 36-pixel white circle as the target and 18-pixel white circles as the distractors were used on a black background of the touch screen and these stimuli were updated every second. For controlling driving workload between participants, a lane keeping assist system that allowed the participants to easily steer the subject vehicle was applied. After the leading vehicle continued to drive 30 meters ahead of the subject vehicle for a period selected randomly between 30 and 50 seconds, it decelerated at 0.4 G at an unanticipated timing for the participants.

(ii) Motivated tailgating scenario: Before starting to drive, the participants were required to imagine an urgent situation in which they would have to quickly drive to the airport to avoid missing their flight. The participants were instructed to continue to drive for approximately 4 minutes. Lane changes were inhibited. The leading vehicle always drove in front of the subject vehicle at approximately 72 km/h (45 mph) and sometimes slowed down at 0.08 G. This created a conflict with the driver's goal to arrive at the destination in time and may have facilitated tailgating behavior.

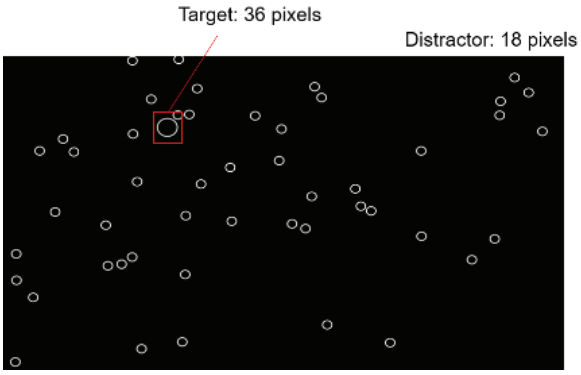

Figure 4. SuRT screen.

Procedure

The experiment consisted of three sessions: practice, evaluation in the distracted driving scenario, and evaluation in the motivated tailgating scenario. In the practice session, the participants were required to familiarize themselves with the simulation environment and the operation of the steering wheel and pedals. After this session, the participants received an explanation of how each HMI works according to hazardous events. In each evaluation session, a combination of the HUW and any one of display color, road sound, seatbelt tightening, and seat vibration HMIs was applied to investigate the effect of the proposed notification system, whereas only the HUW was additionally applied as a baseline condition (in total five HMI conditions). The participant performed five driving iterations under each HMI condition in the distracted driving scenario and a single drive under each HMI condition in the motivated tailgating scenario (see Table 1 for a list of test conditions). The test conditions were randomized in each evaluation session. After every drive, the participants answered nine questions (five for usefulness and four for affective satisfaction) on a scale from -2 to +2 (five grades) to assess subjective acceptance [12].

Table 1.

Test conditions

Session	Task	HMI	Number of drives
Practice	Driving	None	1
Evaluation in the distracted driving scenario	Driving + SuRT	Only HUW (Baseline)	5
		Display color + HUW	5
		Road sound + HUW	5
		Seatbelt tightening + HUW	5
		Seat vibration + HUW	5
Evaluation in the motivated tailgating scenario	Driving	Only HUW (Baseline)	1
		Display color + HUW	1
		Road sound + HUW	1
		Seatbelt tightening + HUW	1
		Seat vibration + HUW	1

RESULTS

Distracted Driving Scenario

In this scenario, when the leading vehicle started to decelerate, almost all the participants were looking at the touch screen to perform the secondary task and did not see the driving situation. After becoming aware of a potential danger, they suspended the secondary task, looked ahead to understand the situation, and decelerated the subject vehicle. To evaluate how soon the participant responded to the hazardous event, the time from when the leading vehicle started to decelerate until the participant looked ahead, released the gas pedal, and started to press the brake pedal was analyzed. The yaw and pitch angles of the gaze were used to determine where the participant

was looking at and distinguish between seeing the monitor (driving situation) and seeing the touch screen for the secondary task.

Figure 5 shows the reaction times with each HMI condition. A one-way ANOVA was carried out to determine whether the response of participants varied depending on the HMI. Significant main effects of HMI on time to looking ahead [$F(4, 397) = 215.48$, $p < .001$], time to gas pedal release [$F(4, 397) = 102.83$, $p < .001$], and time to brake start [$F(4, 397) = 80.38$, $p < .001$], respectively, were found. Subsequent multiple comparison tests (Bonferroni corrected) revealed that, compared to the baseline condition, the participants responded significantly sooner to the hazardous event when any of the display color, road sound, seatbelt tightening, and seat vibration HMIs were activated. The display color, seatbelt tightening, and seat vibration HMIs reduced the reaction time by a greater margin than the road sound HMI did (see Table 2).

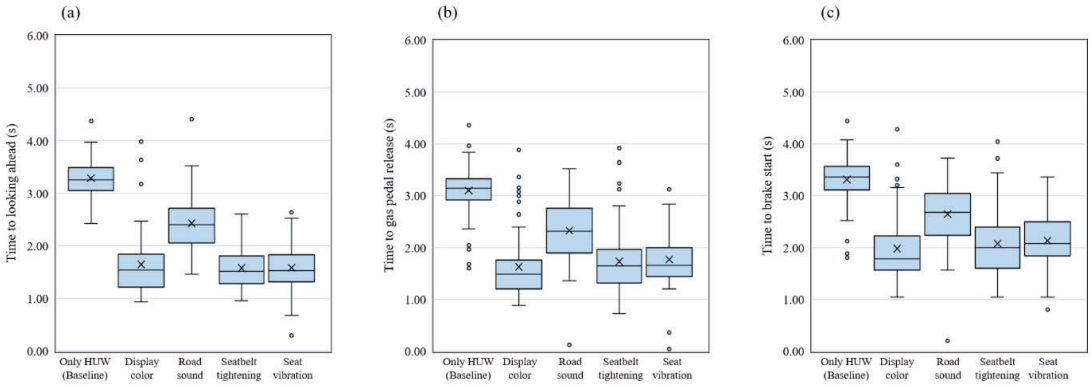

Figure 5. Comparison of reaction time of (a) looking ahead, (b) gas pedal release, and (c) brake start.

To evaluate how hazardous the situation became in consequence of reactions to the hazardous event, the minimum inter-vehicular distance after the leading vehicle started to decelerate was calculated for every drive (see Figure 6). A larger value means a longer distance to the leading vehicle, i.e., a safer situation. A one-way ANOVA and multiple comparison tests were carried out. A significant difference was found [$F(4, 397) = 82.58$, $p < .001$] and the difference in inter-vehicular distance between conditions had a similar tendency as the differences for reaction times (see Table 2).

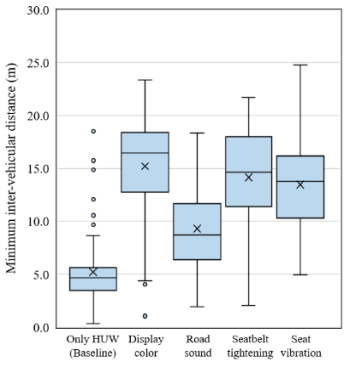

Figure 6. Comparison of minimum inter-vehicular distance.

Table 2.

Multiple comparison test of the reaction time and minimum inter-vehicular distance

HMI condition 1	HMI condition 2	Time to looking ahead		Time to gas pedal release		Time to brake start		Minimum inter-vehicular distance	
		Mean difference	p value	Mean difference	p value	Mean difference	p value	Mean difference	p value
Only HUW	Display color	1.65	p < .001***	1.47	p < .001***	1.33	p < .001***	-10.0	p < .001***
Only HUW	Road sound	.86	p < .001***	.77	p < .001***	.67	p < .001***	-4.11	p < .001***
Only HUW	Seatbelt tightening	1.71	p < .001***	1.37	p < .001***	1.23	p < .001***	-8.96	p < .001***
Only HUW	Seat vibration	1.71	p < .001***	1.33	p < .001***	1.18	p < .001***	-8.27	p < .001***
Display color	Road sound	-.79	p < .001***	-.70	p < .001***	-.67	p < .001***	5.91	p < .001***
Display color	Seatbelt tightening	.069	n.s.	-.10	n.s.	-.10	n.s.	1.06	n.s.
Display color	Seat vibration	.063	n.s.	-.14	n.s.	-.16	n.s.	1.75	n.s.
Road sound	Seatbelt tightening	.86	p < .001***	.60	p < .001***	.57	p < .001***	-4.85	p < .001***
Road sound	Seat vibration	.85	p < .001***	.56	p < .001***	.51	p < .001***	-4.16	p < .001***
Seatbelt tightening	Seat vibration	-.0057	n.s.	-.036	n.s.	-.056	n.s.	.69	n.s.

***Statically significant at $p < 0.001$, n.s. = Not significant at $p > 0.05$

Motivated Tailgating Scenario

Because the participants drove in a different manner according to their preferences in this scenario, the stimulus strength and activation frequency of HMI during 4 minutes of driving can vary depending on the individual participant. To evaluate how far the participants drove from the leading vehicle in consequence to the interaction between HMI activation and participant reaction during a drive, the average Time-Headway (THW) between the subject vehicle and leading vehicle was analyzed for every drive. Furthermore, although the HUW was set with the same threshold as in the distracted driving scenario, it was not activated due to the small deceleration of the leading vehicle in all drives of this scenario. In consequence, the drive with only HUW (baseline condition) equaled a non-HMI drive. Therefore, for evaluation of the effects of each HMI, the variation of the average THW (ΔTHW_{ave}) in each drive with each HMI relative to that of non-HMI drive was calculated for every participant. To consider the participant characterization, based on whether the average THW in non-HMI drive exceeds 1.5 seconds, the participants were divided into two groups: non-aggressive (8 participants) and aggressive (9 participants).

Figure 7 shows ΔTHW_{ave} of each participant group. With the aggressive participants, a trend for an increase of the average THW relative to the non-HMI condition is observed (ΔTHW_{ave} exceeded zero for many participants) and the effect of seat vibration HMI was significant ($t(8) = 2.17$, $p < .05$). To compare the effect of each HMI, a two-way ANOVA was carried out with the factors of HMI and participant group and the analysis indicated no main effect of both HMI and participant group (HMI: [$F(3, 60) = 1.75$, $p = .33$], participant group: [$F(1, 60) =$

9.79, p = .052]). The interaction between HMI and participant group was not significant [F(3, 60) = .39, p = .76].

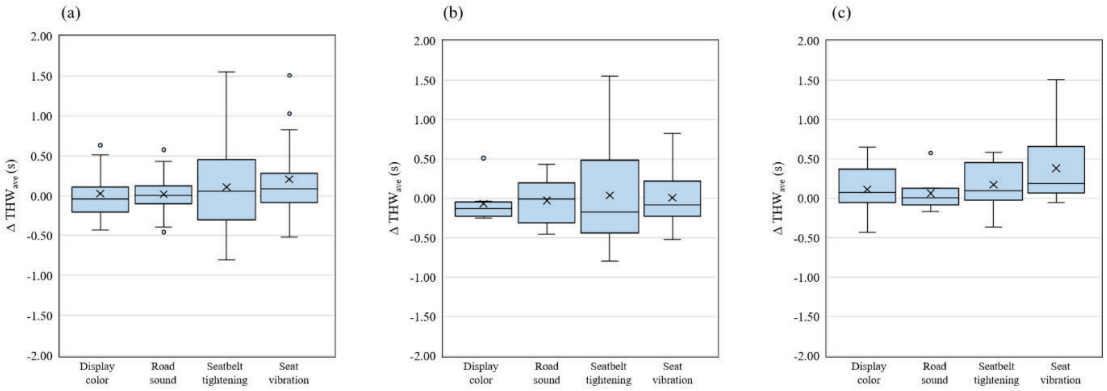

Figure 7. Comparison of ΔTHW_{ave} of (a) all participants, (b) non-aggressive participants, and (c) aggressive participants.

Subjective Ratings of HMI Acceptance

To evaluate system acceptance, based on the analysis method that Van der Laan et al. have reported [12], the average scores of five questions for subjective usefulness and four questions for affective satisfaction were calculated. Figure 8 shows the average scores of all participants on two dimensions, a usefulness scale and a satisfaction scale. Here, the error bars indicate the standard deviations between participants and there is no plot for the baseline condition in the motivated tailgating scenario due to no HMI activation. In the distracted driving scenario, the usefulness scores of the display color, road sound, seatbelt tightening, and seat vibration HMIs were significantly higher than the neutral score (zero), i.e., they were evaluated positive (display color: $t(16) = 8.95$, $p < .01$, road sound: $t(16) = 2.69$, $p < .01$, seatbelt tightening: $t(16) = 17.42$, $p < .01$, seat vibration: $t(16) = 11.30$, $p < .01$). The satisfaction scores of the baseline condition and the display color HMI were significantly higher than zero (baseline: $t(16) = 6.01$, $p < .01$, display color: $t(16) = 2.78$, $p < .01$), whereas the score of the road sound HMI was significantly lower than zero ($t(16) = 2.79$, $p < .01$).

In the motivated tailgating scenario, the usefulness scores of the display color, seatbelt tightening, and seat vibration HMIs were significantly higher than zero (display color: $t(16) = 2.23$, $p < .05$, seatbelt tightening: $t(16) = 5.59$, $p < .01$, seat vibration: $t(16) = 6.34$, $p < .01$). The satisfaction scores of all HMIs were not higher than zero, whereas the scores of the road sound and seat vibration HMIs were significantly lower than zero (road sound: $t(16) = 2.47$, $p < .05$, seat vibration: $t(16) = 1.98$, $p < .05$) and were thus evaluated negative

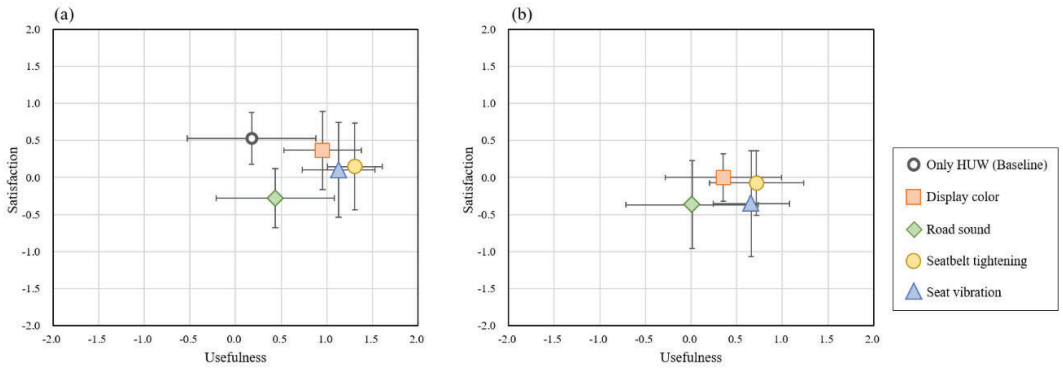

Figure 8. Subjective ratings of HMI acceptance in (a) distracted driving scenario and (b) motivated tailgating scenario.

DISCUSSION

The results demonstrate that, compared to the existing FCW, all HMIs that we proposed reduced participant reaction times to the hazardous event and the situation, accordingly, became safer in the distracted driving scenario. In this scenario, the participants were not able to directly watch the leading vehicle beginning to decelerate because of the secondary task. The proposed system has the features of early signal onset and conveying the degree of front risk. It is considered that, compared to the existing FCW, these features led the participants to become aware of the increased front risk and respond to it sooner. In particular, because the deceleration of the leading vehicle was rapid and the collision risk largely increased in this scenario, the change of stimuli from HMIs seemed to be easy to perceive.

The reaction time with the road sound HMI was longer than that with the other three HMIs (display color, seatbelt tightening, and seat vibration) and both the subjective usefulness and affective satisfaction were negative on average. For the prototype of sound source, not the beep sound but the natural road sound that the driver hears in daily driving was used to avoid annoyance. In this experiment, the participants heard both the environmental sounds of the simulator and the auditory stimulus through the headphones and it seemed that they were difficult to distinguish. However, this issue can be caused by in-vehicle sounds or environmental sounds during actual driving. It is noteworthy that the display color HMI was perceived as both useful and satisfying.

In the motivated tailgating scenario, compared to the baseline condition, only the seat vibration HMI significantly encouraged the aggressive participants to drive farther away from the leading vehicle. In this scenario, the deceleration of the leading vehicle was small and the inter-vehicular distance gradually decreased. Because the stimulus of seat vibration HMI was output in a steady rhythm all the time while driving and the interval between stimuli was changed once TCPA fell below the threshold, the participants were able to notice the start of risk increasing more clearly and respond to the increased risk sooner compared to the other HMIs. However, some participants were sensitive to the stimulus of seat vibration HMI and the average score of affective satisfaction was negative. In fact, a participant commented, “The vibration from the seat is uncomfortable for me.” These results suggest that this scenario requires both ease of perceiving the starting point of stimulus changing and

subjective acceptance, whereas there was no HMI that achieved a good balance of them on its own.

One possible reason why HMIs other than the seat vibration HMI did not achieve significant effects on safe driving in the motivated tailgating scenario is that each driving time (4 minutes) was too short to evaluate such an effect. As mentioned above, besides the small deceleration of the leading vehicle, the degree and frequency of approaching the leading vehicle depended on not only HMI effect but also individual participants in this scenario. ΔTHW_{ave} was calculated for every participant to evaluate the effect of each HMI on driving safety and this quantitative measurement showed consequences of the iterative interaction between HMI activation and participant reaction during a drive. Therefore, a long-term evaluation is considered necessary to determine whether the system has an effect on driving behavior in such a situation with lower criticality. Furthermore, HMIs can be improved to have signal onset at a degree that is not perceived as annoying even in situations with lower criticality. For instance, an HMI that is activated all the time while driving can give information on current status or a small change of risk to the driver. To focus on these issues, our research group has reported another study for long-term system evaluation on a public road [13].

From the results in the distracted driving scenario, the display color HMI seems to be the most balanced HMI between the effects on driving safety and subjective acceptance. In prototyping the display color HMI, we designed the stimulus to be perceivable in the peripheral visual field while driving. However, in cases of severely inattentive driving or drowsy driving, it is not guaranteed that a driver will always perceive such a visual stimulus. On the other hand, the seatbelt tightening and seat vibration HMIs also showed a good balance of the effect on driving safety and subjective acceptance although their satisfaction scores were not necessarily positive. Especially concerning the seatbelt tightening, a participant commented, “When I pressed the brake pedal, releasing of the tension was too late,” which may explain the low score of satisfaction. In this study, we set the thresholds for both start and stop of all HMIs to 4 seconds of TCPA. In consequence, the stimulus stopped too late after the participant started to decelerate the subject vehicle and this time gap is considered to partially lead to low scores of subjective ratings. A promising approach for this issue is to adapt the HMI stop threshold to driving behavior and driver’s attention through combination with not only driving data but also driver’s gaze data from a driver monitor camera. To minimize the gap between HMI activation and driver’s risk perception, once the situation is improved by the driver’s appropriate attention or averting action, the system should stop the stimulus immediately.

Furthermore, the multimodal effect using multiple HMIs is another interesting investigation topic. Although the stimulus strength of each HMI changes depending on the risk level in our concept, the participant may have differently perceived each HMI that utilizes different modality [14]. If we apply multiple HMIs and assign their roles according to the risk level, the information may become more subjectively relevant to the effect of averting front collision risks.

CONCLUSIONS

In this study, we have proposed a new system that assesses potentially hazardous situations in real time and continuously outputs signals with a strength that depends on the risk level. A driving simulator experiment was carried out to investigate the effects of the proposed system on driving behavior and user acceptance. The results

indicate that the proposed system reduced driver reaction times to a developing front collision risk and the situation accordingly became safer compared to a classical FCW in a driving situation with high or medium collision risk. A peripheral visual stimulus that changes the color brightness on the meter display showed high system acceptance in such a driving situation. Future work should aim to achieve more balanced HMI candidates in terms of driving safety and system acceptance in driving situations with lower criticality. We expect further insights from long-term evaluations in which drivers would have more opportunity to become accustomed to the added information sources.

REFERENCES

[1] Cicchino, J. (2017). Effectiveness of Forward Collision Warning and Autonomous Emergency Braking Systems in Reducing Front-to-rear Crash Rates. Accident Analysis & Prevention, 99(A): 142-152

[2] Yue, L., Abdel-Aty, M., Wu, Y., Ugan, J., and Yuan, C. (2021). Effects of Forward Collision Warning Technology in Different Pre-crash Scenarios. Transportation Research Part F: Traffic Psychology and Behaviour, 76: 336-352

[3] Krüger, M. (2022). An Enactive Approach to Perceputual Augmentation in Mobility. Munich: Ludwig-Maximilians-Universität Munich.

[4] Breznitz, S. (1984). Cry Wolf: The Psychology of False Alarms. Psychology Press.

[5] Getty, D. J., Swets J. A., Pickett, R. M., and Gonthier, D. (1995). System Operator Response to Warnings of Danger: A Laboratory Investigation of the Effects of the Predictive Value of a Warning on Human Response Time. Journal of Experimental Psychology: Applied, 1(1): 19-33

[6] Jamson, A. H., Lai, F., and Carsten, O. (2008). Potential Benefits of an Adaptive Forward Collision Warning System. Transportation Research Part C: Emerging Technologies, 16(4): 471-484

[7] Sorkin, R. D., Kantowitz, B. H., and Kantowitz, S. C. (1988). Likelihood Alarm Displays. Human Factors, 30(4): 445–459

[8] Lerner, N., Singer, J., Huey, R., Brown, T., Marshall, D., Chrysler, S., Schmitt, R., Baldwin, C. L., Eisert, J. L., Lewis, B., Bakker, A. I., and Chiang, D. P. (2015). Driver-vehicle Interfaces for Advanced Crash Warning Systems: Research on Evaluation Methods and Warning Signals. (Report No. DOT HS 812 208). Washington, DC: National Highway Traffic Safety Administration.

[9] Wang, J. H. and Song, M. (2011). Assessing Drivers' Tailgating Behavior and The Effect of Advisory Signs in Mitigating Tailgating. In Proceedings of the 6th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Transportation Research Board, Olympic Valley - Lake Tahoe, California, U.S., 583-589

[10] Imazu, H. (2006). Ships Collision and Measures for Safety. Journal of Japan Society for Safety Engineering, 45(6): 1-10

[11] Mattes, S. and Hallén, A. (2009). Surrogate Distraction Measurement Techniques: The Lane Change Test. In M.A. Regan, J.D. Lee, & K.L. Young (Eds.), Driver Distraction. Theory, Effects, and Mitigation (pp. 107-122). Boca Raton, FL: CRC Press.

[12] Van der Laan, J. D., Heino, A., and De Waard, D. (1997). A Simple Procedure for The Assessment of

Acceptance of Advanced Transport Telematics. Transportation Research – Part C: Emerging Technologies, 5(1): 1-10

[13] Krüger, M., Krüger, V., and Mukai, T. (2023). Real-vehicle Long-term Evaluation of Interfaces for Augmenting a Driver's Ability to Anticipate Front Risks. In Papers of the 27th International Technical Conference on the Enhanced Safety of Vehicles (ESV) (Paper Number 23-0307)

[14] Stevens, S. S. (1957). On the Psychophysical Law. Psychological Review, 64(3): 153-181

MONITORING SYSTEM OF DRIVER'S HEALTH CONDITION TO PREVENT TRAFFIC COLLISION CAUSED BY HEALTH CONDITION RISKS AND COGNITIVE DECLINE

Satoru Shinkawa
Hideki Sakai
Hiroshi Ono
Masahiro Kimura
Shigenobu Mitsuzawa
Honda R&D Co., Ltd.
Japan

Keisuke Nakamura
Honda Research Institute Japan Co., Ltd.
Japan

Noriyuki Kimura
Teruaki Masuda
Etsuro Matsubara
Department of Neurology, Faculty of Medicine, Oita University
Japan

Atsuhito Nakamichi
Shirou Yano
Takao Fujino
Usuki-City Medical Association
Japan

Emiko Segawa
Takuma Sato
Yoshitaka Nakamura
Kazuya Nagaoka
Ken Aoshima
Eisai Co., Ltd.
Japan

Paper Number 23-0226

ABSTRACT

Driving risks for elderly drivers are known to be associated with age-related diseases and cognitive decline. Furthermore, daily physical conditions such as drowsiness and fatigue also affect cognitive function and driving behavior. Therefore, in order to prevent traffic accidents involving elderly drivers, it is important to provide personal driver support that takes into consideration the effects of daily physical conditions. In this study, we explored the feasibility of a monitoring system utilizing daily physical condition data that can be assessed by wearable devices on elderly subjects. Focusing on the sleep characteristics that affect the physical condition, we found the relationship between attention function and driving behavior. As a result of the attention function evaluation by the Attention Network Test, irregular sleep time was associated with greater variation in attention function, suggesting that people with irregular sleep time had more unstable attention function. In addition, as a result of the driving behavior evaluation by the Driving Simulator Test, greater variation of the attention function was associated with the larger steering entropy and maximum acceleration of the car. These results suggest that instability of the attention function may cause the rough driving. Combined with the results of relationship between variability of sleep time and attention function, these results suggest that people with irregular sleep time are more likely to engage in rough steering and pedal operation, which may lead to sudden steering and acceleration that can cause accidents. It is also known that elderly people have problems in falling asleep and maintaining sleep than younger people. In order to eliminate traffic accidents involving elderly drivers, a support system that incorporates information on sleep habits will become more important. In recent years, the use of wearable devices has made it possible to objectively acquire daily activity and sleep data, and it is expected to utilize a wider range of daily activity data. In the future, we are planning to acquire actual vehicle driving data to understand the relationship between physical condition and driving behavior in more detail.

INTRODUCTION

Honda aims to achieve zero fatalities in traffic accidents involving Honda motorcycles and automobiles worldwide by 2050. The number of accidents involving elderly drivers is increasing along with the acceleration of the aging society in the world, and it is recognized as social-level problem [1]. Driving risk for elderly drivers is known to be associated with age-related diseases (such as cardiovascular disease, diabetes, sleep disorders, and other lifestyle-related diseases) and cognitive decline [2-3]. Furthermore, daily activity indicators that affect physical condition, such as sleep, exercise, and fatigue have also been reported to affect cognitive function and driving [4-6]. Thus, monitoring of daily activity is considered to be important for personal driver support in addition to conventional medical checkup data. In this study, we explored the feasibility of a monitoring system utilizing daily activity data that can be assessed by wearable devices. We report the relationship between daily activity indicators, attention function, which is important for safe driving, and the changes in driving behavior in elderly drivers.

METHODS

Ethics

This study was conducted in accordance with the Declaration of Helsinki and ethical guidelines for epidemiology research authorized by the Japanese government, and it was approved by Institutional Review Board of Oita University (Clinical Review Board Approval No. 2355-C45), of Eisai Co., Ltd. (Registration No. 2022-0793), and of Honda R&D Biotechnology Ethics Committee (No. 99HM014H). Written informed consent was obtained from all participants.

Study design

Elderly residents of Usuki City (Oita Prefecture, Japan) were asked to wear a smartwatch (VENU 2S, Garmin Ltd., Olathe, Kansas, USA) for two weeks to obtain daily activity data. Each participant was also administered the Attention Network Test (ANT) and the Driving Simulator Test (DST) four times, each at least two days apart during the two weeks. First trial of each assessment was served as practice session.

Participants

ANT was performed on 24 participants (age range 75-86 years, mean = 79.54, standard deviation (SD) = 3.39), of whom 16 (age: mean = 73.32, SD = 4.97) completed the DST. All participants held a driver's license (Table. 1).

Table 1. Characteristics of the participants

Characteristic	All, n = 24	DST completed, n = 16
Age (years), range	75 - 86	75 - 86
Age (years), mean (SD)	79.54 (3.39)	78.81 (3.52)
Sex (male), n (%)	18 (75)	13 (81.25)
BMI (kg/m2), mean (SD)	24.24 (2.38)	24.28 (1.88)
Warning[a], n (%)	3 (12.5)	1 (6.25)
Crush[a], n (%)	1 (4.16)	0 (0)

a Self-reported number of the experiences in the past 2 years.

Wearable sensor data

The daily activity and related data were assessed using a smartwatch, VENU 2S (Garmin), and calculated by built-in algorithms developed by Garmin [7-9]. Sleep/wake parameters were estimated based on time data and RR-intervals (RRI), heart rate variability (HRV), respiration rate and wrist/body movement data assessed by optical sensor combined with accelerometer data. The number of steps were estimated by accelerometer data. Stress levels (0–100) were estimated primarily using a combination of RRI and HRV data.

Attention function assessment

Using the ANT, we evaluated three attentional networks that constitute the human attentional function; the alerting network, which is thought to be involved in maintaining arousal, the orienting network, which is thought to be involved in selective attention to sensory stimuli, and the executive network, which is thought to be involved in resolving conflicting information [10]. The test program was created and executed using Unity (Unity Technologies, San Francisco, USA) and displayed on a monitor with a resolution of 1920 × 1080.
The test used three Cue conditions (No cue, Center Cue, and Spatial Cue). In the Spatial Cue condition, an asterisk indicated the location of the next target. Participants were asked to identify the direction of the arrow quickly and accurately in the middle of the target and press the corresponding keyboard button (left button for target arrow pointing left, right button for target arrow pointing right). Three target types were used for the evaluation (Neutral, Congruent, and Incongruent) (Figure 1).

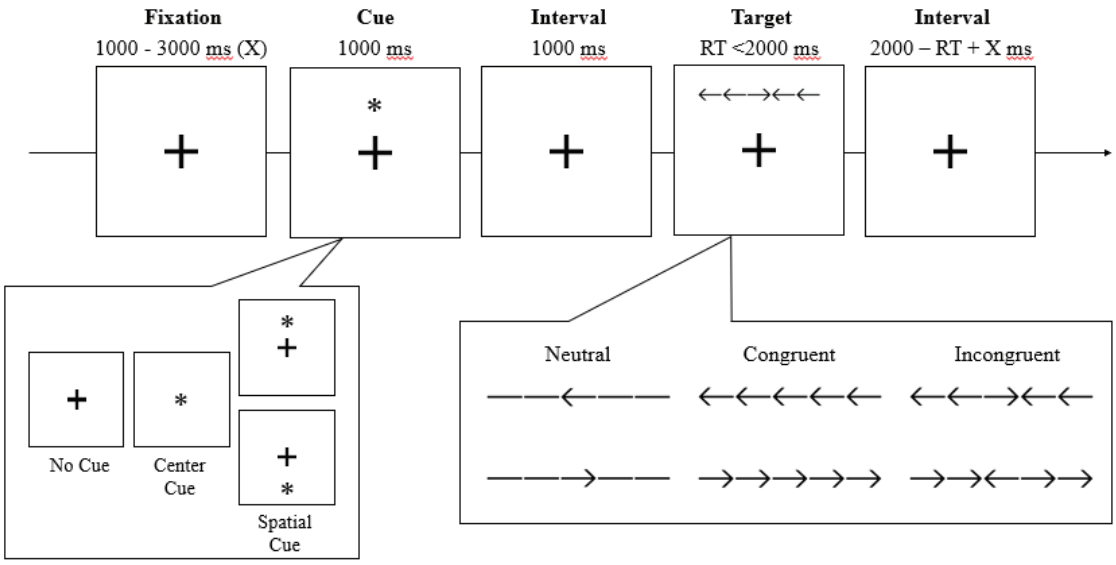

Figure 1. Schematic of Attention Network Test.

Scores for each of the three attentional functions were calculated using the following formula based on the Reaction Time (RT) of the two related parameters.

Alerting Effect = RT (No cue) − RT (Center Cue)
Orienting Effect = RT (Center Cue) − RT (Spatial Cue)
Executive Effect = RT (Incongruent) − RT (Congruent)

Higher Alerting Effect or Orienting Effect scores reflect the ability to use cues more efficiently, indicating better alerting or orienting attention function. On the other hand, higher Executive Effect scores indicate poorer function, indicating less ability to resolve conflicts between discrepant perceptual information.

Driving behavior evaluation

Driving behavior was evaluated using HONDA Safety Navi (Honda Motor Co., Tokyo). For this test, a scenario consisting of an urban area and a mountain road was created and conducted.

Participants underwent a 3-minute practice session followed by 3 x 5-minute main test sessions. If a subject reported feeling simulator sickness during the practice session, further testing was immediately stopped. The test included scenarios depicting situations such as pausing in an urban area, traffic lights, parked vehicles, and car jumps at intersections. The mountain road also consisted of a Winding Road consisting of an ascent and a descent (Figure 2). Input variables such as steering, accelerator, and brake recorded every 10 ms on the DST control PC and data on output changes such as vehicle coordinates and speed were acquired and used for analysis. Steering entropy was calculated according to Nakayama *et al.* (1999) [11].

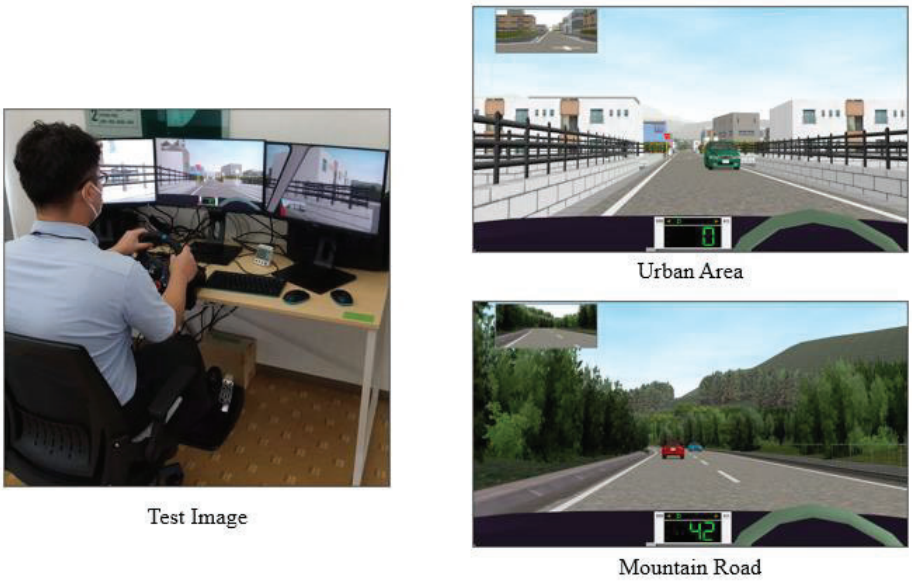

Figure 2. Schematic of Driving Simulator Test.

Statistical Analysis

Statistical analysis performed using python and scipy. The Pearson's correlation analysis was used to evaluate the relationships between daily activity data, attention function, and driving behavior. In all cases, $p < 0.05$ was considered statistically significant.

RESULTS

Daily activity and attention function

Figure 3 shows data on total sleep time, number of steps, and stress level obtained using a smartwatch in each participant. We first tried to examine the relationship between sleep and attentional function based on previous reports [12-14].

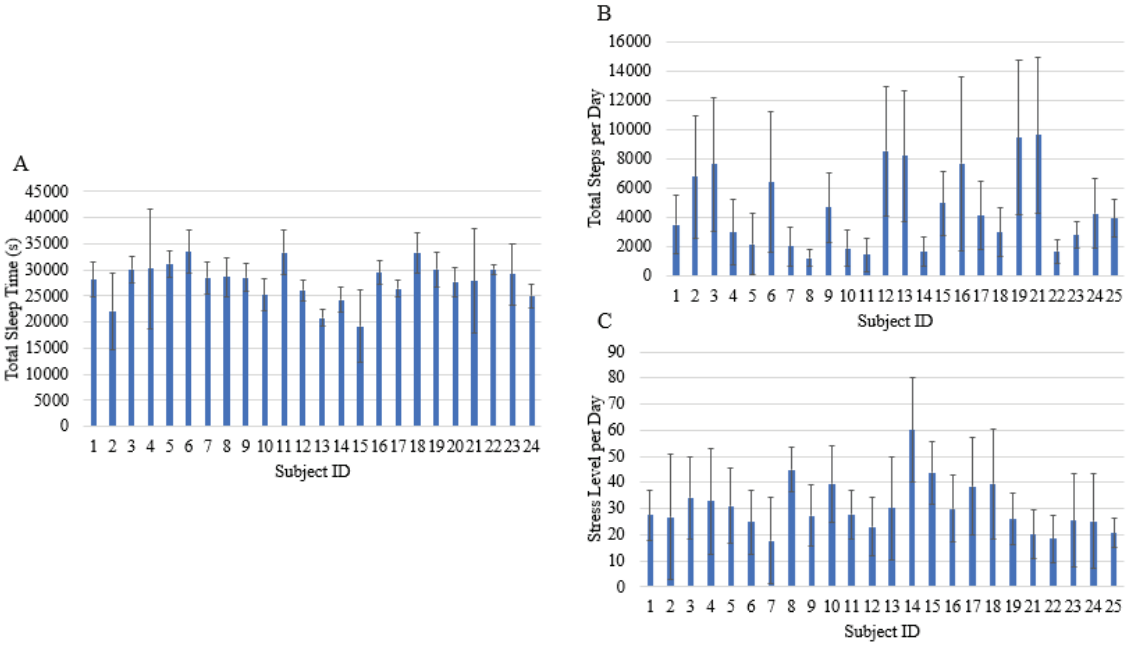

Figure 3. Daily activity log assessed by smartwatch. Each value represents mean ± SD.

Two sleep parameters, mean sleep duration and sleep duration variability (Figure 3A), were used to compare with the results of the ANT. The results showed that sleep duration variability is highly correlated with attentional function (Figure 4D-F) than mean sleep duration (Figure 4A-C) in all three attentional functions, alertness, orientation, and executive functions. These results suggest that irregular sleep duration may cause the unstable attentional function.

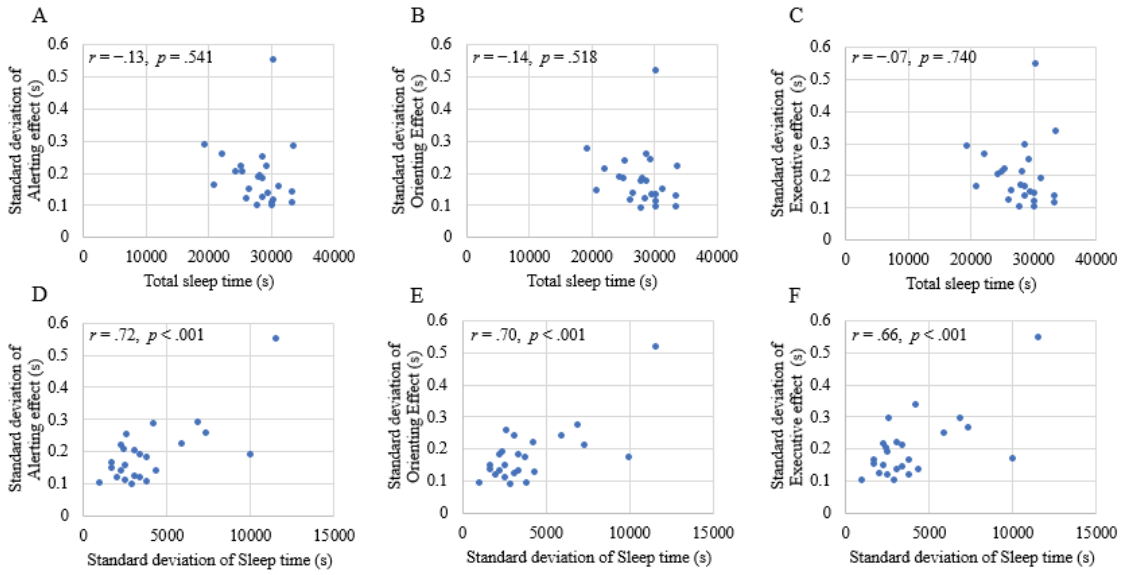

Figure 4. Relationship between sleep parameters and attention function.

Attention function and driving behavior

Driving behavior analysis using DST was conducted to examine the effects of attention function instability on driving behavior. This study focused on steering and pedaling on mountain roads.

Steering entropy was used to evaluate steering operation [11]. We found the significant correlations between steering entropy and variability for all three attention functions, alertness, orientation, and executive functions (Figure 5A-C). For pedaling, the average of the maximum acceleration for each of the nine tests was used for the evaluation. As a result, we also confirmed the relationships between maximum acceleration and variability of all three attention functions (Figure 6A-C). In short, both higher steering entropy and maximum acceleration were associated with greater variation of attention functions.

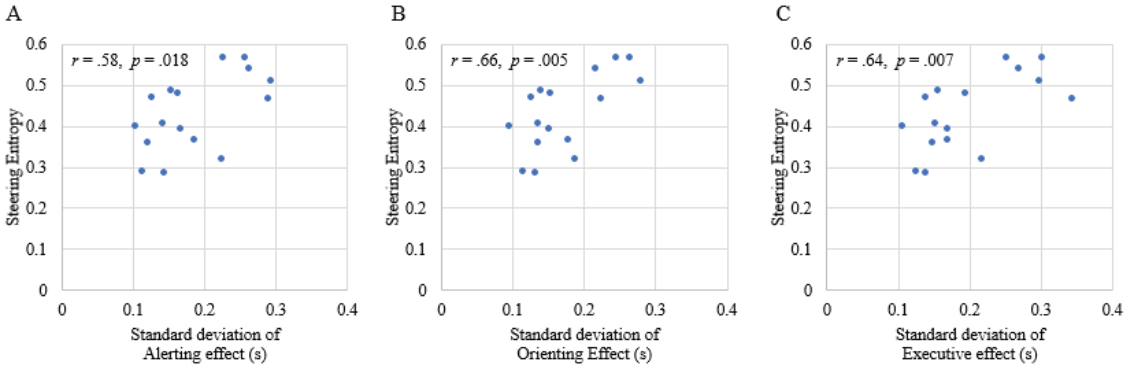

Figure 5. Relationship between attention function and steering entropy.

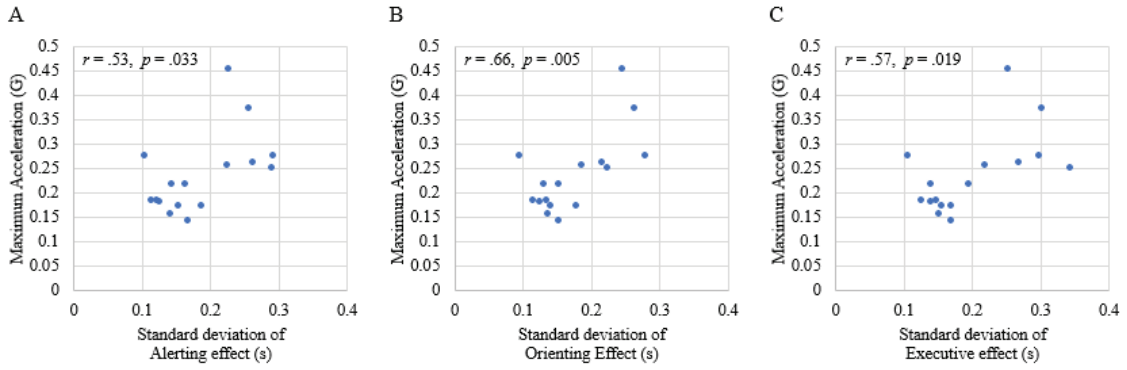

Figure 6. Relationship between attention function and maximum acceleration.

DISCUSSION

The present study focused on relationships between sleep, attention function, and driving behavior. We found that the variability of daily sleep duration had more influence on attentional function than the mean sleep duration (Figure 4). Continuous sleep loss and circadian rhythm disruption are reported to cause poor attention function [15-16]. An association between disturbed circadian rhythm and cognitive dysfunction has been also reported [17], and thus, we are planning to further explore the relationships between irregular sleep time, circadian rhythm disturbances, and instability in attention function.

The results on attention function and driving behavior showed that those with unstable attention function tended to have larger values of steering entropy and maximum acceleration on mountain roads (Figure 5 and 6). Combined with the results of relationship between variability of sleep time and attention function, these results suggest that people with irregular sleep time are more likely to engage in rough steering and pedal operation, which may lead to sudden steering and acceleration that can cause accidents.

It has been reported that elderly people have sleep related problems such as difficulties in falling asleep and to maintaining sleep than for younger people [18]. In addition, the frequency of sudden and unexpected driving risk is thought to be increased in the elderly than in younger people, and it can be hypothesized due to a combination of sleep problems, age-related cognitive decline, and unstable attentional function. Therefore, those relationship should be carefully addressed to develop the driver support systems utilizing daily physical condition data to prevent accidents among the elderly.

In recent years, wearable devices have made it possible to easily obtain data on daily activities. Further accumulation of a wider range of daily activity data in the future will make it possible to study the effects of daily driving risk in more detail. As a next step, we are planning to conduct correlation analysis with actual driving data to evaluate the importance of monitoring system in real world.

Currently, as advanced driver assistance systems related to the risk of physical condition while driving, we are conducting research on medical emergency stop systems using driver monitoring cameras, driver availability/sleepiness monitoring systems, etc. In order to further improvement of the effectiveness of these systems, highly accurate estimation of the driver's condition is essential. The use of drivers' daily data as input data for this purpose is expected in the future.

CONCLUSION

This study examined the potential of the driver support system to utilize daily activity data. It was found that among the elderly, irregular sleep time may cause the instability of the attentional function. Furthermore, the results of driving behavior analysis revealed that unstable attention function may lead to rough steering wheel and pedal operation. These findings indicate the possibility of prediction of driving risk based on daily activity data.

REFERENCES

[1] Matsuyama, T., Kitamura, T., Katayama, Y., Hirose, T., Kiguchi, T., Sado, J., Kiyohara, K., Izawa, J., Okada, N., Takebe, K., Watanabe, M., Miyamoto, Y., Yamahata, Y., & Ohta, B. (2018). Motor vehicle accident mortality by elderly drivers in the super-aging era: A nationwide hospital-based registry in Japan. Medicine, 97(38)

[2] Falkenstein, M., Karthaus, M., & Brüne-Cohrs, U. (2020). Age-related diseases and driving safety. Geriatrics, 5(4), 80.

[3] Fraade-Blanar, L. A., Ebel, B. E., Larson, E. B., Sears, J. M., Thompson, H. J., Chan, K. C. G., & Crane, P. K. (2018). Cognitive decline and older driver crash risk. Journal of the American Geriatrics Society, 66(6), 1075-1081.

[4] Lowrie, J., & Brownlow, H. (2020). The impact of sleep deprivation and alcohol on driving: a comparative study. BMC public health, 20(1), 1-9.

[5] Kirk-Sanchez, N. J., & McGough, E. L. (2014). Physical exercise and cognitive performance in the elderly: current perspectives. Clinical interventions in aging, 9, 51.

[6] Guo, M., Li, S., Wang, L., Chai, M., Chen, F., & Wei, Y. (2016). Research on the relationship between reaction ability and mental state for online assessment of driving fatigue. International journal of environmental research and public health, 13(12), 1174.

[7] Freitag, L., Clijsen, R., & Hohenauer, E. (2021). The Physiological and Perceptual Stress Response During Data Collection in Altitude: A Single-Case Report of A Healthy Researcher. Archives of Clinical and Medical Case Reports, 5(6), 821-831.

[8] Polysomnography, P. S. G. (2019). A Sleep Analysis Method Based on Heart Rate Variability.

[9] Myllymaki, T. (2014). Stress and recovery analysis method based on 24-hour heart rate variability. Firstbeat Technologies Ltd.

[10] Fan, J., McCandliss, B. D., Fossella, J., Flombaum, J. I., & Posner, M. I. (2005). The activation of attentional networks. Neuroimage, 26(2), 471-479.

[11] Nakayama, O., Futami, T., Nakamura, T., & Boer, E. R. (1999). Development of a steering entropy method for evaluating driver workload. SAE transactions, 1686-1695.

[12] Chee, M. W., & Chuah, Y. L. (2007). Functional neuroimaging and behavioral correlates of capacity decline in visual short-term memory after sleep deprivation. Proceedings of the National Academy of Sciences, 104(22), 9487-9492.

[13] Chua EC-P, Fang E, Gooley JJ (2017) Effects of total sleep deprivation on divided attention performance. PLoS ONE 12(11): e0187098. https://doi.org/10.1371/journal.pone.0187098

[14] Lo, J. C., Groeger, J. A., Santhi, N., Arbon, E. L., Lazar, A. S., Hasan, S., ... & Dijk, D. J. (2012). Effects of partial and acute total sleep deprivation on performance across cognitive domains, individuals and circadian phase.

[15] Van Dongen, H., Maislin, G., Mullington, J. M., & Dinges, D. F. (2003). The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. Sleep, 26(2), 117-126.

[16] McGowan, N. M., Uzoni, A., Faltraco, F., Thome, J., & Coogan, A. N. (2020). The impact of social jetlag and chronotype on attention, inhibition and decision making in healthy adults. Journal of Sleep Research, 29(6), e12974.

[17] Wright Jr, K. P., Hull, J. T., Hughes, R. J., Ronda, J. M., & Czeisler, C. A. (2006). Sleep and wakefulness out of phase with internal biological time impairs learning in humans. Journal of Cognitive Neuroscience, 18(4), 508-521.

[18] Chaput, J. P., Dutil, C., & Sampasa-Kanyinga, H. (2018). Sleeping hours: what is the ideal number and how does age impact this?. Nature and science of sleep, 10, 421.

THE IMPACT ON DRIVING PERFORMANCE FROM GRADED COGNITIVE LOAD WITH VISUO-SPATIAL AND PHONOLOGICAL PROCESSING OF VISUAL AND AUDITORY INPUT.

Hiroko Adachi, Yusuke Muramatsu, Naohiro Sakamoto, Hajime Ohya
Honda R&D Co., Ltd.
Japan

Paper Number 23-0229

ABSTRACT

The majority of human factors in traffic accidents are the result of cognitive error. Errors of cognition are produced by the relationship of the cognitive load of the traffic environment and vehicle interior environment with the driver's information processing. The cognitive load while driving is made up of the loads from the sense organs of sight and of hearing. The resources used for processing of visuo-spatial information and phonological information are independent, and it has been proposed that each processing resource has its capacity. It has been reported in previous research that when the cognitive load increases, driving becomes unstable. On the other hand, it has been reported in other research that when the cognitive load becomes high, driving becomes stable. Considering that cognitive load has been reported as an influence that both increases and decreases performance, it is conceivable that performance varies with the type and magnitude of the cognitive load from each category of information, and that a moderate degree of load exists under which performance reaches its highest level. For this paper, a driving simulator was used to study the influence on driving performance caused by graded cognitive load from the visuo-spatial process and phonological process of input from the sense of sight and sense of hearing. In testing, drivers drove on a course with a series of gentle curves while responding to n-back tasks that use visual/visuo-spatial process and auditory/phonological process. The result was that in the case of n-back tasks using visual/visuo-spatial processing, driving performance was diminished as the difficulty of the n-back task increased. However, in the case of n-back tasks using auditory/phonological processing, driving performance did not change when the difficulty of the n-back task increased. Also, although the load under which performance reaches its highest level was not determined, it was confirmed that auditory n-back tasks do have loads under which performance tends not to change. This is thought to be because the visual/visuo-spatial process used in driving and other information processes tend not to influence each other, while the same information processes did interfere with each other. The conclusion is that, in order to maintain stable driving performance, it can be considered important that the cognitive load on the driver does not interfere with the processing of visual/visuo-spatial information while driving.

INTRODUCTION

Honda aims for zero fatalities in traffic accidents involving Honda motorcycles or automobiles worldwide by 2050. Of human factors resulting in fatal accident cases in Japan, 68.8% are due to cognitive error such as not confirming safety, intrinsic absence of attention ahead, extrinsic absence of attention ahead, and so on [1]. Driving scenes have various different items of information that drivers should perceive. Drivers perceive not only the lane they are driving in, but also traffic signs, traffic participants, the vehicle interior environment, and so on. The traffic environment information that is being perceived is constantly changing as driving proceeds. On country roads, there is less traffic environment information, and distracted driving and other such errors cause pedestrians to be overlooked. On the other hand, there is more traffic environment information in urban areas, and information overload can result in delay in noticing pedestrians. These kinds of errors of cognition occur through the relationship between the cognitive load from the traffic environment and vehicle interior environment and the driver's information processing [2]. Information processing of the cognitive load takes place by means of the working memory. Alan Baddeley and Graham Hitch have proposed a working memory that is considered to be made up of a visuo-spatial sketch pad and phonological loops [3]. According to the multiple resource model proposed by Christopher D. Wickens, the resources for processing visuo-spatial information and those for phonological information are independent, and each processing resource is proposed to have its capacity [4]. These suggest that different information processing domains tend not to influence each other while information processing domains that are the same interfere with each other. During driving, visual/visuo-spatial processing is considered to make up most of the processing. For that reason, the input of visuo-spatial process information during driving can be considered likely to influence driving.

There have also been a number of research reports that cognitive load has an influence on driving performance. According to Uno, et al., when cognitive load rises to a high level, variation in lateral movement of the vehicle becomes larger when following a vehicle ahead and driving is said to become unstable [5]. On the other hand, according to Johan Engstrom, et al., when cognitive load rises to a high level, there are fewer standard deviations of lateral position during lane keeping tasks, and driving is reported to be stable [6]. In other words, cognitive load can be considered to have an influence that both increases and decreases performance. In addition, the Yerkes–Dodson law hypothesized that when arousal and load increase, performance increase, but when arousal and load increase too much, then performance decreases instead [7]. Based on this hypothesis, it can be considered possible that there is a degree of load under which performance reaches its highest level.

From these proposals and hypotheses, it can be considered that cognitive load has an influence on driving performance, and further that it is possible that the way performance is influenced when the load is in a separate domain that does not overlap with processing for driving differs from when it is in the same domain that does overlap with that processing. It can also be considered that there is a moderate degree of load under which performance reaches its highest level. For this paper, a driving simulator (DS) was used to study the influence on driving performance from the graded cognitive load of visuo-spatial processing and phonological processing of input from the sense of sight and sense of hearing. This testing was reviewed and passed by Bioethics Committee Meetings for Hondas R&D activities. (Bioethics Committee No. 99HM-020H)

METHOD

System

Driving task (lane center tracing task) Driving performance was evaluated by use of the DS. The test participants drove in the middle of the roadway on a winding course made up of a series of curves. In order to continuously place a cognitive load on them and also to continuously measure their driving performance, the curvature of the winding course curves was gradually changed in a design intended to require drivers to be constantly steering.

Cognitive tasks (n-back tasks) In order to control the cognitive load, n-back tasks were used as the cognitive tasks. N-back tasks are typical tasks used to impose a cognitive load, and there are many cases of their use in other research as well [8]. In the n-back task, the subject is given numbers or letters at regular intervals, and if the number or letter is the same as the one n times before, the subject is to respond with a positive sign (○) and if it is different, then the subject is to respond with a negative sign (×) (Fig. 1). In n-back tasks, using a higher number n makes the load larger, and so it is possible to control the load. For the present research, a preliminary experiment was conducted with reference to previous research and a number or letter was presented at 4-second intervals [9]. A single n-back trial consists of a total of 37 issues with seven right answers. What was presented in the n-back tasks was a graphic that uses the visual sense to employ visual/visuo-spatial processing that overlaps with the information processing during driving, and a voice that uses the sense of hearing to employ auditory/phonological processing that does not overlap with driving (Fig. 2). The graphic, taking previous research as a reference [10], was a figure with a 4×4 grid, as shown in Fig. 3, containing red circles at two locations. Nine types of these figures were prepared and displayed randomly. Figure 4 shows where the figure was displayed. It was confirmed through advance testing that this display location does not overlap with the location of the line of sight during driving and the location is also easy to check when driving through left and right curves, and this configuration was determined. For the voice, nine integers from one to nine were played back randomly. In order to control the conditions, headphones were worn in the same way during the graphic n-back tasks. During the graphic n-back tasks, however, no voice sound was played, and the headphones were muted.

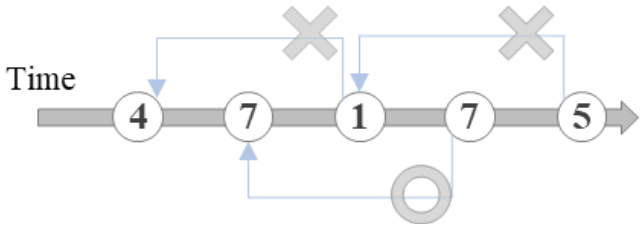

Figure 1. Example of 2-back task.

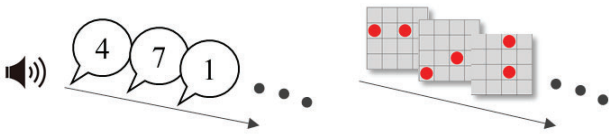

Figure 2. Content presented in n-back tasks.

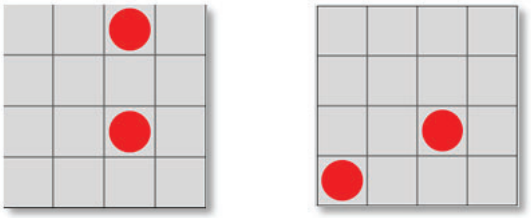

Figure 3. Example of graphic in graphic n-back task.

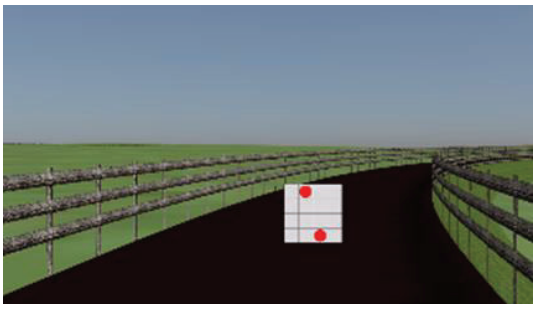

Figure 4. Graphic n-back task display location.

Subjective evaluation In order to confirm the extent to which test participants engaged in the tasks, the test participants subjectively evaluated the degree of task achievement and the degree of effort. For the degree of achievement, evaluations were made of the n-back tasks and driving tasks severally. The evaluations used the visual analogue scale (VAS), where a score of zero indicates a task was extremely unachieved and a score of 100 indicates it was extremely achieved. For the degree of effort, a score of zero indicates extremely little effort was made and a score of 100 indicates extreme effort was made. These two subjective evaluation indices were evaluated on a range from zero to 100.

Test Procedure

A total of 18 men and women (average age 37.6 years) whose consent for testing was obtained participated in the tests. This research was conducted with review and approval by Bioethics Committee Meetings for Hondas R&D activities (Bioethics Committee No. 99HM-020H). Test participants were also given an explanation of the purpose of the research using the consent forms and their cooperation was requested. Test participants whose consent was obtained also signed the consent forms.

Figure 5 shows the flow of the test as a whole. To accustom the test participants to the n-back tasks, familiarization with n-back tasks was conducted in advance using 3-back tasks. The test participants drove a winding course in the DS environment. They were instructed to keep their position in the middle of the course roadway as much as possible. The driving speed was set at a fixed 50 km/h, and test participants only operated the steering. To accustom the test participants to driving in the middle of the roadway, they were fully familiarized with the DS steering operation and the appearance of the field of vision. For this familiarization, both the driver view (Fig. 6) and the overhead view (Fig. 7) were displayed together, and the test participants were made able to drive in the middle of the course roadway while checking the view of the subject vehicle's position and the field of vision in

the driver view. After the test participants' driving familiarization, the driving for the test itself was done. The test itself was conducted in 10 trials in order to have a mixture of each type of n-back task. The graphic and sound n-back tasks were both conducted under the five conditions of no n-back, 1-back, 2-back, 3-back, and 4-back. However, the order of the n-back tasks was made random in order to do away with order effects. The test of n-back tasks under five conditions was taken as one set, and two sets were conducted. The driving for a single trial consisted of a reference interval in which only driving is done and an evaluation interval in which n-back tasks and driving are done simultaneously. In order to keep responses to n-back tasks in the evaluation interval from interfering with driving, test participants responded using the paddle shift button on the steering wheel instead. After the driving ended, test subjects made subjective evaluations of their own driving and the n-back tasks, which completed one trial.

Graphic n-back tasks and sound n-back tasks were conducted on separate days.

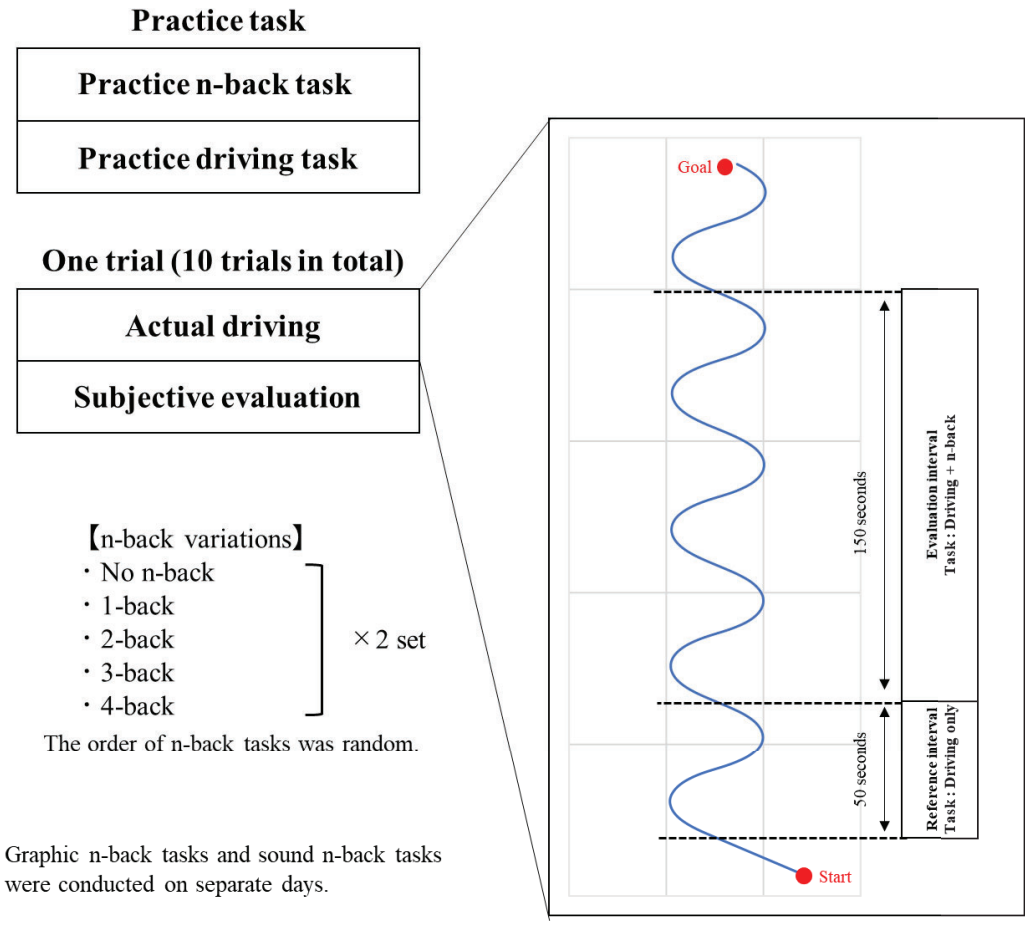

Figure 5. Test flow.

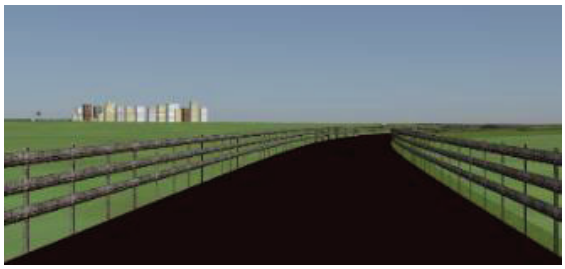

Figure 6. Display when driving in the middle of the course roadway (driver view).

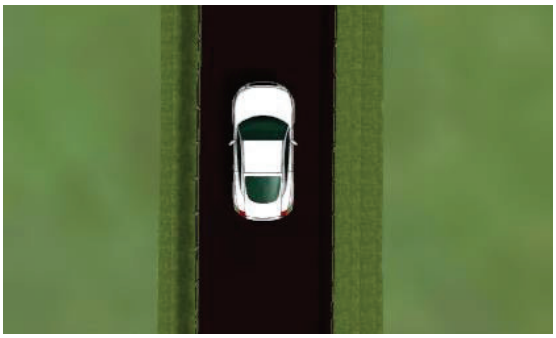

Figure 7. Overhead view during familiarization drive.

ANALYSIS METHOD

Evaluation of Cognitive Tasks (n-back Task Response Reaction Time/Number of Right Answers)

In order to check whether the cognitive load on test participants was successfully being imposed gradually, the performance of the n-back tasks was evaluated. For performance, the n-back task response reaction time and the number of right answers were calculated. The response reaction time was taken to be the time from when the n-back task was presented until the response was made by the paddle shifter. The number of right answers was taken as the number of responses for which the response given was that the number or the graphic presented was the same as that n times before. The number of right answers can be for a maximum of seven tasks.

Driving Performance Evaluation (Lateral Offset from Course)

In order to confirm the influence of n-back tasks on driving, the reference interval of each trial was used as a standard in evaluating the driving performance in the evaluation interval. It was reported from previous research that, compared to when no cognitive load is imposed, variation in lateral movement of the vehicle becomes larger when following a vehicle ahead when a cognitive load is imposed. This can be considered to be because the number of times of significant deviation from the course increases due to drivers' inability, as a result of the cognitive load, to notice the lateral offset until it has grown large. In order to evaluate the degree of which the task of driving in the middle of the course is achieved, the position of lateral offset from the middle of the course roadway was calculated (Fig. 8). Fig.9 shows the calculation method of the amount of lateral offset for one trial. The average lateral offset position of the reference interval in the data for all trials by that individual was reduced from the lateral offset position data of the evaluation interval of one trial. The absolute value of this subtracted

lateral offset position data of the evaluation intervals was averaged, and the amount of lateral offset was calculated for one trial. As there are two trials for each n-back task, the first and second trials are averaged and these amount of lateral offsets were taken as the driving performance evaluation values for each n-back task.

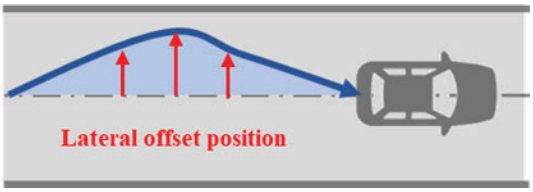

Figure 8. Lateral offset position from the middle of the course roadway.

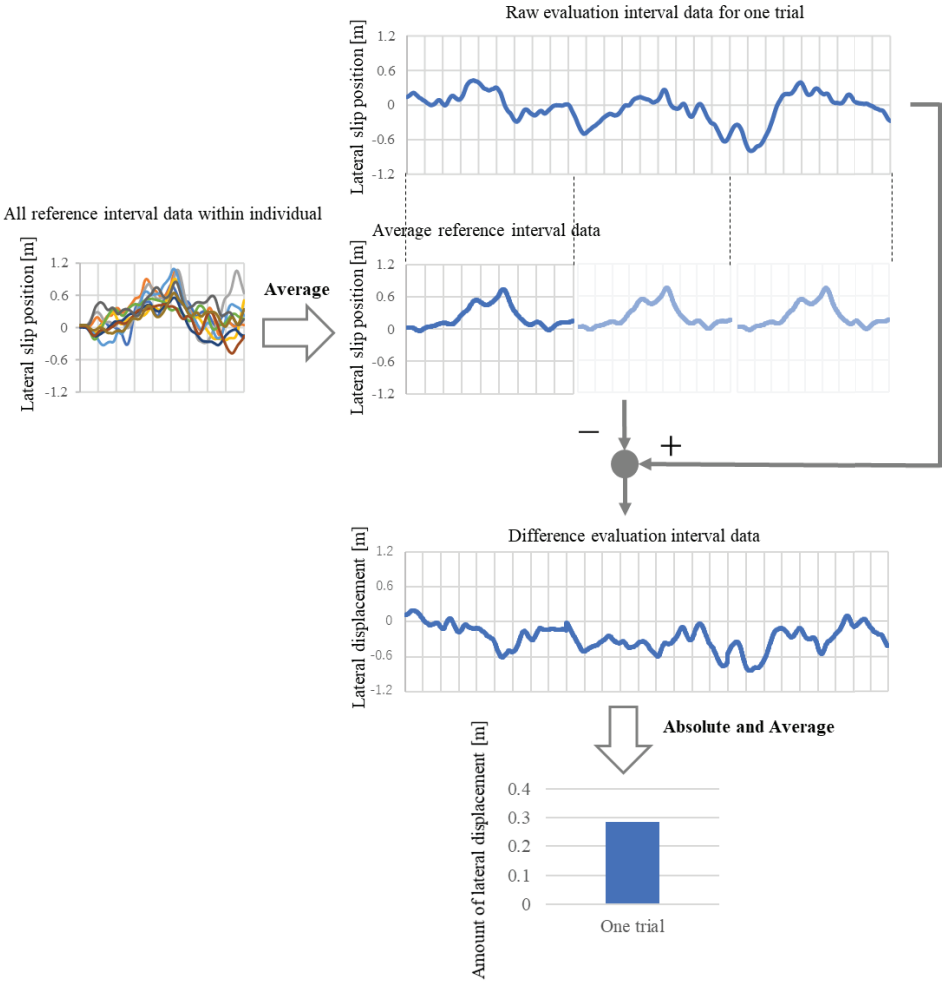

Figure 9. Amount of lateral offset calculation method for one trial.

RESULTS

Cognitive Task Evaluation (n-back Response Reaction Time/Number of Right Answers and Subjective Evaluation)

Figures 10 and 11 show the response reaction times and the number of right answers to the visual and auditory n-back tasks that were the cognitive tasks. For both the visual and auditory n-back tasks alike, increasing the size of the number n resulted in slower reaction times and the number of right answers decreased. Differences were apparent in the visual and auditory n-back task scores.

In addition, Fig. 12 and Fig. 13 show the average subjective achievement and subjective effort for all test participants in each n-back task. The subjective achievement results show that as the n numbers grew higher, the tasks were not achieved for both visual and auditory n-back tasks. The n-back task scores and subjective achievement evaluation suggest that the n-back tasks were successful in gradually changing the test participants' cognitive load. In both the visual and auditory n-back tasks, the subjective effort indicated that the greatest effort was achieved with the 2-back tasks.

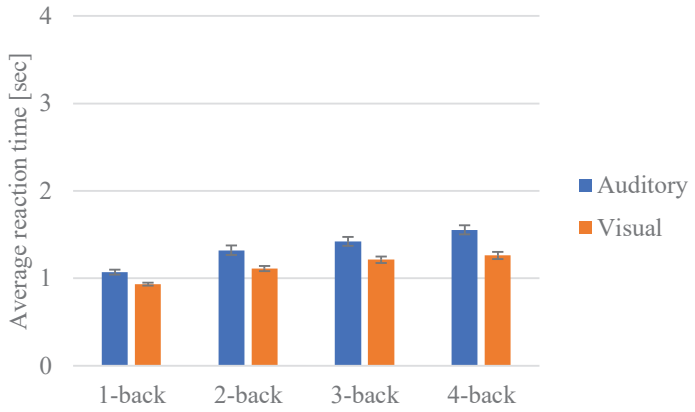

Figure 10. Average reaction times in visual and auditory n-back tasks.

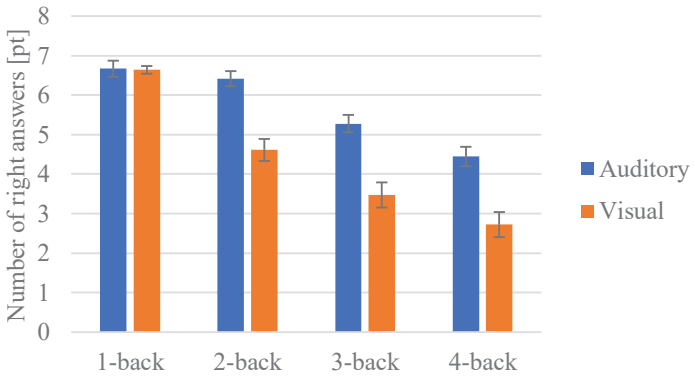

Figure 11. Average number of right answers in visual and auditory n-back tasks.

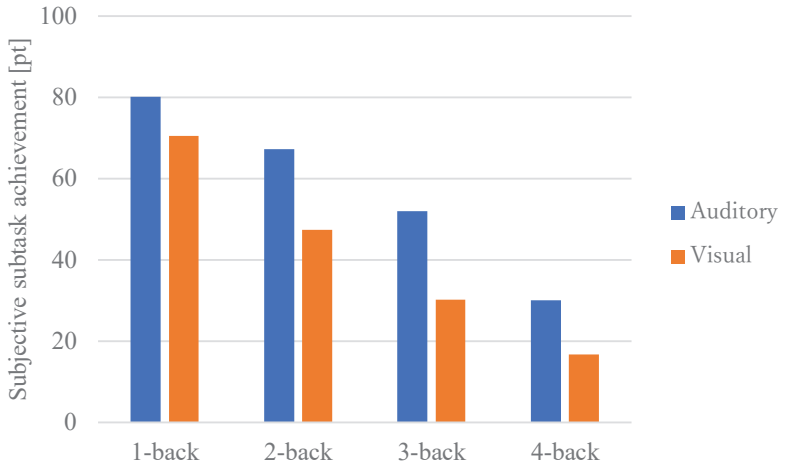

Figure 12. Subjective achievement in visual and auditory n-back tasks.

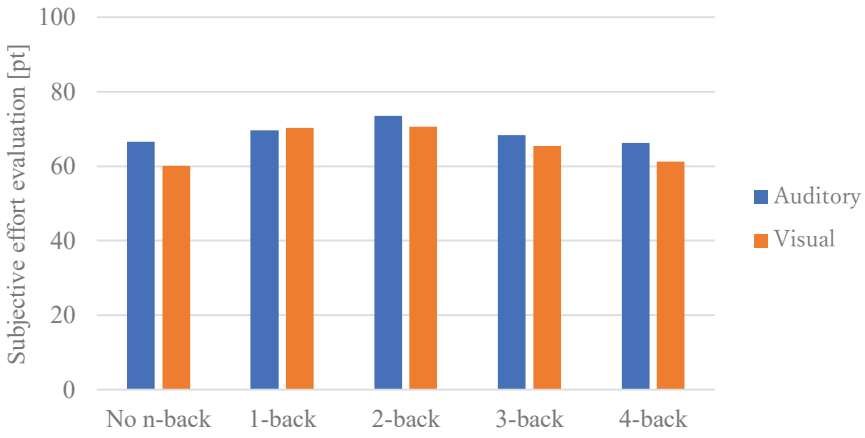

Figure 13. Subjective effort of driving and n-back tasks combined.

Relationship of Degree of Cognitive Load (n-back Tasks) and Driving Performance

Figures 14 and 15 show the amount of lateral offset in the course roadway during the evaluation interval for the visual and auditory n-back tasks, respectively. In the visual n-back tasks, when t-tests were conducted from 1-back to 4-back tasks with respect to the no n-back tasks, there was a significant difference between the no n-back tasks and all of the n-back tasks. As the cognitive load increased, the amount of lateral offset increased. In the auditory n-back tasks, when t-tests were conducted from 1-back to 4-back tasks with respect to the no n-back tasks, the amount of lateral offset did not show a large, significant difference between the no n-back tasks and 1-back to 4-back tasks.

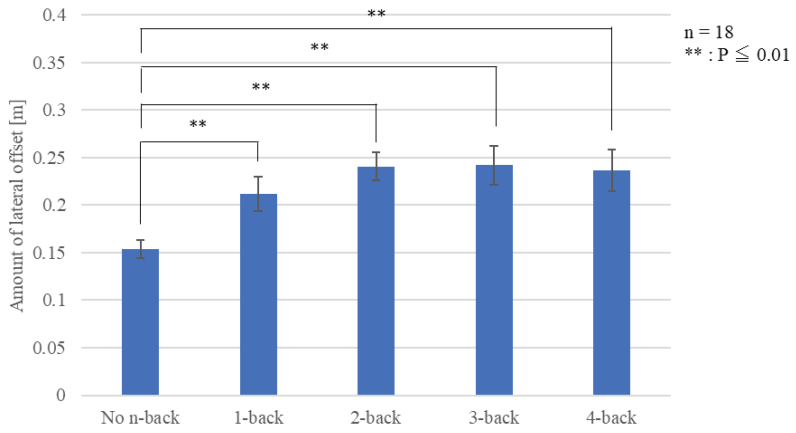

Figure 14. Amount of lateral offset in the visual n-back tasks.

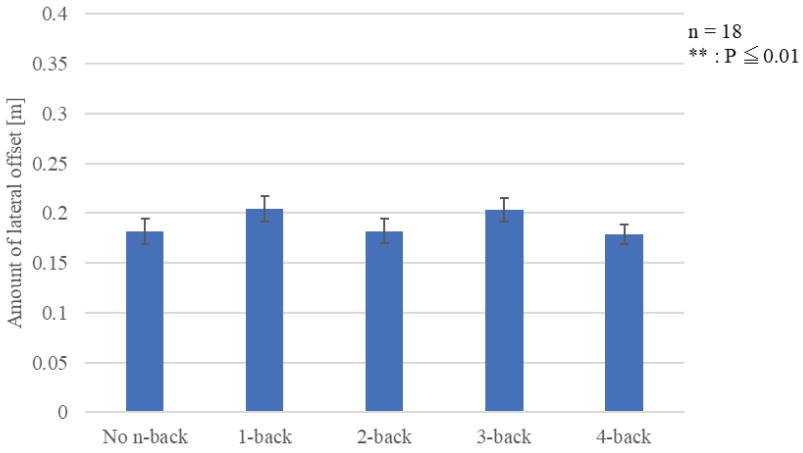

Figure 15. Amount of lateral offset in the auditory n-back tasks.

Differences in Cognitive Processing Domains and the Relationship with Driving Performance

In order to compare the driving performance when visual and auditory n-back tasks were performed, the 1-back tasks, that the difference in the amount of lateral offset between the visual and auditory tasks was slight, were taken as a reference and the change in the amount of lateral offset with the n-back tasks was confirmed. Figure 16 shows the lateral offset relative ratio of each n-back task with the 1-back task as the reference. When t-tests were performed on visual n-back tasks and auditory n-back tasks, a significant difference in the amount of lateral offset between them was observed in the 2-back and 4-back tasks. In the visual and auditory tasks, there was a tendency for the amount of lateral offset that was observed to be larger in the visual tasks.

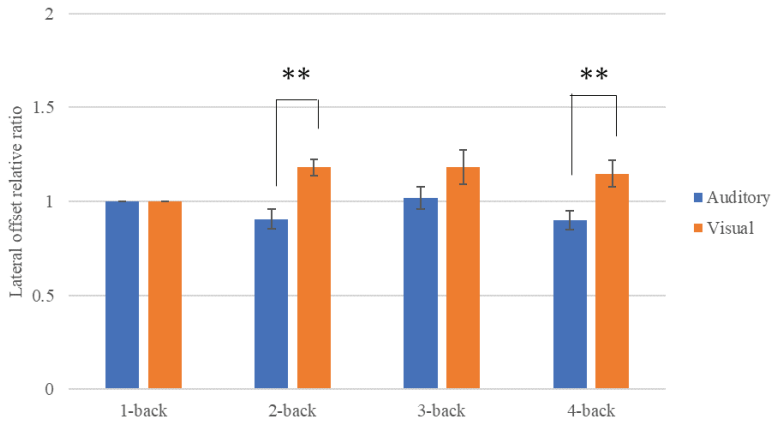

Figure 16. Relative ratio of the amount of lateral offsets in each n-back task with 1-back task as reference.

DISCUSSION

Relationship of Degree of Cognitive Load (N-back Tasks) and Driving Performance

Verification was conducted regarding the influence of cognitive load on driving performance and whether or not there is a moderate degree of load under which performance reaches its highest level.

It was suggested that the influence on steering operation that accompanies increases in the difficulty of auditory n-back tasks is different from the influence on steering operation that accompanies increases in the difficulty of visual n-back tasks. As shown in Fig. 14 and Fig. 15, driving performance in the visual n-back tasks decreased more than in the no n-back tasks. On the other hand, driving performance in the auditory n-back tasks showed no change from the no n-back tasks. However, no tendency for performance to reach its highest level was observed here, either. This is thought to be because differences in the cognitive processing domain result in different influences on driving performance.

Differences in Cognitive Processing Domains and the Relationship with Driving Performance

Verification was conducted regarding whether there is a difference in how driving performance is influenced in separate domains that do not overlap with processing for driving and in the same domain that does overlap with that processing. The result, as shown in Fig. 16, was that the auditory n-back tasks interfered less with steering operation than the visual n-back tasks. Significant differences occurred in the 2-back and 4-back tasks. The n-back task difficulty was more moderate in the auditory 2-back tasks than in the visual 2-back tasks and the subjective effort was higher, and this is thought to be the reason why there was no difference in driving performance between the auditory 2-back tasks and no n-back tasks. There was also the opinion that 4-back tasks were too difficult for the test participants, who therefore gave up on the n-back tasks and gave priority to the driving task. It is conceivable that since the auditory/phonological processing used in the auditory n-back tasks is in a separate domain from the visual/visuo-spatial processing for the driving task, participants were able to concentrate on the driving task so that there was no difference in driving performance from the no n-back tasks.

By contrast, the visual/visuo-spatial processing used in visual n-back tasks is in the same domain as the visual/visuo-spatial processing for the driving task. It is conceivable that for this reason, a simultaneous balance

with the driving task could not be achieved, and so driving performance in the n-back tasks decreased more than in the no n-back tasks. From this, it can be inferred that there are cases when the load from information presentation that employed auditory/phonological processing did not cause driving performance to become unstable, so that there were cases when a simultaneous balance with the driving task could be achieved.

CONCLUSIONS

For this paper, the influence on driving performance from the graded cognitive load of visuo-spatial processing/phonological processing of input from the sense of sight and sense of hearing was studied. As a result, it was found that driving performance changes with the cognitive load. A degree of cognitive load under which performance reaches its highest level was not confirmed, and performance degenerated under the cognitive load from visual/visuo-spatial processing. It was confirmed, however, that under the cognitive load of auditory/phonological processing, which is in a separate domain from the visual/visuo-spatial processing used in driving, there are degrees of load under which performance tends not to change. From this, it was confirmed that there are different tendencies for performance to change according to the type of domain of the load being imposed. It is conceivable that this is because information processes in different domains do not readily influence each other, whereas information processes in the same domain did interfere with each other. In order to maintain stable driving performance, it is conceivably important that the cognitive load imposed on the driver does not interfere with the information from the sense of sight while driving. It may be considered necessary for automobile manufacturers to take the modalities of information equipment into consideration in creating HMIs. As one of the limitations of research, there is the fact that not all processing domains based on the multiple resource model can be examined. It may also be considered necessary to examine the combination of a visual/phonological processing domain and an auditory/visuo-spatial processing domain that does not overlap with the visual/visuo-spatial processing domain that is used in driving.

REFERENCES

[1] ITARDA (Institute for Traffic Accident Research and Data Analysis): MAJOR ACTIVITIES : ITARDA, JAPAN TRAFFIC ACCIDENTS DATABASES(J-TAD), 2020, https://www.itarda.or.jp/english/activities (Accessed 2022.09.23)

[2] Takubo, N. 2005. “An Analysis of Traffic Accident Data for Mental Workload and Human Error by Drivers.” IATSS Review, Vol.30, No.3, September: 299-308.

[3] Baddeley, A. D. and Hitch, G. J. 1974. “Working Memory.” Psychology of Learning and Motivation,Vol.8, 47-89

[4] Wickens, C. D. 2008. “Multiple Resources and Mental Workload.” Human Factors, Vol.50, No.3, June: 449-455

[5] Uno, H., Ohotani, A., Asoh, T. and Nakamura, Y. 2010. “Comparison of Task Demand Assessment Techniques for In-vehicle Driver Interface to Present Information.” Transactions of the Society of Automotive Engineers of Japan, Vol.41, No.2, March: 539-544

[6] Engstrom, J., Johansson, E. and Ostlund, J. 2005. “Effects of visual and cognitive load in real and simulated

motorway driving.” Transportation Research Part F: Traffic Psychology and Behaviour, Vol.8, Issue.2, 97-120
[7] Yerkes, R. M. and Dodson, J. D. 1908. “The relation of strength of stimulus to rapidity of habit-formation.” Comparative Neurology and Psychology, Vol.18, 459-482
[8] Melnicuk, V., Thompson, S., Jennings, P. and Birrell, S. 2021 “Effect of cognitive load on drivers’ State and task performance during automated driving: Introducing a novel method for determining stabilization time following task-over of control.” Accident Analysis and Prevention, Vol.151, 105967
[9] Esaki, S., Obinata, G., Tokuda, S., Mori, N. and Makiguchi, M. 2011. “Estimation of Driver's Fatigue Using Physiological Measures and Principal Component Analysis.” The 54th Japan Joint Automatic Control Conference, Vo.54, 61-61
[10] Kunimi, M. and Matsukawa, J. 2009. “Age-related changes in processing and retention in visual working memory on the N-back task.” The Japanese Journal of Psychology, Vol.80, Issue.2, 98-104

23-0279

NON-INVASIVE BLOOD ALCOHOL DETECTION USING NEAR INFRARED SPECTROSCOPY AND CHEMOMETRIC TECHNIQUES

Salvatore Brauer[1]
[1]Joyson Safety Systems, USA

Gary Ritchie[1]
[1]Joyson Safety Systems, USA

Len Cech[1],
[1]Joyson Safety Systems, USA

Emil Ciurczak[2]
[2]Doramaxx Consulting, USA

John Coates[3]
[3]John Coates Consulting, USA

ABSTRACT

This paper presents an update on the research, development, and manufacturing of a novel passive contact-based Near-Infrared Alcohol Sensor (NIR-AS) for non-invasively measuring Blood Alcohol Concentration (BAC) in human subjects and thus, provides potential for application in support of the new US Infrastructure Investment and Jobs Act bill, section 24220, signed into law on 11/15/2021, once it is enforced.

Alcohol-impaired driving remains a global problem. According to the most recent published report in 2020, U.S. motor vehicle crashes, alcohol-impaired fatalities represent over 30% of the total fatalities; a 14% increase over 2019 and a 29% increase relative to Vehicle Miles Traveled (VMT). The Infrastructure Investment and Jobs Act bill, section 24220, cites statistics on the societal and human costs of alcohol impaired driving and specified intent to make BAC sensors standard equipment in all new U.S. cars in the future. The NIR-AS design and process for analyzing performance in quantifying BAC builds on the R&D carried out in support of the Driver Alcohol Detection System for Safety (DADSS). The published research from DADSS provides valuable technical guidance and performance targets for BAC sensing in motor vehicles. Blood testing is the established gold standard for measuring driver BAC. Although blood testing is the most accurate reference for comparison against NIR-AS (or any new BAC sensor), it is highly invasive, time consuming, and cost prohibitive. Breathalyzers are well established sensors for estimating BAC, however, they also have performance limitations in practical, real-life conditions. Even so, based on published research, including DADSS, breathalyzers can provide an appropriate surrogate reference under controlled clinical and analysis conditions, for analyzing the performance of any new BAC sensor. The NIR-AS sensor described in this paper targets the passive detection performance requirements specified by DADSS.

An alcohol dosing Design of Experiments (DOE) was carried out using a set of Near Infrared Alcohol Sensor (NIR-AS) prototypes with human subjects using a repeat low level alcohol dosing protocol. BAC reference data was also collected using several law enforcement grade and commercial breath analyzers. NIR-AS spectra were processed and analyzed using commercially available and proprietary software.

The DOE resultant data was analyzed using commercially available software packages to produce chemometric models. The paper presents model performance statistics including root mean square standard error of calibration (RMSEC), root mean square standard error of prediction (RMSEP), and square of the correlation coefficient, R^2, for the NIR-AS calibration. A global model employing multiple sensors was tested across the same DOE and performance statistics are presented. Using NIR-AS, it is shown that BAC can be measured at varying concentrations of alcohol within the human body, including low alcohol dosing levels. Further improvements on the NIR-AS design and function will also be presented.

Based on our results, there is significant correlation between BAC breathalyzer and NIR measurements at low dosing levels. The results demonstrate a high correlation between NIR-AS spectra and reference breathalyzers and achieve low RMSEP, RMSEC, and RMSECV. NIR-AS, with continued development, can be a potential tool for assessing driver alcohol impairment in support of ADAS and/or ADS countermeasures.

INTRODUCTION

The determination of the Percent Blood Alcohol Concentration (%BAC) in human subjects using Near Infrared (NIR) spectroscopy and the multivariate analysis technique known as chemometrics has been established [1]. In 2001, using NIR transflectance spectroscopic measurement of analytes (e.g., fat) in milk samples, Norris declared that several criteria must be met for an accurate and precise measurement to be made by NIR [2]. Applying these concepts for the measurement of the %BAC in humans:

1. Alcohol must be able to be detected by NIR at a very low levels ($\leq$ 0.08% or eight one hundredths of one percent, 800 ppm) means that there is approximately 0.08 g of alcohol for every 100 mL of blood. This is the legal definition of alcohol impairment in most states in the United States. Very low levels of an analyte can be detected and measured by NIR diffuse reflection spectroscopy (0.02% - 0.07% BAC) below the 0.08% legal limit.
2. As NIR is not a primary analytical method, but a correlation technique, accurate constituent data (ground truth) is required for developing NIR prediction models of low analyte values.
3. Sampling errors also must be overcome to obtain high accuracy.
4. Sampling errors can be greatly reduced by averaging the spectra from multiple samples of the same constituent level.
5. A narrow bandpass spectrometer is not essential to measure a narrow bandwidth constituent.
6. BAC must be uniquely separable against all other analytes (e.g., specificity).

This paper discusses two experiments (defined further in this paper as ***Surrogate*** and ***Human***) which demonstrate that all six requirements can be met for the detection and the quantification of %BAC. By first analyzing a laboratory surrogate and then using two human subjects (palmar-side finger) and two NIR-AS spectrometers at low alcohol levels using chemometrics and as ground truth, a breathalyzer model Draeger Alcotest 5820 (Houston, TX). The paper includes procedural discussion, the principle components, loading plots, correlation, and regression coefficients from PCA and PLS, and the correlations and regression plots from a recursive chemometrics method called Derivative Quotient Math (DQM).

.

1. Laboratory surrogate (***Surrogate***)

   In the first experiment, it will be shown that the surrogate, consisting of alcohol and water absorbed into a cotton matrix, ranging in concentrations from 0.01% to 0.10 % (~ 100 ppm to 1000 ppm) can be measured as % alcohol directly, using a wide bandpass spectrometer (32 nm) with a Signal to Noise (S/N) $\geq$ 1400:1 @ 1700 nm, or 31.46 dB, and sufficiently resolved from water (the largest interfering absorber in human subjects) by the method of Derivative Quotient Math (DQM).

2. Human subject dosing (***Human***)

   In the second experiment, an attempt to detect and estimate quantity of alcohol will be carried out using spectra from two different NIR-AS spectrometers (same design used in the ***Surrogate*** test) and quantified using an evidentiary breathalyzer at low levels 0.02% - 0.07% (200 ppm to 700 ppm) as %BAC in human subjects.

CHEMOMETRICS

Chemometrics was defined by Svante Wold in a 1971 grant application and mentioned again in 1972 [3], as "*The art of extracting chemically relevant information from data produced in chemical experiments is given the name of chemometric*" in analogy with biometrics, econometrics, etc."

Nearly three decades elapsed between Wold's definition and when Karl Norris studied NIR spectra with a chemometric application referred to as the Derivative Quotient Math (DQM) and concluded [1] that "*an optimized second derivative ratio makes it possible to obtain a linear correlation to an analyte from spectral data from diffuse transmission and diffuse reflectance measurements.*" This further implied that the DQM pre-treatment optimized the spectral data in such a way that the processed data then fits the Beer–Lambert–Bouguer law relationship [4].

INSPECTING ALCOHOL NIR SPECTRUM AND DERIVATIVES

The spectrum of an absolute alcohol[1] (100% ethanol) transflectance reference spectrum is shown in figure 1, along with the spectrum second derivative annotated with prominent wavelength absorption bands shown [5].

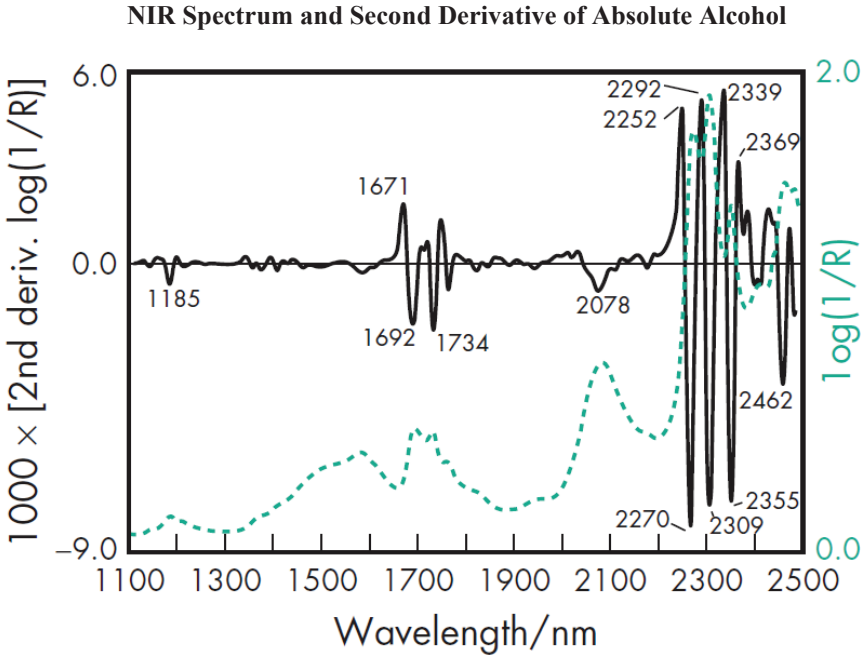

Figure 1. Plots of Absolute Alcohol (a) log(1/R) vs wavelength, dashed line, scale on right-hand side) and (b) second-derivative log(1/R) vs wavelength (solid line, scale on left-hand side). Near-infrared spectrum of alcohol (absolute) measured by transflectance (1mm path-length).

[1] From the Handbook of Pharmaceutical Excipients Sixth Edition. Near-infrared (NIR) spectra of liquid samples were measured using a FOSS NIRSystems 6500 spectrophotometer (FOSS NIRSystems Inc., Laurel, MD, USA). Liquid samples were measured by transflectance using a gold reflector (2 x 0.5mm optical path-length, FOSS) placed in a 45mm silica reflectance cell against air as the reference. Spectra are presented as plots of (a) log(1/R) vs wavelength (dashed line, scale on right-hand side) and (b) second-derivative log(1/R) vs wavelength (solid line, scale on left-hand side). R is the reflectance and log(1/R) represents the apparent absorbance. Second-derivative spectra were calculated from the log(1/R) values using an 11-point Savitzky-Golay filter with second-order polynomial smoothing.

The alcohol second derivative reference displays thirteen prominent bands in the region between 1100 nm and 2500 nm, where all these features are represented with a sensor resolution of 2 nm. This region is the C-H_n first overtone. The molecular formula for alcohol (ethanol) is CH_3CH_2OH. Therefore, the C-H_n first overtone corresponds to the CH_3 and CH_2 (methyl and methylene) group of the alcohol molecule [6].

Table 1 compares the reference standard bands with the five bands of the alcohol second derivative bands measured on the wide-band spectrometer shown in figure 2 (11-point Savitzky-Golay filter with second-order polynomial smoothing). The NIR-AS prototype devices, discussed in the rest of this paper, have a wavelength range of 1350 nm to 2550 nm, with a wavelength resolution of 32 nm.

The alcohol spectrum, 1st ($d^1A/d\lambda^1$) and 2nd derivative ($d^2A/d\lambda^2$) are shown in figure 2 from the wideband spectrometer. Second-derivative spectra were calculated using an 11-point Savitzky-Golay filter using a second-order polynomial smoothing, indicating the derivative of the optical values (usually the absorbance) with respect to wavelength ($d^nA/d\lambda^n$). The spectra taken with the spectrometer were converted from Wavenumber (cm^{-1}) to Wavelength (nm) and Transmission to Absorbance (log 1/T).

Of the requirements previously mentioned, the one having the greatest significance to this research is number six: BAC must be uniquely separable against all other analytes (e.g., specificity). Water is the largest absorber interfering with BAC determination, and useful to investigate several chemometric approaches previously mentioned in other analogous applications. The authors are not aware of reports which investigate the use of the Derivative Quotient Math (DQM) for the determination of Blood Alcohol concentration (BAC) in humans. The algorithm possesses unique capabilities especially suited for measuring BAC in humans.

NIR SPECTROSCOPY SCATTER CORRECTION

The wavelength-dependent redirection of light scattered through a complex media can be defined simply as scatter in the NIR application discussed in this paper.

Davies [7], in presenting an explanation of the origin and use of derivatives in spectroscopy, concluded that, while derivatives are useful for removing extraneous signals from NIR spectra, the resulting spectra still contain multiplicative effects of scatter.

Derivative Quotient Math (DQM) Explanation

Karl Norris, “The Father of NIR,” explained the DQM mechanism in “Norris on Norris Regression,” described in reference 0 as follows:

“*First, there are two distinct items involved. The first is the gap derivative (sometimes called the Norris derivative by mistake), the second is the "Norris Regression", which may or may not use derivatives.*

The "Norris Regression" is a regression procedure to remove the effects of varying pathlengths among samples because of scatter effects. This is accomplished by incorporating a divisor into the regression term. The divisor can be the absorbance at another wavelength, a difference between the absorbance at two wavelengths, a first derivative, or a second derivative. The single wavelength divisor does not work well in many cases because that signal contains offset variations as well as multiplier variations, and we only wish to sense the multiplier signal [Multiplicative Scatter].”

More recently, with the publication of the Fourth Edition of the Handbook of Near-Infrared Analysis [9], Hopkins offers further insight as to why applying derivative pre-treatments alone are insufficient for removing multiplicative effects from light scattering: He asks, “*... why use Derivative Quotients? Simply stated, the ratios effectively cancel*

the multiplicative effects caused by differences in scattering between samples. The ratios will also cancel differences between instruments that are due to differences in spectral bandwidth. In addition, it has been observed that DQM models generally require only 1 or 2 terms, possibly 3 terms. This may make such calibrations very robust…"

When considering the use of derivatives to correct scattering, spectral pre-treatments based on derivatives remove most spectral baseline offsets. However, simple derivative-based pre-treatments cannot remove multiplicative effects. The second derivative-based pre-treatments can largely remove linearly (or nearly linearly) sloping baselines.

Again, considering Figure 2: the Alcohol reference NIR spectrum (Bottom), first derivative (Middle) and second derivative (Top) of alcohol measured by diffuse reflection on the wide bandpass spectrometer (NIR-AS-212,), the highlighted lines represent peak find solutions using the chemometrics software tool Solo (Eigenvector Research Inc., Manson, WA).

The alcohol wavelengths and their positions are annotated textually. The wavelengths for the raw spectrum are positive, zero for the first derivative (points at zero), and negative for the second derivative.

A flow diagram outlining the DQM algorithm is available in reference 8. A MATLAB script, dqm1, was applied on the ***Surrogate*** and ***Human*** data sets. The program executes searches of gaps (segments) or smoothing point intervals (smt) that are tested over intervals individually selectable for the numerator and denominator.

The optimal gap segment size for the wavelength range, using the benchmark devices in the wavelength range between 1350 nm - 2550 nm, can be estimated on (or near) the band centered at 2295 nm (Figure 3). The number of points in the half-peak width of this segment is equal to six (6). The peak height has 12 points from peak to base line. Using these values for the gap and smoothing search, the DQM will not exceed 6 data points for the gap segment search, or 12 for the smoothing interval (see Figure 3).

Further from Hopkins: "*Derivatives can correct the offset and slope differences that are found in sets of spectra of diffusely scattering samples. However, they cannot remove the multiplicative effects. It was observed by Karl Norris that the ratio of derivatives terms can remove the multiplicative effects. He used simple multiple regression to find the terms of quotients employing optimal wavelengths for the numerators and denominators."*

Hopkins explains [10], that choosing an optimal gap size can be approximated by "*selection of a convolution interval about the same size as the number of points in the half-band width of the sharpest band in the wavelength range in which you are working*."

NIR spectrum, (Bottom) First Derivative (middle) and Second Derivative (Top) absolute alcohol from NIR-AS (unit #212) spectrometer.

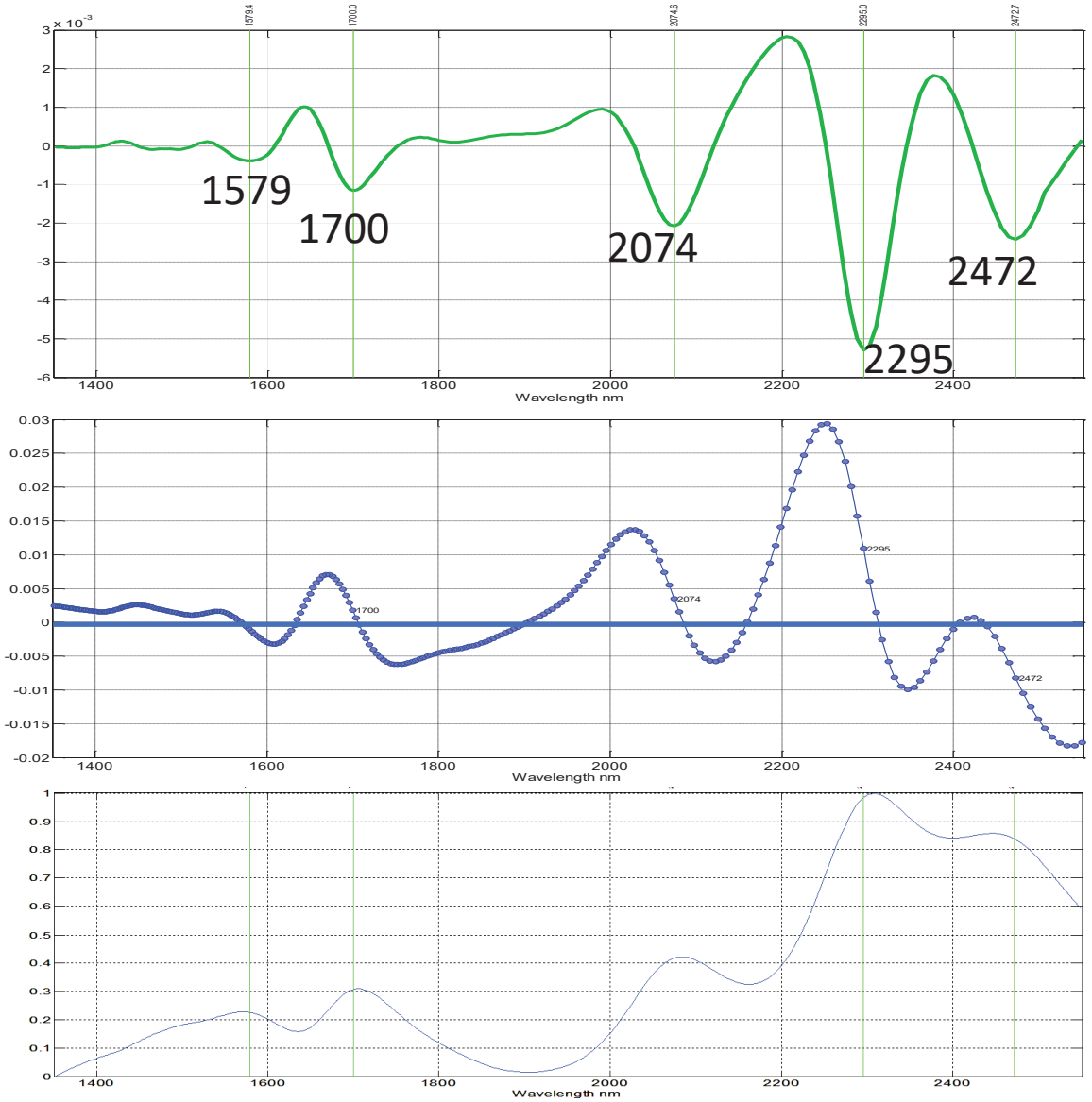

Figure 2. Shows the alcohol NIR spectrum, (Bottom) First Derivative (middle) and Second Derivative (Top) of from NIR-AS (unit #212) spectrometer.

Table 1.
Compares the reference standard bands to the five bands of the alcohol second derivative bands measured on the wide band spectrometer shown in figure 2.

Alcohol NIR Second Derivative of Reference Wavelengths (nm)	Alcohol NIR Second Derivative of NIR-AS-212 Spectrometer Wavelengths (nm)
1185	
1671	1579
1692	1700
1734	
2078	2074
2252	
2270	
2292	2295
2309	
2339	
2355	
2369	
2462	2472

INSPECTING THE ALCOHOL NIR SPECTRA USING THE DQM APPLICATION

Having established the known alcohol NIR absorption bands from the wideband spectrometer above, the DQM program was applied to the ***Surrogate*** spectra in order to find the terms of derivative quotients employing optimal wavelengths for the numerators and denominators.

Experiment 1: Laboratory *Surrogate* (Low concentrations (0.01% - 0.1%) of alcohol and water)

Material and Equipment

The following chemicals were used to conduct the NIR measurements: pure lab grade Alcohol (PHARMCO-AAPER, Brookfield, CT) and distilled water. A micropipette, (Wilmed LabGlass, MED Plus, Vineland, NJ) was used to perform the serial dilutions. 100% cotton medical-grade gauze pads were used as the sampling matrix for these tests.

Procedure

Stock Standard
Serial dilutions were from 0.10% alcohol by volume down to 0.01% in steps of 0.01%. A 0.10% solution stock solution was prepared (500uL alcohol into 500mL water) with a micropipette. Serial dilutions were made with a 10 – 100uL micropipette.

NIR Measurement

To present the samples to the detector window, small uniform cotton squares were prepared and fit into a small holder on the NIR-AS sensor. A spectral sample of the cotton was measured as the reflectance background correction. After the dilutions were made, 300ul of solution was pipetted onto the cotton swab and then placed into the holder. Alcohol Spectra were collected (a total of 10) using 10-second dwell times. This process was repeated for each subsequent concentration.

Chemometric Analysis

The DQM parameters selected for searching the wavelengths from 2000 nm - 2550 nm (alcohol range) (at 32 nm, there are 80 wavelengths) used were:

1. # of Calibration Samples: N22
2. # of Validation Samples: N21
3. Number of terms (1 or 2): 2
4. Derivative order (d): Term #1: 1D, Term #2: 2D
5. Number of differential gaps (gap): 3
6. Number of smoothing points (smt): 6

Recall that, from the inspection of the absolute alcohol NIR spectra, it was determined that the optimum gap and smoothing for peaks in a spectrum is predicated on the half-width of a peak. For alcohol, it was determined that, based on the second derivative spectrum, the most intense absorption band is found at 2295 nm. Based on this fact, the number of data points in the spectrum (80), and studies on the NIR absorption of alcohol and water [11], the selection of DQM parameters were optimized.

With respect to alcohol detection, our research into the optimum range for the detection of alcohol in the presence of water and other absorbers that can interfere with the detection, showed that for the NIR-AS prototype the selection of the 2000 nm – 2550 nm (CH - CH combination) portion of the full range was required in order to identify principle components, latent variables, and to determine the correlation of breathalyzer to the NIR-AS spectra to generate the regression coefficient (PLS model) of the surrogate calibration data set specific for the determination of alcohol.

Second Derivative of the absolute Alcohol NIR Spectra (11, 2, 2D) @ 2295 nm

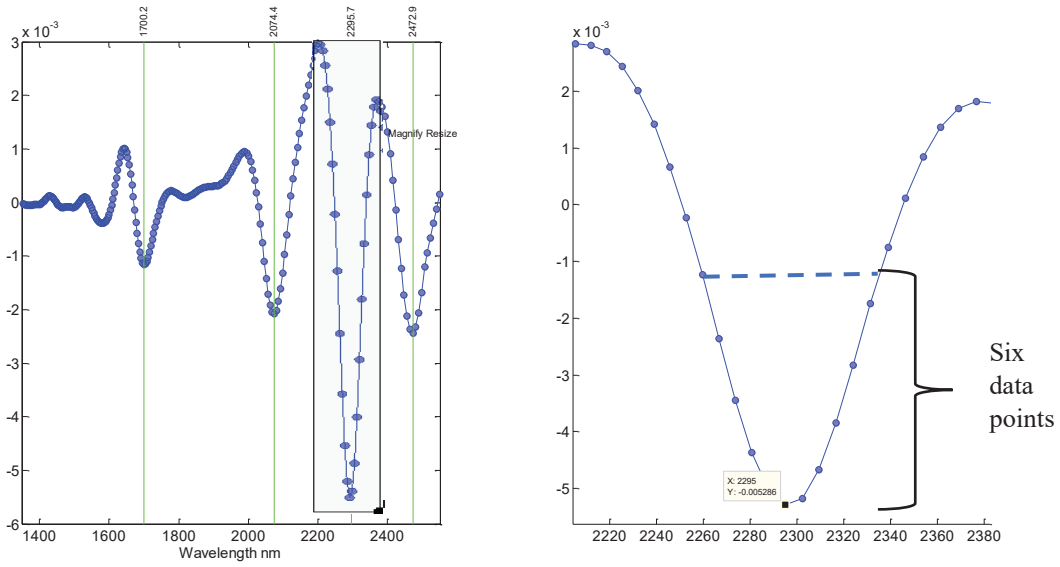

Figure 3. The optimum gap and smoothing for peaks in a spectrum is predicated on the half-width of a peak of interest. Alcohol peak at 2295 nm shows expanded view in order to count data points for band.

The ***Human*** experiment analysis will show that the range from 2000 – 2550 nm**, uses these same wavelengths for the selection of alcohol.** The final results following implementing the dqm1 program ***Surrogate*** experiment 1 (surrogate) were:

%Alcohol Regression Results

SEC = 0.008 RSQ = 0.9375 N = 22 RMSEP = 0.009 SEP = 0.009 bias = 0.000 RSQ = 0.9215 N = 21

2-Term Model: WAEXP2 N22 CAL 4.modl, Ders normalized
Term #1: (2D 2273.6842 NM, gap=1, smt =0)/(2D 2198.8024 NM, gap=3, smt =6)
Term #2: (2D 2392.1824 NM, gap=3, smt =6)/(2D 2028.7293 NM, gap=1, smt =5)
Coefficients B0, B1, B2, .. = -0.61818 0.54646 -0.017699

Figure 4 shows the surrogate NIR spectra and Extended Multiplicative Scatter Correction (EMSC) spectra used in experiment 1. Note that the spectral concentrations (0.01%, 0.03%, 0.05%, 0.07%, 0.09%) do not increase linearly from low to high alcohol concentration with wavelength, but changes due to scatter, instrument, sampling, and other effects previously mentioned.

The spectra are corrected for multiplicative scatter using Extended Multiplicative Scatter Correction (EMSC) [12] prior to analysis following a DQM analysis of unprocessed spectra, giving poor ratio quotient results for the % alcohol at five concentration levels.

Table 2 lists the DQM results for the 2-term and 1-term models for the determination of surrogate concentrations by NIR.

Figure 5 shows the optimal derivatives selected for alcohol by the DQM algorithm. The DQM Matlab application dqm1 (MathWorks, Inc., Natick, MA) was used to find the terms, quotients employing optimal wavelengths for the numerators and denominators terms. Figure 6 shows the calibration from the optimal wavelengths using DQM. The output for the Standard Error Calibration (SEC) and Regression Coefficient plots for the DQM results of the surrogate are shown in Figure 7.

%Alcohol NIR Spectra and Extended Multiplicative Scatter Correction (EMSC) (CAL N = 22)

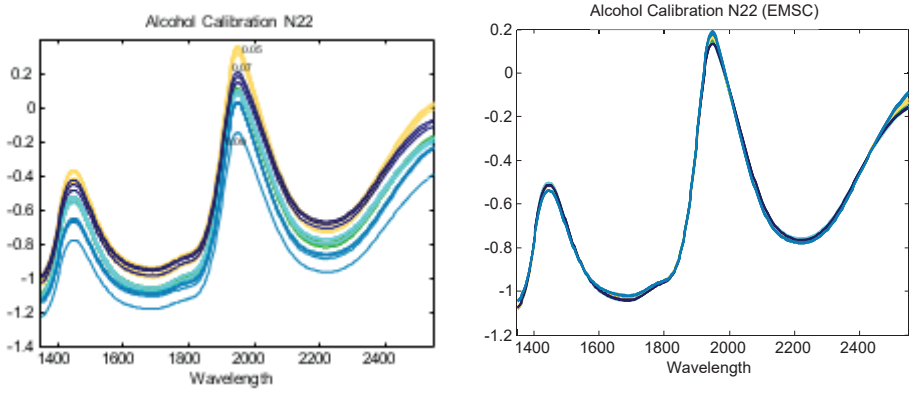

Figure 4. Shows the surrogate NIR spectra and Extended Multiplicative Scatter Correction (EMSC) spectra used in experiment 1.

Optimal Derivatives Selected For %Alcohol (Spectra (N22) By the DQM Algorithm (2-Term Model 2D/2D)

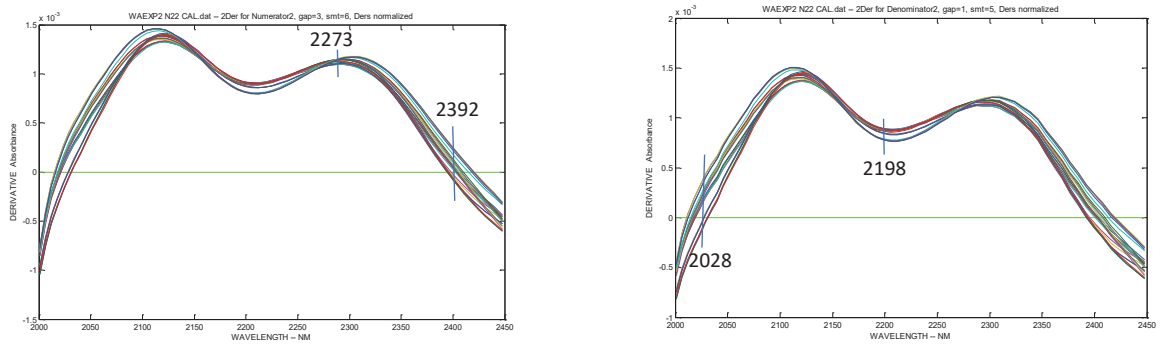

Figure 5. Optimal (2-term model 2d/2d) derivatives selected for %alcohol surrogate (N22) by the DQM algorithm. Note the prominence of regions of the spectra where wavelengths have been selected correspond to regions of no scatter: the cross-over regions of the spectra. Other wavelengths center on or near a region of the alcohol analyte where known NIR absorption is expected to occur.

Surrogate Concentration Scatter Plot CAL (N22) And VAL (N21)

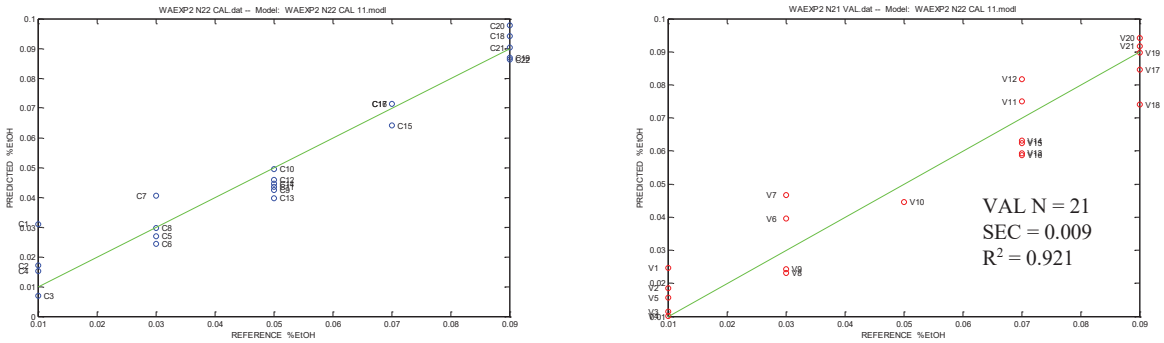

Figure 6. shows the surrogate concentration scatter plot from the DQM calibration and validation as a result of the gap, smoothing (smt) derivative quotient wavelength search.

TABLE 2.

DQM Results for the 2-Term and 1-Term Models for the Determination of Surrogate Alcohol Concentrations by NIR

Wavelengths	**2-Terms**	**GAP**								
Derivative Ratio	**SEC (%BAC)**	**N1**	**smt**	**D1**	**smt**	**N2**	**smt**	**D2**	**smt**	**R^2**
1D / 1D Term 1 2384 / 2239 Term 2 2062 / 2062	0.012	2	0	3	5	3	0	2	0	0.846
1D /2D Term 1 2062 / 2309 Term 2 2225 / 2392	0.011	3	6	1	1	1	0	1	1	0.876
2D /1D Term 1 2309 / 2057 Term 2 2338 / 2017	0.010	1	1	3	6	1	5	3	6	0.883
2D / 2D **Term 1 2273 / 2198** **Term 2 2392 / 2028**	**0.008**	**1**	**0**	**3**	**6**	**3**	**6**	**1**	**5**	**0.938**
	1-Term	**GAP**								
	SEC (%BAC)	**N1**	**smt**	**D1**	**smt**					**R^2**
1D /1D Term 1 2384/ 2239	0.013	2	0	3	5					0.798
1D /2D Term 1 2062 / 2309	0.013	3	6	1	1					0.796
2D/1D Term 1 2309 / 2957	0.013	1	1	3	6					0.816
2D /2D Term 1 2273 / 2198	0.015	1	0	3	6					0.738

Standard Error Calibration (SEC) and Regression Coefficient Plots for the DQM results of the surrogate

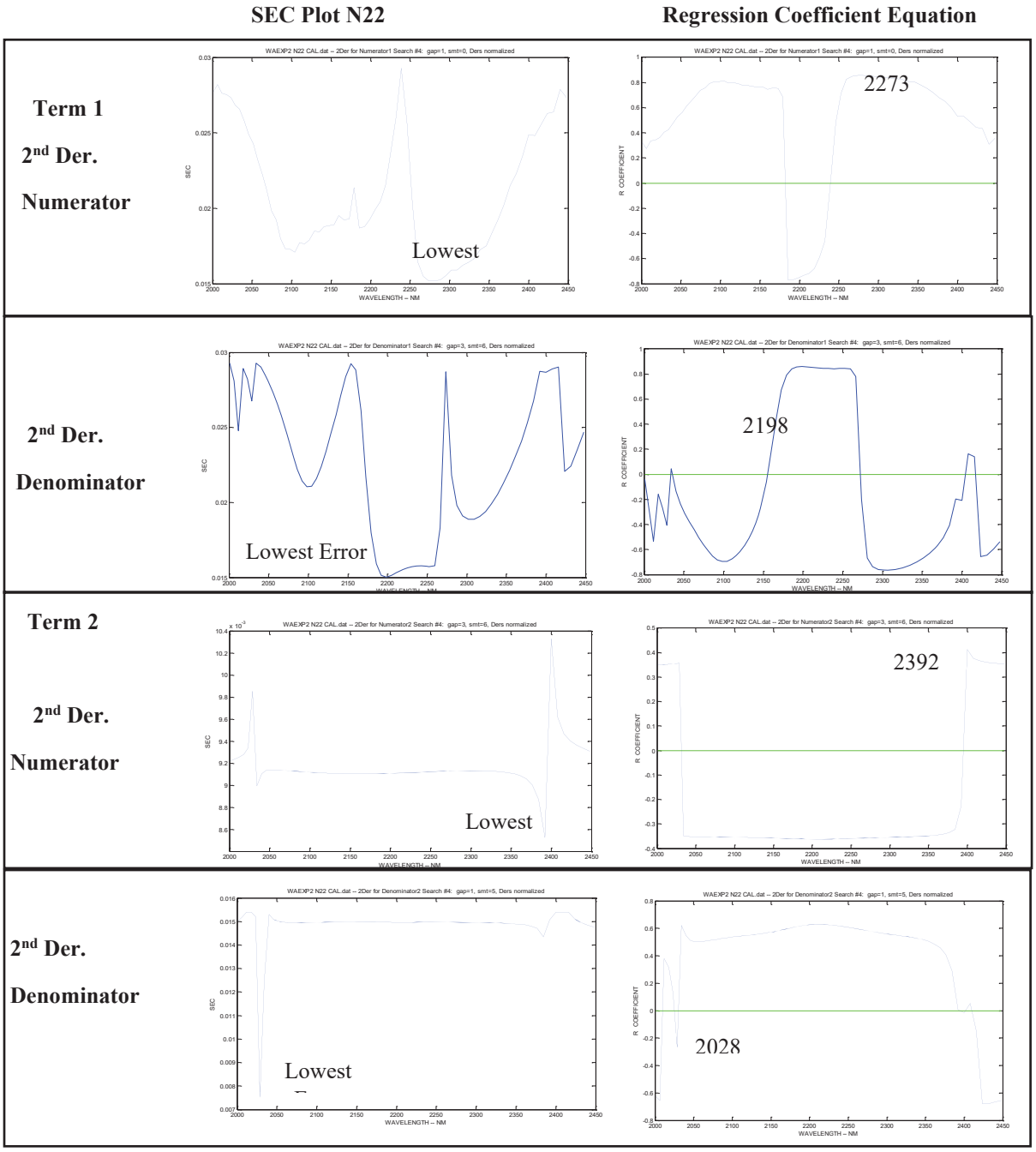

Figure 7. Standard Error Calibration (SEC) and Regression Coefficient Plots for the DQM results of the surrogate mixtures

The best model is found with the lowest standard error of the calibration (SEC) and a coefficient of determination (R^2) approaching 1.0. The SEC was 0.008 and R^2 was 0.938 for sample N = 22 for concentrations in the range 0.01%, 0.03%, 0.05%, 0.07% and 0.09% (Figure 6). The wavelengths selected to be specific for the alcohol should be on or close to those wavelengths specific for alcohol as shown in figure 2. For this analysis, those bands are the alcohol bands at 2273 and 2028 nm.

Based on the results of ***Surrogate***, we will show that the results for the ***Human*** experiment over the calibration range from 2000 – 2550 nm is the only range showing specificity for alcohol using the NIR-AS

IDENTIFICATION OF ALCOHOL BY PRINCIPLE COMPONENT ANALYSIS

Since the DQM provides excellent correlation of the ***Surrogate*** data to the NIR spectra (low error, high coefficient of determination), it may be reasonable to assume that this may be also true for the ***Human*** data. However, prior to testing this hypothesis, a very useful tool to probe for alcohol following a NIR measurement and correlation with the corresponding breathalyzer values is to review the Principle Components (PCs), and residual analysis.

Principal Component Analysis (PCA)

The basic procedure fundamental to most chemometric analysis is the technique known as Principal Component Analysis (PCA). As reviewed by Wold *et al*. [13], and first explained by Pearson [14], the problem at hand is applicable for the use of PCA in determining the differences between groups of spectra. Wold *et al.* provide a general approach for extracting the dominant patterns from data matrices. They are:

1. Formulate the problem statement by asking why the data matrix (in this study, spectra and alcohol doses) were collected in the first place.
2. What is the purpose for the experiments and measurements?
3. Specify, before the analysis. what kinds of patterns are expected to be found.
4. The decomposition of the independent variables (X-wavelengths) and the dependent variables (Y-absorptions) matrices comprising the spectra results in a set of loading and scores describing the variance. (See section on Loadings and Scores).
5. In examining the resulting scores plot, look for outliers (spectra that do not fit any observed pattern), but do not remove outliers without understanding their underlying cause.
6. Use the resulting Principal Components to guide continued investigation or chemical experimentation, not as a result.

For the NIR-AS system under study the general approach is answered by the following statements.

Problem Statement: Development of a passive touch-based Near-Infrared (NIR) sensor for non-invasively measuring the Blood Alcohol Concentration (BAC) in the driver of a vehicle.

Purpose for the experiments and measurements: A NIR touch-sensor potentially offers non-invasive, non-destructive and rapid measurement times. This novel sensor is intended to meet the passive detection requirements of the Driver Alcohol Detection System for Safety (DADDS), improve driver safety by providing a non-intrusive means of notifying a driver or applying some other countermeasure when their estimated %BAC may exceed established threshold(s). When BAC values are modeled, it is intended that the model follows Beer–Lambert–Bouguer Law which describes a linear relationship between the spectral absorbance and the concentration, molar absorption coefficient and optical coefficient of a solution.

However, measuring BAC in human fingers is non-trivial. The finger is a highly scattering, chemically complex matrix, composed of tissues varying in thickness and densities, contributing to scatter. The purpose, therefore, is to find, through well planned and designed experiments, conditions that model this nonlinear phenomenon by reducing scatter, identify and minimize interfering absorbers (e.g., hemoglobin), and linearly correlate only the highly scattered spectra to the % BAC values obtained by suitable reference (in this study, an evidentiary breathalyzer, model Draeger Alcotest 5820 (Houston, TX). In future studies, there is a plan to use Headspace Gas Chromatography with Flame Ionization Detector (HS-GC-FID) on drawn blood taken simultaneously with the NIR-AS and breathalyzer measurements for validation of the NIR-AS touch sensor.

What kinds of patterns are expected to be found: Alcohol dosing curves (Y-matrix) obtained from breathalyzer measurements, when linearly correlated with the NIR-AS spectra (X-matrix) that are obtained simultaneously, result in correlation, score, and loading plots are a function of the relationship of the X – Y matrices. When corrected for variances from the NIR-AS spectra and breathalyzer values, using chemometric preprocessing techniques, the result is a linear regression curve from the calculated NIR-AS wavelength regression coefficients that can be useful for predicting %BAC.

Describe the expected results from the scores and loadings with respect to the observed variance: NIR %BAC results, when linearly modeled by correlating alcohol concentration absorbances and NIR spectral wavelengths, can be used to analyze future unknown NIR-AS spectra obtained through the finger for %BAC.

Describe any observed outliers and explain their root cause: NIR-AS Spectra and %BAC results not within the models 95% confidence limits must be investigated and root cause determined.

Propose future use of the resulting principal components to guide the ADS program: Unknown NIR-AS spectra may be compared with the model PCA scores to classify the spectra as either "free of alcohol" or "containing alcohol" and subsequently analyzed by regression analysis to estimate how much alcohol is or is not present.

Loading and scores: Plots of spectra, scores, loadings, and residuals may be used to extract the relevant information pertaining to the alcohol analyte and provide scientific evidence for the presence or absence of the compound in blood. As such, Nørgaard *et al.* [15] provide a methodology for achieving the objective as stated by Pearson, "to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane." Principal Component Analysis (PCA) can be used to "estimate the latent spectra (loadings) and determine the corresponding concentrations in the samples (scores) from the measured spectra."

The following analysis will be used to demonstrate if NIR spectroscopy can be used to detect the presence or absence of alcohol at low concentrations. Utilizing PCA of the spectra, Partial Least Squares (PLS) followed by DQM will be explored to analyze the spectra for the Percent Blood Alcohol Concentration (%BAC).

Experiment 2: *Human* (Methodology Described by Nørgaard *et al.*)

The data set used for accessing the utility of NIR spectroscopy for the analysis of %BAC is composed of two subjects on two benchmark devices (NIR-AS-103 and NIR-AS-215) and are presented in Table 3.

The %BAC ranges from 0.017% to 0.075%, from $N_{Spectra}$ = 1,171, averaged (Coadd) = 586, and then split into calibration (N = 291) and validation (N = 295) samples. Outlier removal, using Robust PLS, was performed on the

calibration, while manual removal was used on the validation set. Wavelength ranges from 1350 nm - 2550 nm (257 variables), 32 nm bandpass. The Signal to Noise (S/N) was determined to be (≥ 1400 : 1 @ 1700 nm) or 31.46 dB.

TABLE 3.
Design of Experiment for Two Human Subjects Measured on Two Benchmark Devices (N = 1,171)

Device	**Subject 1 NIR-AS-103**	**Subject 2 NIR-AS-215**
	Calibration (N = 683)	**Calibration (N = 488)**
%BAC	0.031% – 0.075%	0.017% – 0.074%
Coadd Data (Before Sorting)	**Calibration (N = 586)**	
N (X-Matrix) (Outliers removed)	(N = 206)	
%BAC	0.017% - 0.074%)	
N (Y-Matrix)	(N = 206)	
Wavelengths 1350 nm - 2550 nm	257	
Split from Coadd Data	**Validation (N = 291)**	
N (X-Matrix) (Outliers removed)	(N = 127)	
%BAC	0.019% - 0.075%	
N (Y-Matrix)	(N = 127)	
Wavelengths 1350 nm - 2550 nm	257	

Figure 8 shows plots of raw and Extended Multiplicative Scatter Correction (EMSC) from the ***Human*** data calibration spectra, measured against a 99% Spectralon® reflection standard, following outlier removal. Spectra are colored by the %BAC levels. The alcohol dosing range (after sorting on ascending values) is shown for each subject in Figure 9.

As seen in figure 8, the BAC observed in the raw and EMSC treated spectra are not varying with alcohol absorption as one goes from the lowest to the highest value BAC, but from the scatter created from the tissue comprising the finger, the measurement placement and other non-absorption sources.

As mentioned in the introduction to this section that, the loadings and the sample residual plots may be analyzed using the approach recommended by Nørgaard *et al*. for "estimating the latent spectra (loadings) and determines the corresponding concentrations in the samples (scores) from the measured spectra."

Plot of Raw and Extended Multiplicative Scattered Correction Spectra From Humans Following Alcohol Consumption Calibration (N206) Validation (N127)

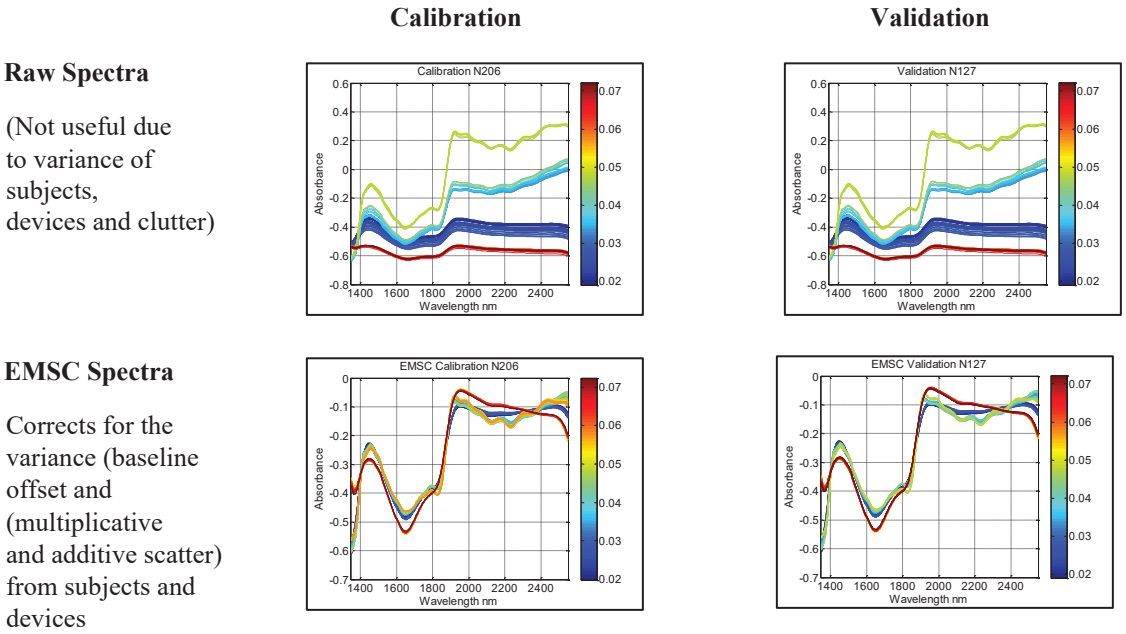

Figure 8. Plot of Raw and Extended Multiplicative Scattered Correction Spectra (Human)

Alcohol Dosing Plots Extended Multiplicative Scattered Correction Spectra After Sorting (*Human*)

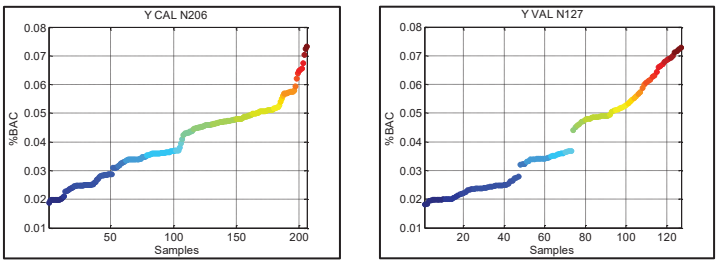

Figure 9. Sorted alcohol breathalyzer value plots for Calibration (N206) and Validation (N127)).

Figure 10 shows the PCs of the BAC values from human subject data of three selected samples, a low, middle, and high %BAC. Note that the PCA model is calculated on the calibration samples (N206) X-variables only (wavelengths); the Y-matrix breathalyzer values are not used. To the left, in column one, the raw spectra are shown for calibration sample 0.030 %BAC (#50), sample 0.050 %BAC (#149), and sample 0.070 %BAC (#241). Column two shows the mean spectrum over all calibration sample. The mean spectrum is identical for all samples.

The first loading vector is the spectral structure that is best at describing the variation in the EMSC data (Figure 8). No other structure can explain more of the variation in the data than the first loading vector. The first loading is common to all samples; what makes the samples different is the content or concentration of this structure in their spectrum. This concentration is called the score value. The score value for 0.030 %BAC is 22.6, for 0.050 %BAC is 47.5, and 0.070% is 10.0. The 288 remaining samples in the data set have different score values. Multiplying the

loading vector with the score values for samples #50, #149 and #241 are the best descriptions one can obtain for these samples, when the loading vector should also describe the other samples. Other observations are the following:

- The shape of the first loading vector is the inverse of the EMSC treated spectra. This explains the greatest variance of the spectra (N206), and is describing all non-absorbing phenomena, (i.e. physical offset of the spectra related to finger placement, differences in individuals, and different NIR device).
- The shape of the second loading vector resembles the human subject finger EMSC spectra. Prominent features of this spectrum are seen around the 1450 nm and 1940 nm wavelengths. These are known wavelengths for the NIR absorption of water.
- The residual spectrum shows, as one goes down the table with increasing alcohol concentration, the wavelengths between 1350 nm – 2000 nm increase at those wavelengths associated with water.
- The residual spectrum also increase between 2000 nm – 2550 nm as one goes from 0.05% to 0.1% alcohol.

Loading Plots of %BAC Alcohol Consumption (Range = 0.030% – 0.070%)

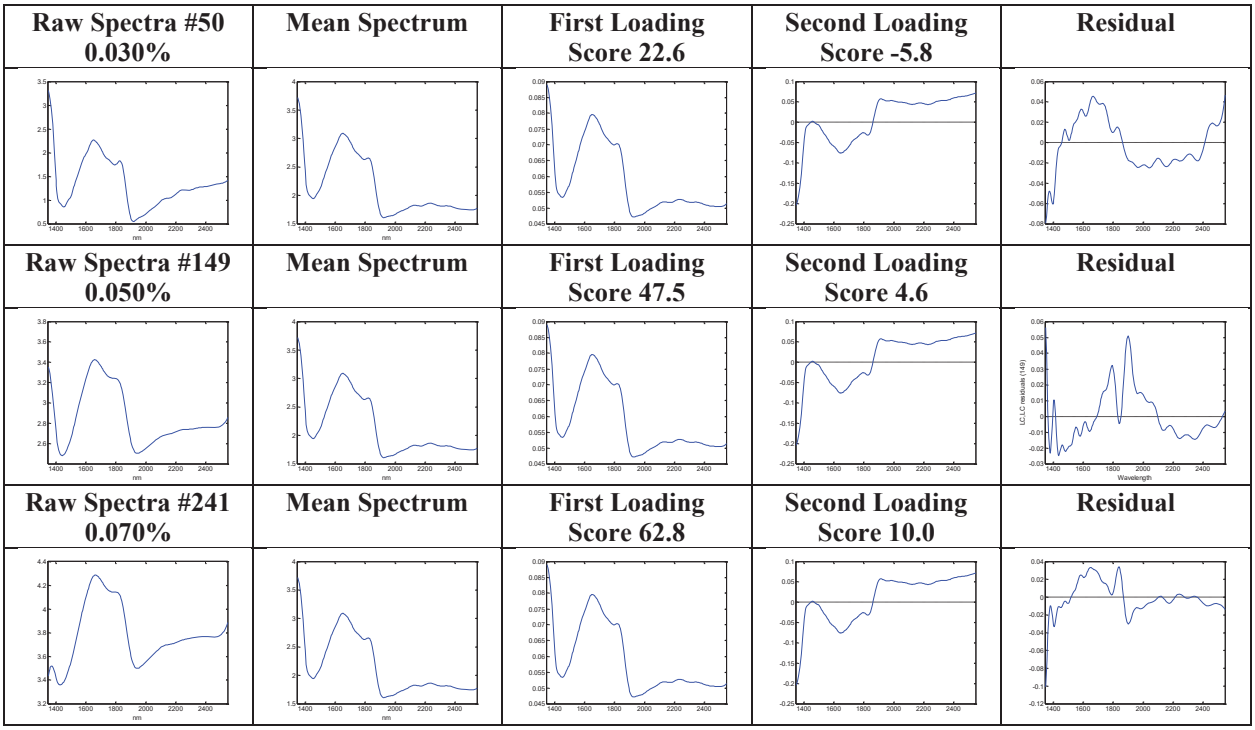

Figure 10. PCA of %BAC Alcohol Consumption (Range = 0.030% – 0.070%)

The second loading is the structure that describes the second most variation in the data set. The vector has the special property of being orthogonal (perpendicular) to the first loading. Once again, the sample diversity is reflected in the loading score value.

The part of the variation in the data set not described by the first two loading vectors is represented by the residuals (Figure 10, column five). The residuals are specific for each sample and can be used for the detection of deviating sample patterns. Note that, as one moves down the residual column, the features of the residual describing all other samples other than the sample selected look less like the loadings as they increase from 0.030% to 0.070%. By comparing the size of the residuals with the variation of the EMSC data, one can calculate the variance explained for each Principal Component.

An important observation in the residual plots as the concentration of alcohol is increasing moving down the residual column, the intensity of the peaks other than those of alcohol in the 1350 nm – 2000 nm region begin dominating the whole absorption spectrum. Prior to testing the DQM on experiment #1, the PLS calibration was tried on the set using the full wavelength range. The resulting derivative plots show the wavelength range being calibrated on is from 1350 nm – 2550 nm, and clearly shows the effect of water in this region around 1450 nm and 1940 nm (see figure 11).

So, while PCA is useful, in this case, finding loading vectors describing the absorption spectrum of water in the finger, these is no clear indication of where alcohol is in the spectrum using the whole wavelength range 1350 nm – 2550 nm.

***Surrogate* experiment, PLS calibration on %Alcohol NIR Extended Multiplicative Scatter Correction (EMSC) (N = 21) Set using the full wavelength range (1350 nm – 2550 nm)**

2nd Loading **Correlation** **Regression Coefficient**

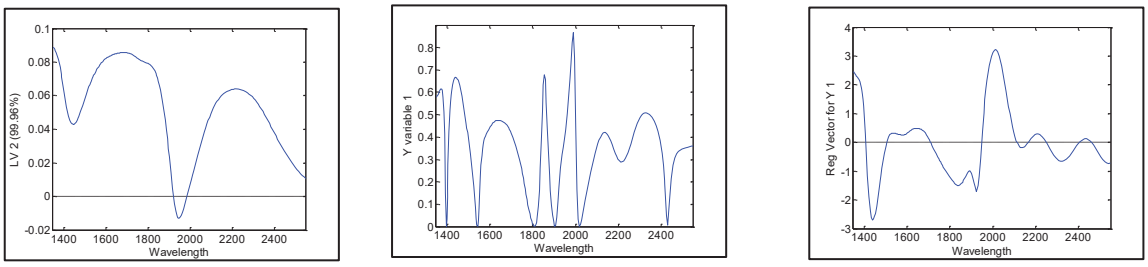

Figure 11. Surrogate Experiment, calibration full range 1350 nm – 2550 nm, shows 2nd loading explaining 99.96% model variation due to water, correlation plot showing water and the regression coefficient showing variables contributing most to the model. Note the large negative coefficient at ~1450 nm and large positive coefficient around ~1940 nm, associated to water.

As can be seen in figure 12, water absorption dominates the correlation and regression even though alcohol and other absorbers from hemoglobin, fat, and other analytes are present in the human. The relationship to these wavelengths is a function of the surrogate concentrations at 0.01%, 0.05% and 0.10%. Hence, the correlation and regression from the wavelengths associated with water strongly exhibit absorption effects (1350 nm – 2550 nm) greater than that of alcohol across the full spectrum.

Can the same patterns from the loading, correlation, and regression coefficient be observed in the Human experiment (BAC% estimation for two subjects and two devices) presented in table 3? Figure 12 shows that the same patterns seen from the surrogate data across the full range spectrum can also be seen in the %BAC human subject plots leading to the conclusion that the spectra from the ***Human*** data set behaves in the same manner as the spectra from the ***Surrogate*** model.

That is, the full wavelength range 1350 nm – 2550 nm is detecting and calibrating on water and other absorbing components of the fingers, not alcohol. The alcohol wavelength range 2000 nm – 2550 nm is detecting and calibrating on alcohol and other absorbing components of the fingers, but not calibrating on water since there are no NIR absorption bands of water in this range.

Experiment 2, 2. Human subject dosing PLS %BAC 2nd Loading, Correlation Plot & Regression Coefficient Calibration (N = 206) Wavelength Range (1350 nm – 2550 nm)

2nd Loading **Correlation** **Regression Coefficient**

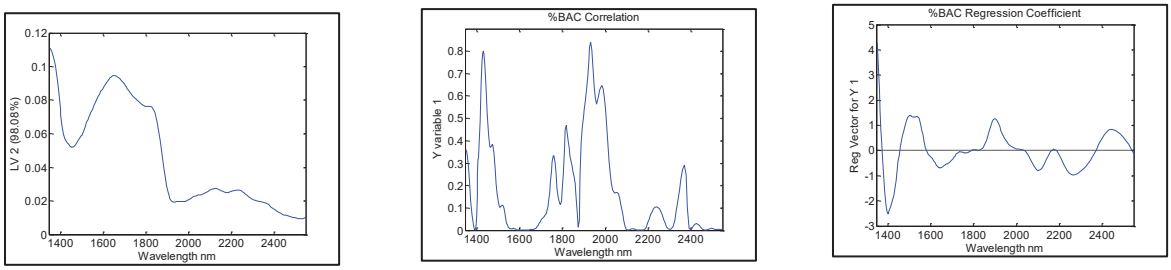

Figure 12. Experiment 2 Human subject dosing calibration range 1350 nm – 2550 nm. Compare this with figure 11. The 2nd loading explains 98.08% model variation due to water, correlation plot showing water and the regression coefficient showing variables contributing most to the model. Note the negative coefficients at approximately 2050 nm, 2300 nm, and a positive coefficient around 2450 nm, is undoubtably an affect from the alcohol.

What is the takeaway from the ***Surrogate*** experiment and subsequent review of the ***Human*** experiment being analyzed in PLS using the full wavelength and alcohol range (2000 – 2550 nm) for calibration? Correlation is not causation. Figure 13 shows the regression scatter plots for the surrogate and human subject data sets from a full range and an alcohol range calibration.

PLS Regression Scatter Plots (Experiment #1 and #2)

***Surrogate* Data Set 1** ***Human* Data Set 2**

Surrogate N22 **%BAC 2 Subject / 2 Devices N206**

7(1350 nm – 2550 nm) **(2000 nm – 2550 nm)** **(1350 nm – 2550 nm)** **2000 nm – 2550 nm)**

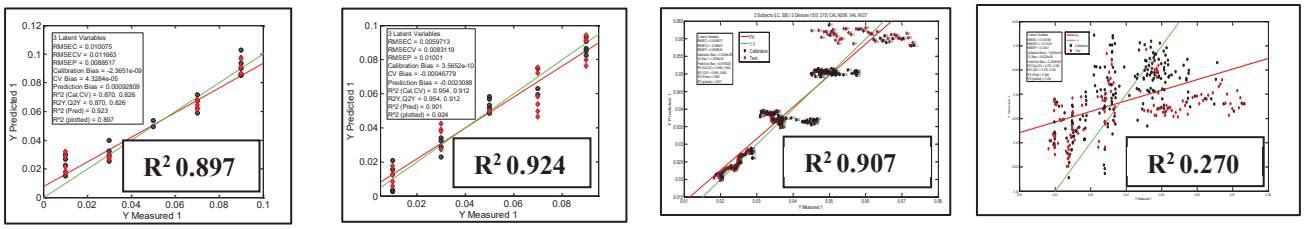

Figure 13. PLS regression plots of calibration (left) and validation (left middle) of surrogate data set, compared to calibration (right middle) and validation (right) of human data set.

The correlation for the 2000 nm – 2550 nm range for data set 2 falls from $R^2 = 0.907$ to $R^2 = 0.270$ when using the full wavelength range. This is in contrast to the ***Surrogate*** experiment, where the calibration correlation went up from $R^2 = 0.897$ to $R^2 = 0.924$. This leads to the conclusion that the correlation for %BAC (***Human***) was caused mainly by the water in the 1350 nm – 2550 nm range. When the calibration wavelength range was shortened to the alcohol range in the ***Surrogate*** experiment, alcohol was the dominant component contributing to the calibration as shown in figure 18. Note that alcohol now dominates the regression coefficient, as evidenced by the dramatic change in the first loadings. They no longer resemble water providing strong evidence that its influence has been reduced in the 2000 nm – 2550

nm range). The explanation for the cause of the correlation for both the ***Surrogate*** and ***Human*** spectra has now been correctly determined using loadings, correlation, and regression coefficient plots.

PLS on *Human* experiment, %BAC 2nd Loading, Correlation Plot & Regression Coefficient Calibration (N = 206) Wavelength Range (2000 nm – 2550 nm)

2nd Loading **Correlation** **Regression Coefficient**

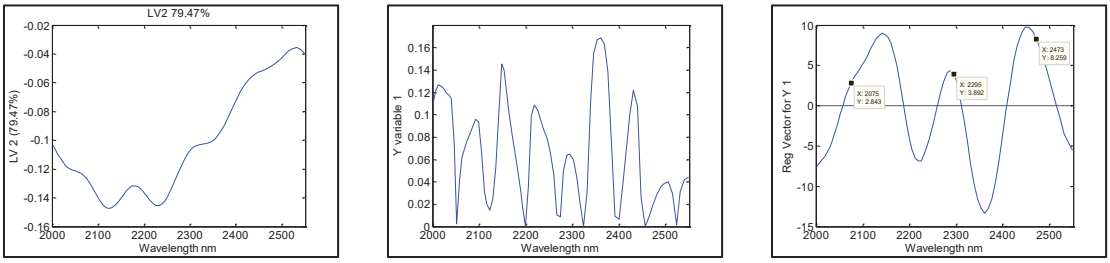

Figure 14. Shows the loadings, correlation, and regression vector plots for human subject data set 2 (2000 nm – 2550 nm) calibration range.

Figure 14 shows that while the regression coefficient is using the wavelengths of those associated with alcohol (see figure 2 and table 1). The second loading from figure 14 explains 74.47% of the model variation, but it does not follow the same pattern as in the surrogate or the %BAC human (N206) full-range calibration. It can only be concluded that other absorbers and non-absorbing phenomenon such as noise and scatter, are also contributing to the variance seen in the model. Moreover, the very low value and random appearing correlation plot may be showing that wavelength multicollinearity is impacting the calibration even more than that already present in the full 1350 nm -2550 nm range. This is not unexpected and can be explained by the reduced number of wavelength variables in the 2000 nm – 2550 nm range (80 compared to 257, when the full range is used), and the decrease of signal to noise.

Experiment 2 (*Human*): The DQM Equation Applied to %BAC Estimation for Human Subjects

Data from ***Human*** experiment (%BAC 2 Subject / 2 Devices) N206 can now be tested by the DQM regression program, following the parameters used on the ***Surrogate*** experiment, except for two differences. First, because the correlation obtained in the data set 2 %BAC 2 subject /2 devices (CAL N206) VAL (N127), it was postulated that using the large number of samples for the 2000 nm - 2550 nm range as was originally used for the full range calibration (N206) was introducing nonlinearities. The data set was trimmed down to CAL N159 and VAL (98) using combined leverage, KNN and manual outlier removal. Second, the spectra from the ***Human*** experiment have significantly lower signal to noise ratio then the ***Surrogate*** experiment, the gap / smoothing was set to a maximum of 6 and 12 respectively, based on the analysis of the alcohol peak at 2295 and measuring the number of data points in the peak at half-width (see figure 3).

The DQM parameters selected for searching the wavelengths from 2000 nm – 2550 nm (alcohol range) (at 32 nm resolution, there are 80 wavelengths) used were:

1. # of Calibration Samples: N159
2. # of Validation Samples: N98
3. Number of terms (1 or 2): 2
4. Derivative order (d): Term #1: 1D, Term #2: 2D
5. Number of differential gaps (gap): 6
6. Number of smoothing points (smt): 12

Figures 15, 16, and 17 give the correlation scatter plot, the derivatives, the SEC, and regression coefficients plots from the analysis. Table 4 provides the DQM final result for the 2-term model for the Determination of %BAC on two subjects and two devices. The final results following implementing the dqm1 program were:

%BAC Regression Results

SEC = 0.005 RSQ = 0.8230 N = 206

RMSEP = 0.004 SEP = 0.004 bias = 0.002 RSQ = 0.9501 N = 98

2-Term Model: %BAC S1 S2 N206 CAL 31.modl, Derivatives normalized

Term #1: (1D 2198.8024 NM, gap=6, smt=10)/(2D 2147.3685 NM, gap=6, smt=12)

Term #2: (1D 2338.8535 NM, gap=2, smt=2)/(2D 2057.1429 NM, gap=4, smt=11)

Coefficients B0, B1, B2, .. = 0.064373 -0.002534 -0.00018087

Optimal Derivatives Selected For %BAC Spectra CAL (N159) VAL (98) By The DQM Algorithm

(2-Term Model 1D/2D)

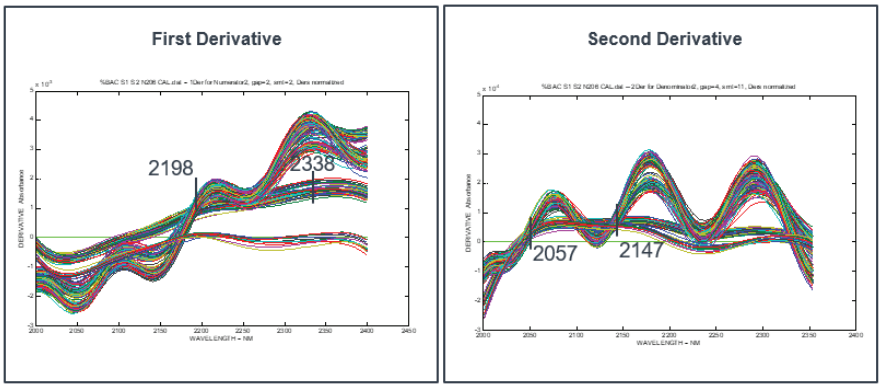

Figure 15. Optimal Derivatives Selected For %BAC Spectra CAL (N159) VAL (98) By the DQM Algorithm. As was shown for the surrogate DQM model experiment 1, note the prominence of regions of the spectra where wavelengths have been selected that correspond to regions of no scatter: the cross-over regions of the spectra. Other wavelengths center on a region of the alcohol analyte where known NIR absorption is expected to occur.

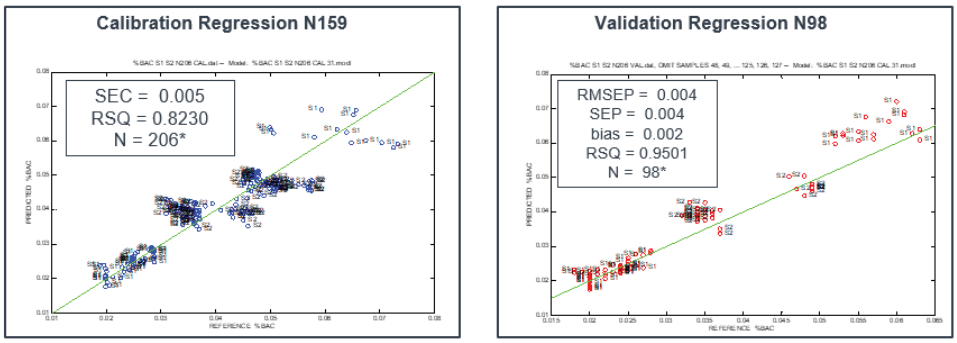

***Figure 16. %BAC Spectra CAL (N159) VAL (98) Scatter Plot*Standard Error Calibration (SEC) and Regression Coefficient Plots for the DQM results of the %BAC Spectra CAL (N159) VAL (98)**

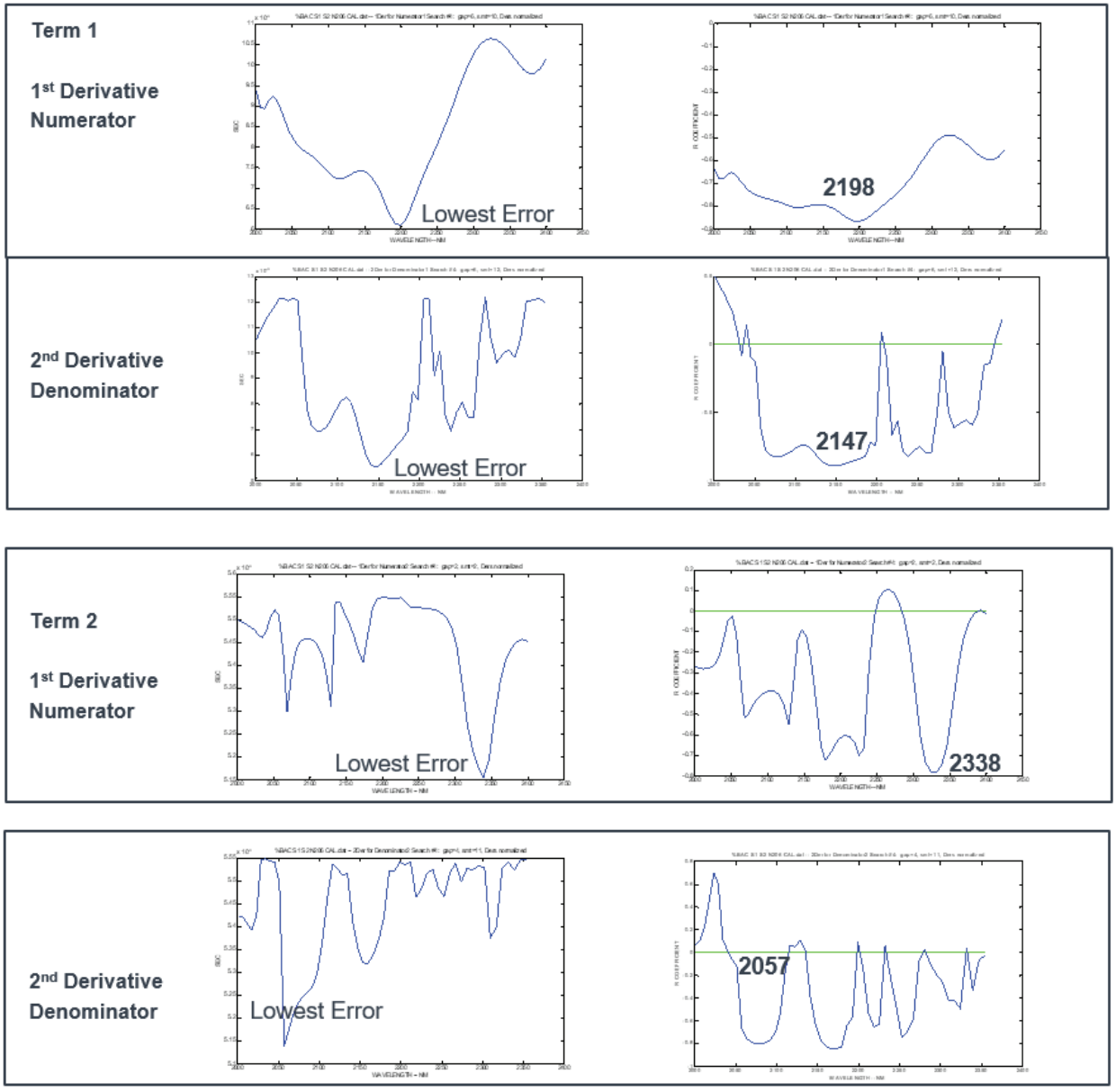

Figure 17. Standard Error Calibration (SEC) and Regression Coefficient Plots for the DQM results of the Human Subjects

TABLE 4.

DQM Results for the 2-Term and 1-Term Models for the Determination of %BAC in two subjects and two devices by NIR

Wavelengths	**2-Terms**	**GAP**								
Derivative Ratio	**SEC (%BAC)**	**N1**	**smt**	**D1**	**smt**	**N2**	**smt**	**D2**	**smt**	**R^2**
1D /1D Term 1 2252 / 2034 Term 2 2172 / 2062	0.005	4	12	2	12	6	0	2	0	0.810
1D / 2D **Term 1 2198 / 2147** **Term 2 2338 / 2057**	**0.005**	**6**	**10**	**6**	**12**	**2**	**2**	**4**	**11**	**0.823**
2D / 1D	0.006	5	0	3	0	5	2	5	1	0.767
2D / 2D	0.006	1	0	1	0	5	9	6	12	0.739
	1-Term	**GAP**								
	SEC (%BAC)	**N1**	**smt**	**D1**	**smt**					**R^2**
1D / 1D	0.006	4	12	2	12					0.754
1D / 2D	0.006	6	10	6	12					0.793
2D / 1D	0.008	5	0	3	0					0.597
2D / 2D	0.007	1	0	1	0					0.645

PLS calibration on %Alcohol NIR Extended Multiplicative Scatter Correction (EMSC) (N = 21) Set using the alcohol wavelength range (2000 nm – 2550 nm)

2nd Loading **Correlation** **Regression Coefficient**

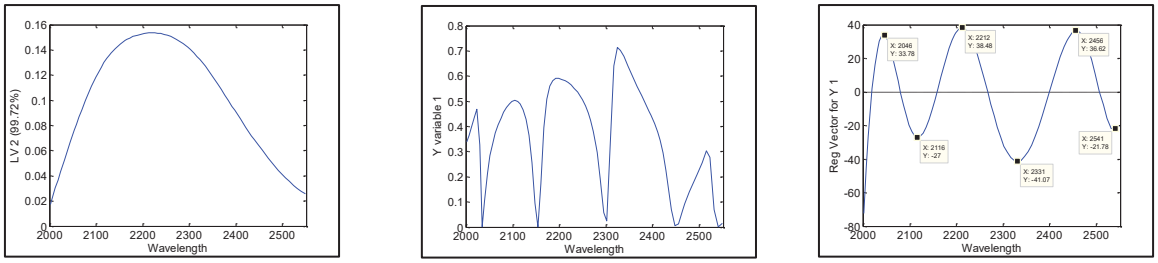

Figure 18. PLS calibration on %Alcohol NIR Extended Multiplicative Scatter Correction (EMSC) CAL (N22) Surrogate data set using the alcohol wavelength range (2000 nm – 2550 nm). The second loading explains 99.72% of the model variance and is now dominated by positive wavelength variables specific for alcohol. Note the coefficients around wavelengths closely associated with alcohol.

DISCUSSION

The PCA, PLS, and DQM analysis applied to NIR spectra obtained from surrogate mixtures of alcohol : water and two human subjects on two benchmark devices demonstrate several aspects of the measurement of %BAC that fulfill the requirements for an accurate and precise measurement to be made by NIR. However, challenges still remain considering instrument Signal to Noise S/N, limited spectral resolution, instrument variability, and measurement sampling variability contributing to additional scatter from nonlinear phenomenon related to biological tissue heterogeneity. Observations from these experiments and analysis using the NIR-AS prototype sensors are the following:

1. The correlation has been shown to be attributed to wavelengths that are specific for the absorption of alcohol in the wavelength range from 2000 nm – 2550 nm. The first and second latent variables of the calibration data set can be used to explain the high linear correlation and model wavelengths variance contribution to the model.
2. PCA of the residual spectra is useful for attributing the cause of the correlation from increasing concentration of the water or alcohol in these experiments.
3. The breathalyzer measurements are accurate and precise, at least sufficient to be used on several different subjects, correlated to different devices and are reproducible over time despite the high variance of the human data set.
4. The analysis presented in this paper is standard protocol for working through PLS model results. The analysis of loadings, correlation, regression coefficients, and residual plots are necessary for establishing the cause of the PLS correlation model. The difference between water NIR absorption versus alcohol NIR absorption show up in the loadings and regression coefficients indicating the source of the correlation of x-matrix wavelengths to Y-matrix breathalyzer values when either the full wavelength or the reduced wavelength scale are selected for calibration.
5. In this study, the basic experimental design of pairing laboratory *in vitro* ***surrogate*** (alcohol-water mixture) studies with ***human*** studies, in order to have a simple frame of reference (low noise, no scatter) for interpreting the spectra and resulting model information for analysis was the key to understanding the complex nature of %BAC determination in humans. This is a major milestone for this experiment and the adoption of this simple approach should be useful in guiding the next recommended R&D phase: human dosing and clinical analysis of breath and blood.
6. PLS analysis shows that there is a significant correlation between BAC breathalyzer values and NIR measurements at low dosing levels. While the use of PLS for calibrating multiple human subjects and NIR devices looks promising in the 2000 nm – 2550 nm wavelength range, further work is ongoing for improving the NIR-AS, design of experiments, alcohol dosing spectral preprocessing and ongoing model variance studies over time.
7. A correlation on the surrogate data set was made using a wideband NIR device (NIR-AS) to measure very low concentrations of alcohol in the range from 0.01% to 0.01%. The same range was used for the determination of the %BAC in humans and gave similar results. This was useful for studying and understanding the cause of the loadings, regression coefficient and correlation plots in a low noise and non-scattering environment. This approach is a significant finding that will aid in improving the human studies, device design and software performance for future NIR-AS devices.
8. Sampling averaging leads to improved correlations. The design of experiment for the human subject data set takes into consideration one of the requirements of measuring low amounts of a constituent on a wide band device.
9. Averaging the spectra from multiple samples of the same constituent level in order to reduce the sampling error, and hence the impact of spectra preprocessing and final model variance due to unwanted physical effects.
10. Attention paid to the handling of spectral outliers and the design of individual data sets are paramount if good correlation models are to be robust.

The introduction of the use of the DQM regression tool for both calibrating and diagnosing data sets used by both PLS and PCA approaches offers renewed insight into the causes of correlation in determining %BAC. The method can be

used to target the identification of derivative ratios at optimal wavelengths that are specific for the alcohol analyte by adjusting the wavelength range, derivative order, gap, and smoothing parameters. DQM offers a specific and sensitive (detection of the analyte in the presence of interfering absorbers and the elimination of multiplicative scatter) method for the accurate and precise determination of BAC in humans by NIR spectroscopy.

Norris [16] describes the utility of DQM moreover, he explains the method as it applies to single, and two term 1st and 2nd derivatives combined with gap and smoothing pretreatment and is excerpted below.

Near infrared spectra of diffusely reflecting samples [e.g. human skin] are characterized by:

- poor reproducibility
- poor linearity
- high noise and high sensitivity to sample measurement geometry.

Typical calibration procedures for such spectra involve pretreatment with multiplicative scatter correction or standard normal variant correction, first or second derivative, and partial least squares regression. In the case of determining BAC in NIR, spectral preprocessing using EMSC followed by DQM, has proven successful as shown above. Studies are ongoing for exploring the utility of the DQM for the determination of BAC using multiple subjects and devices for understanding how DQM can be used in conjunction with other techniques, to arrive at a better understanding of BAC measurements, correlations to breath, blood, or other biological markers identified for alcohol intoxication and the goal of achieving a chemometric predictive model for %BAC by NIR.

CONCLUSION

The determination of blood alcohol concentration by NIR has been shown to primarily the result of wavelengths in the 2000 nm – 2550 nm region of the spectrum. Based on the known NIR absorption bands from alcohol reference material, and the known NIR absorption spectrum of water, it was shown by exclusion and inclusion experiments of portions of the wavelength range that correlate solely to water (1350 nm – 2000 nm) and to alcohol (2000 nm – 2550 nm), that the cause of the correlation and regression equation could be attributed. More importantly, the PLS equation in the wavelength range from 2000 nm - 2550 nm was shown to not be able to regress on two subjects measured on two devices. The reasons for this are not immediately obvious, but sources of variance from multiple subjects and multiple devices as well as diminished signal to noise and absorption from other constituents (i.e., hemoglobin, protein, fat and other NR absorbers in human tissue) are likely the leading causes. This immediately led to the design of the experiments for including the surrogate for analysis and trying to understand the cause of this.

Again, the authors introduced DQM for the reason of trying to understand the cause of the alcohol correlation failing in the human study, but succeeding in the surrogate study using PLS chemometrics. The output of the DQM reveals several attributes of NIR-AS spectra.

1. Wavelength selection for the optimal detection of alcohol correspond to spectral regions of no scatter (the crossover regions) seen in the surrogate and human spectra.
2. The effect of the selection of wavelength range for inclusion for DQM analysis also shows that DQM will select regions of spectra that have a low error for wavelengths in the crossover regions of the spectra, not necessarily from the analyte of interest. For instance, limiting the range to 2000 nm – 2550 nm for gap, smoothing, derivative selection, and optimization will correlate and regress on the alcohol for both the surrogate and human data set. However, when the full wavelength range is selected (1350 nm – 2550 nm), water is selected for calibrating against the breathalyzer values just as in the PLS algorithm as shown for both the surrogate and human data.
3. The regression vector explaining the correlation of the final DQM model can be shown by the PLS analysis of the same data set using loadings, correlation, and regression coefficients that the analyte in the DQM model is in fact alcohol based on known NIR absorption of reference spectrum.

Further experiments have already been carried out and additional are underway for refining the surrogate experiment (i.e., controlling evaporation and understanding the contribution of scatter from the cotton matrix), controlling NIR measurement of the finger, and optimizing the NIR test device for enhanced signal to noise and resolution. The next

phase for beta testing the devices will focus on completing the pre-clinical study of multiple human subjects on multiple test devices and regression analysis of the data by DQM, PLS, and PCA. The goal of providing a global model for the determination of BAC in humans, which we believe is achievable based on this and other decisive and targeted studies designed to understand the cause of correlation and regression of the BAC model.

REFERENCES

[1] 25th International Technical Conference on the Enhanced Safety of Vehicles (ESV), Paper Number 17-0304, DRIVER ALCOHOL DETECTION SYSTEM FOR SAFETY (DADSS) – PRELIMINARY HUMAN TESTING RESULTS, Lukas, S., Zaouk, A., Ryan, E., McNeil, J., Sheperd, J., Willis, M., Dallal, N., and Schwartz, K., (2017)

[2] K.H. Norris, NIR news Vol. 13 No. 3 (2002)

[3] Wold S. Spline-funktioner-ett nytt vertyg i data-analysen. Kemisk Tidskrift. 1972; 84(3): 34– 35

[4] Thomas G. Mayerhöfer, Susanne Pahlow, and Jürgen Popp, The Bouguer-Beer-Lambert Law: Shining Light on the Obscure ChemPhysChem 2020, 21, 1–19

[5] Rowe, R.C., Sheskey, P.J. and Quinn, M.E. (2009) Handbook of Pharmaceutical Excipients. 6th Edition, Pharmaceutical Press

[6] Ma, L., Peng, Y., Pei, Y. *et al.* Systematic discovery about NIR spectral assignment from chemical structural property to natural chemical compounds. Sci Rep 9, 9503 (2019)

[7]A.M.C., Davies, Back to basics: spectral pre-treatments – derivatives, Spectroscopy Europe, pgs. 32-33. Vol. 19 No. 2 (2007)

[8] [http://www.impublications.com/discus/messages/5/237.html?1172691604].

[9] Ciurczak, E.W., Igne, B., Workman, Jr., J., & Burns, D.A. (Eds.), Chapter 14 Derivative Quotient Method Regression, or the Norris Regression, Handbook of Near-Infrared Analysis (4th ed.). CRC Press, pgs. 269 -287, (2021)

[10] [http://www.impublications.com/discus/messages/5/237.html?1172691604],

[11] Wanga, X., Baoa, Y., Liua, G., Li, G., Lin, L., Study on the Best Analysis Spectral Section of NIR to Detect Alcohol Concentration Based on SiPLS, Procedia Engineering 29, 2285 – 2290, 1877-7058 © 2011 Published by Elsevier Ltd. doi:10.1016/j.proeng.2012.01.302 (2012) Available online at www.sciencedirect.com

[12] Martens, H., Nielsen, J. P., and Engelsen, S. B., Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures, Anal. Chem. 75, 394 – 404, (2003)

[13] S. Wold, K Esbensen and P. Geladi, Chemometrics and Intelligent Laboratory Systems, 2 (1987) 37-52

[14] K. Pearson, On lines and planes of closest fit to systems of points in space, Philosophical Magazine, (6) 2 559-572, (1901)

[15] L. Nørgaard, R. Bro and S. Balling Engelsen, Principal Component Analysis and Near Infrared Spectroscopy, A white paper from FOSS

[16] K.H. Norris, NIR news Vol. 12 No. 3 (2001)

DRIVER ALCOHOL DETECTION SYSTEM FOR SAFETY (DADSS) – RISK BASED APPROACH TO ALCOHOL SENSING OUTCOMES MODELING

Timothy Allen, Maura Campbell, Kelly Ozdemir, Kianna Pirooz, Michael Willis, Abdullatif K. Zaouk
KEA Technologies, Inc.
U.S.A.

Susan Ferguson
U.S.A.

George Bahouth, Rebecca Spicer
Impact Research
U.S.A.

Scott E Lukas
Behavioral Psychopharmacology Research Laboratory (BPRL), McLean Imaging Center, McLean Hospital
McLean Hospital, Belmont, MA, 02478,
U.S.A.

Robert Strassburger
Automotive Coalition for Traffic Safety
U.S.A.

Paper Number 23-0291

ABSTRACT

A large number of fatal crashes every year in the United States are caused by alcohol-impaired drivers The Automotive Coalition for Traffic Safety and the National Highway Traffic Safety Administration entered into a research agreement to explore the feasibility of developing a passive in-vehicle alcohol detection system, known as the Driver Alcohol Detection System for Safety, with the goal of significantly reducing the incidence of drunk driving. This paper presents an analysis of the net benefit that could be achieved by installing such technology in the passenger vehicle fleet, using a risk-based approach to model potential outcomes. This outcomes model will calculate the net benefit of, and the public policy challenges associated with, more widespread use of non-invasive technology. Such an approach can be beneficial in determining the merits of the new technology and could be used to help guide public policy with respect to implementation. Furthermore, the technical data can be used to further refine the Driver Alcohol Detection System for Safety performance specifications.

INTRODUCTION

Motor vehicle crashes involving alcohol-impaired drivers result in a large number of deaths and injuries in the United States every year. In 2020, there were an estimated 38,824 motor vehicle crash deaths, of which almost 30 percent involved fatally injured drivers with blood or breath alcohol concentrations at or above 0.08 g/dL, the legal limit in all but one U.S. state (National Highway Traffic Safety Administration (NHTSA), 2022, Insurance Institute for Highway Safety (IIHS), 2022). To address this continuing problem, a Cooperative Research Agreement was instigated in 2008 between the Automotive Coalition for Traffic Safety (ACTS) and NHTSA, to explore the feasibility of developing passive in-vehicle technology that will detect driver breath (BrAC) and blood alcohol concentration (BAC) and ultimately prevent drunk driving when drivers exceed a preset limit. This is known as the Driver Alcohol Detection System for Safety (DADSS) program (Ferguson et al., 2011, Zaouk et al., 2019). The goal is to prevent the vehicle from being driven if the driver's BrAC or BAC is at or above 0.08 g/dL, the legal limit in all but one U.S. state, although other limits can be adopted.

Before such systems can be implemented, policy makers must define performance criteria for these devices. It clearly would be preferable to prevent all incidences of drunk driving. However, there is a concern that trying to prevent the vast number of drunk drivers from driving may result in incorrectly identifying a large number of drivers who are below the limit, thus inconveniencing them. The question is, to what degree would such an outcome affect acceptance of the technology.

Many other vehicle technologies that are currently in use have been shown to decrease crashes and save lives, yet it is acknowledged that they do not perform perfectly. For example, frontal airbags do not always deploy as expected. They occasionally fail to deploy in higher speed crashes or deploy at lower speeds than anticipated, resulting in injuries and deaths that would otherwise not be expected (Ferguson, 1996, 1998). Nonetheless, many studies have confirmed their life saving benefits (Ferguson, et al., 1995, Lund and Ferguson, 1995). NHTSA estimates that as of 2017, 50,457 lives have been saved by frontal airbags (National Center for Statistics and Analysis, 2020). Automatic Emergency Breaking (AEB) systems also do not perform perfectly. Studies that tested the rear-end crash performance of AEB systems when encountering a stationary vehicle at speeds of 30 and 40 mph, have shown that the technology is only about 85% effective in preventing collisions with a stationary vehicles at 30 mph and about 30% effective at 40 mph (AAA, 2022). However, research using real-world crash data, have shown that AEB systems reduce front-to-rear crash rates by 43% and front-to-rear injury crash rates by 45% (Cicchino, 2017). Thus, vehicle safety technologies do not have to be perfect to reduce crashes, injuries, and deaths. Nevertheless, both air bags and AEB technologies still receive public support.

There is one example of a new technology that did not have a successful introduction in the vehicle fleet, because the potential negative impact of the technology on drivers was not sufficiently taken into consideration ahead of its introduction. When driver seatbelt interlocks were mandated in the 1970s there was a public outcry about driver inconvenience. As a result, Congress eliminated the requirement, and the technology was removed from vehicles (New York Times, 1974).

Thus, policy makers must balance the two competing requirements such that alcohol detection systems should have a high rate of success in preventing alcohol impaired driving, but should not inconvenience drivers to the extent that there is an unwillingness to adopt the new technology. If the performance requirements are set too high, manufacturers may have difficulty meeting the requirements or the costs may be prohibitively high. On the other hand, if the requirements are set too low, ineffective devices may be implemented that may be less effective in preventing deaths. Each of these could result in low levels of public acceptance of the device so that their use will be inhibited. Determining what requirements to set that are effective enough, but not so demanding that it prevents manufacturers from developing and implementing the technology, is fundamentally a policy decision. This paper discusses the tools that can be used by policy makers to guide their decision-making process.

RISK-BASED ANALYSIS OF OUTCOMES

Risk-based analysis of outcomes is a method for calculating outcomes based on the intended use of a device, a model of the population of users that will use it, and is based on the sensitivity and the specificity of the device. Sensitivity of the alcohol detection device determines the likelihood that impaired drivers will not be allowed to drive (thus potentially saving lives – known as the benefit). On the other hand, specificity determines the likelihood that sober drivers (or drivers with BrACs/BACs under the set limit) will not be allowed to drive, thus inconveniencing the driver (referred to as harm for the purposes of the calculation). For a given sensitivity and specificity the net value of the use of that device (Benefit - Harm) can be calculated. In addition, general performance acceptance criteria can be set which would give device manufactures, automobile makers, government rule makers, and other interested parties guidance on the minimum performance requirements for a device, such that the device would have a positive impact for users.

As noted above, sensitivity and specificity are indicators of the success in the detection of whatever it is that the device is intended to detect. In the case of a breath-based alcohol detection system the sensor is designed to measure whether a driver's BrAC is above a predetermined limit. If the device reports the driver's BrAC is above the limit, the result is referred to as positive. If the device reports the driver's BrAC to be below the limit, the result is referred to as negative. If the drivers true BrAC is actually known through the use of a reference test device, it is possible to

calculate the success of the device that is being evaluated. If the driver's BrAC is above the limit and there is a positive result this is referred to as a True Positive (TP). If on the other hand, the driver's BrAC is above the limit but there is a negative result this is referred to as a False Negative (FN). Sensitivity is the term used to describe the overall success rate with subjects that are impaired. This is calculated as the fraction of all impaired subjects who were tested that gave a True Positive result (Sensitivity = TP/(TP+FN)).

Conversely, if the driver has a BrAC below the limit and the device being tested gives a negative result this is a True Negative (TN). Similarly, if the driver has a BrAC below the limit, but the test device gives a positive result this is referred to as a False Positive (FP). Specificity is the term used for the overall success rate with test subjects that are unimpaired. This is calculated as the fraction of all unimpaired subjects that were tested which gave a True Negative result (Specificity = TN/(TN+FP)). Figure 1 provides an illustration of these concepts.

	Unimpaired/Allowed to operate the vehicle	Impaired/Not allowed to operate the vehicle
Test Result = Positive	False Positive (FP)	True Positive (TP)
Test Result = Negative	True Negative (TN)	False Negative (FN)
	Specificity = TN/(TN+FP)	Sensitivity = TP/(TP+FN)

Figure 1. Table of possible outcomes from the use of a device

CALCULATION OF THE BENEFIT AND THE HARM FROM THE USE OF A DEVICE

In a risk-based analysis of outcomes the value of a device is calculated as the benefit derived from using the device minus the harm done by using the device. Value = Benefit – Harm. As noted above, there are four possible outcomes: True Positive, True Negative, False Positive, and False Negative. Looking at each of these in turn it can be determined if these outcomes are beneficial, harmful, or neutral. In the case of the DADSS sensor a True Positive result is when a driver with a BrAC above the limit receives a positive test. Depending on the intended use of the device, a range of outcomes might occur; the car could be prevented from being put in gear, could be limited to lower top speed using a "limp mode," or the driver could simply be given a warning. All of these outcomes would be positive as they would likely result in a reduction in the number of impaired drivers. In contrast, in the case of a False Positive, resulting from a driver with a BrAC below the limit receiving a positive test, the outcome would result in inconveniencing the driver. As noted above, the amount of the inconvenience would depend on the implementation. A warning could potentially be merely an annoyance, but if the car is prevented from moving, the inconvenience would be much more significant. A True Negative, when the unimpaired driver receives a negative result is a neutral result, there is no benefit in this case, but also there is no inconvenience to the driver. Similarly, in the case of False Negatives the result is considered neutral. Currently, there is no in-vehicle system preventing impaired drivers from driving, so if an impaired driver receives a negative result from the sensor, the situation would

be no different than the current situation. For risk-based analyses, in cases where a device is being proposed to replace an existing device, False Negatives have to be taken in to account because the test subject might have had a positive result which is no longer detected with the new device. However, since the DADSS system would be a new technology that is not replacing an existing system, this outcome is considered neutral for the purposes of this analysis. In summary, for a DADSS system, a True Positive is considered a beneficial outcome, a False Positive is a harmful outcome, and both True Negatives and False Negatives are considered neutral.

The overall benefit is the value of the benefit multiplied by the number of times that benefit is obtained, which is the total number of True Positive results for a given population. The sensitivity (i.e., the rate of True Positive results for the positive population, TP/(TP+FN)) is multiplied by the total number of the positive population. Thus, as shown in equation1, the benefit of using a DADSS system is Benefit = Value of Benefit * Sensitivity * Population of Drivers over the limit. Similarly, the overall harm of using the DADSS system is the cost of the harm multiplied by the number of times that the harm occurs. The total number of times the harm occurs is the total number of False Positive results in the population, times the size of the population of drives with drivers under the limit. Because Specificity (i.e., TN/(TN+FP) is the rate of True Negative results, the rate of harm is the complement of specificity, Rate of Harm = 1-Specificity. The resulting harm of using a DADSS sensor is thus Harm = Value of harm *(1-Specificity)* number of drives with drivers under the limit. The total value of using the device is shown in equation 1 below.

Value = Value of Benefit * Sensitivity * Population of impaired Drives –

Value of harm *(1-Specificity)* population of unimpaired Drives

Equation 1. The Value equation for an Alcohol Detection system using risk-based analysis of outcomes.

HOW SENSITIVITY AND SPECIFICITY RELATE TO ACCURACY AND PRECISION OF A DEVICE.

The risk-based analysis of outcomes value equation uses sensitivity and specificity to calculate the net benefit of a device. In practice, it is very burdensome to test individual devices for sensitivity and specificity because these are population-based statistics. Thus, in order to accurately measure the sensitivity and specificity of a device, large scale studies with large number of subjects would be required. As a result, policy makers will need to define other requirements in a device model specification, such as accuracy and precision, which can be measured in a laboratory setting. Therefore, it is important to understand how accuracy and precision relate to sensitivity and specificity.

Sensitivity and specificity are measures of the success rate of a device with respect to the entire population that are tested. This success depends on the accuracy and precision of the device. Accuracy is how close a given set of measurements are to their true value, while precision is how close the measurements are to each other, in other words the standard deviation. Because we expect the error of a device to be random, the entire population of the results for a specific test device will be normally distributed and the expected results can be graphed with a Gaussian, or bell curve. The curve will be wider or narrower depending on the precision and accuracy of the device. A more accurate and precise device will have a narrower distribution of results, while a less accurate and less precise device will have a wider distribution of results. Because devices are calibrated, it would be expected that, on average, the bias is zero, and that any inaccuracies in the individual devices would result in a general broadening of the curve for the whole population.

The assumption of the normal distribution of the test results is that the variability is due to the normal random error in measurements. This does not account for erroneous results that may occur if a device is not functioning as intended, for example, if the device breaks. It also does not account for non-random errors (for cause errors) that occur, for example, from an interfering substance. If results from either of these sources of error are a significant contribution to the total results, then the calculated sensitivity and specificity will not match the sensitivity and specificity actually measured in practice, undermining the validity of the model calculations. If, on the other hand, these non-random errors are not a significant contribution to the total number of results, for example because the device is very robust or has a mechanism for detecting device failures, and it does not report such results, then the non-random errors are not a major contributor to model and can be ignored. The DADSS program has done extensive testing of breath-based alcohol sensors in both laboratory and field settings. It has been found that the

sensitivity and specificity as measured in natural settings with a wide range of subjects and in a variety of environmental conditions, matches what is predicted from the calculated accuracy and precision.

How accuracy and precision are related to sensitivity and specificity can be shown graphically. For example, in Figure 2, a curve of results that would be expected from repeated testing of a population of drivers with a true BrAC of 0.07%, (less than the 0.08% limit) is shown. The width of the curve depends on the precision of the device; that is, a more precise device will have a narrower distribution of results, while a less precise device will have a wider distribution of results. Since these drivers all have a BrAC of 0.07% they should be allowed to drive if they are registering a True Negative. If the result is found to be 0.08% and greater, the result is a False positive. Because specificity is the percent of the results that are True Negatives, the percent of the area under the curve to the left of 0.08% is the expected specificity of the device (90%) that has the precision and accuracy that resulted in the Gaussian curve in the graph.

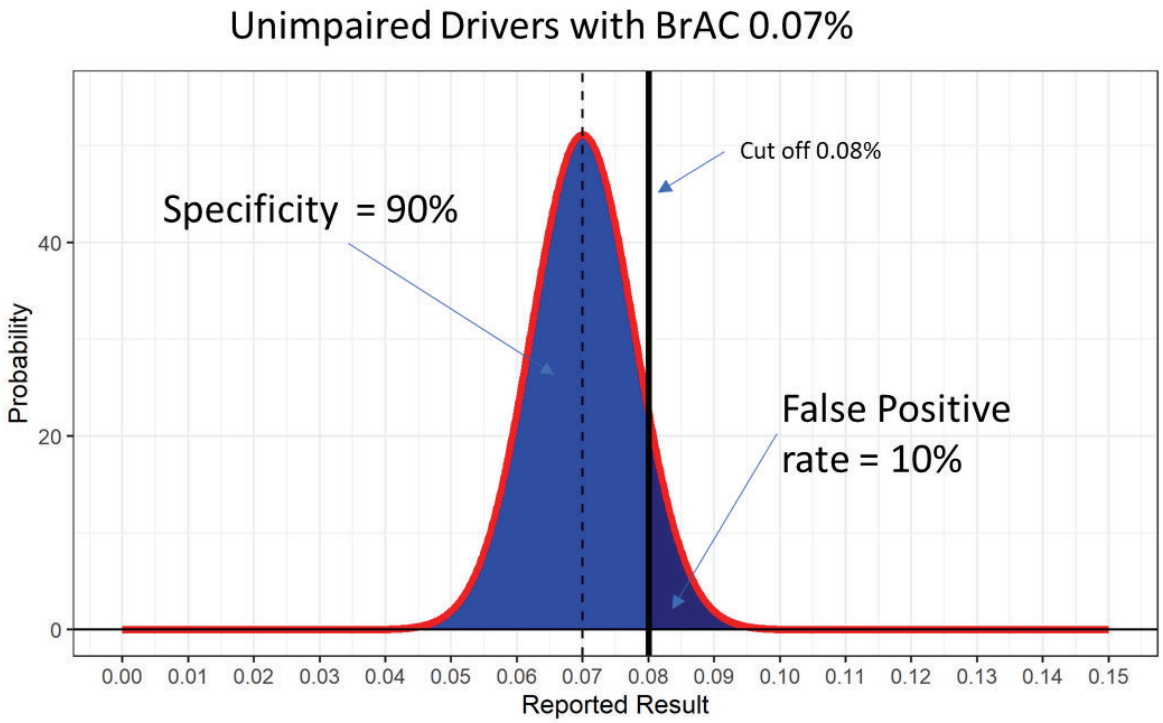

Figure 2. Unimpaired drivers with BrACs of 0.07%

In Figure 3, a curve of results that would be expected from repeated testing of a population of drivers with a true BrAC of 0.095% is shown. Because all these drivers have a BrAC of 0.095% they should not be allowed to drive. If the device reports a result of greater than 0.08% then the result is a True Positive, but if the result is less than 0.08% the result is a False Negative. Because Sensitivity is the percent of the results that are True Positives the percent of the area under the curve to the right of 0.08% (95%) is the expected sensitivity of a device that has the precision and accuracy that resulted in the Gaussian curve in the graph.

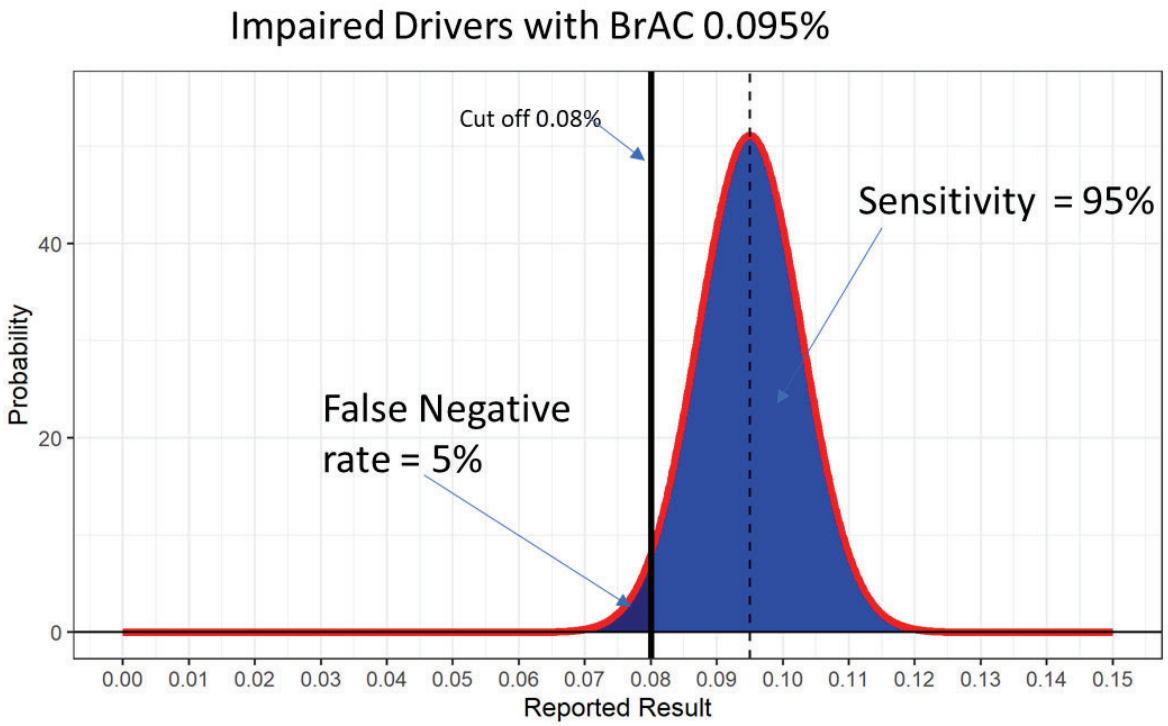

Figure 3. Impaired drivers with BrACs of 0.095%

Sensitivity and specificity depend enormously on the population that is being tested. For example, in Figure 2, the device has a specificity of 90% when testing a population of drivers with BrACs of 0.07% - a result very close to the cut off of 0.08% used to designate drivers as impaired. Using a device that had exactly the same accuracy and precision but instead testing a population of drivers that had BrACs of 0.05% (see Figure 4), should result in a much higher specificity because a much lower percentage of the results would be across the cutoff of 0.08%. Such a device would have a specificity of 99.99% with drivers that had a BrACs of 0.05%. As shown in Figure 5, with sober drivers (BrAC of 0.00%) the device would have a specificity of > 99.99999%.

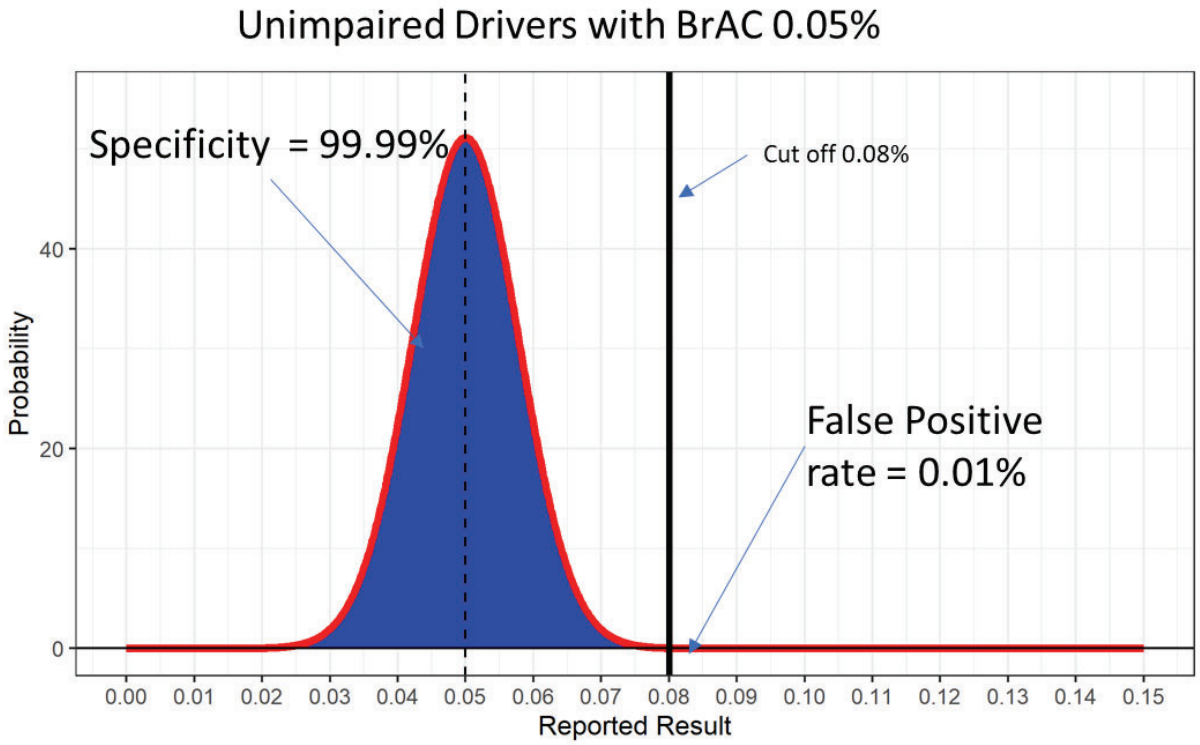

Figure 4. Unimpaired drivers with BrACs of 0.05%

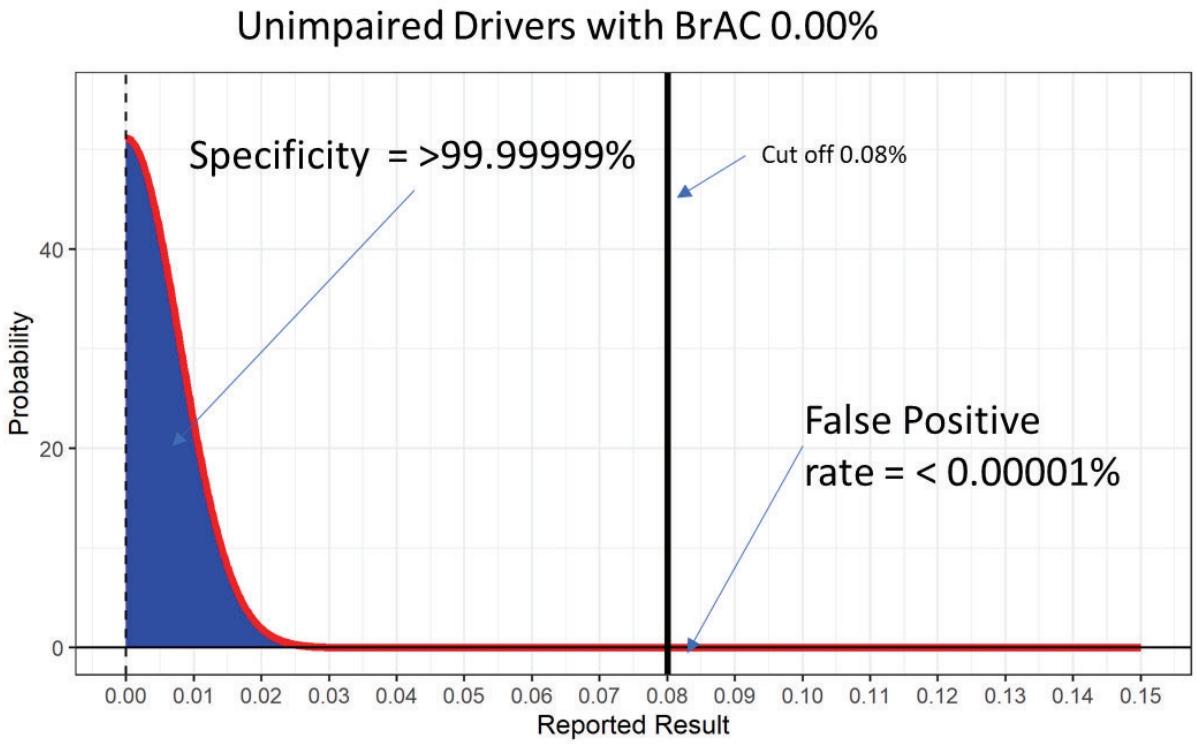

Figure 5. Unimpaired drivers with BrACs of 0.00%

An important concept of risk-based analysis of outcomes is that the sensitivity and specificity used to calculate the value to the population must be the sensitivity and specificity of the device for the whole population of people that will use the device. Thus, results used to calculate sensitivity and specificity must be weighted to account for the actual population of users. Sensitivity is additive by population if there are different populations, for example, drivers with different BACs. For example, if a device had a sensitivity of 50% with a given population A and 75% with a different population B, and if the A population was 10% of the total population and the B population was the remaining 90% the device sensitivity for the whole population is

50% (sensitivity of pop A) x 10% (fraction of total pop that is pop A)

+

75% (sensitivity of pop B) x 90% (fraction of total pop that is pop B)

=

72.5% overall sensitivity

Analyses of data from the 2013-2014 National Roadside Survey (Ramirez et al, 2016) combined with data from the 2017 National Household Travel Survey (FHWA, 2017), shows that there are approximately 600,000,000 drives each day in the U.S. Of these drives, there are an estimated 97.89% of trips that are driven sober. Of the drives by drivers who have some level of blood alcohol:

- 59% (1.25% of all drivers) are between 0.001 and 0.049 % BrAC,
- 15% (0.31% of all drivers) are between 0.05 and 0.079 % BrAC, and
- 26% (0.55% of all drivers) are over 0.08 %BrAC.

In the case of the example device which has a specificity of 99.99999% for sober drivers, 99.99% for drivers with 0.05% BrACs and 90% for drivers with a 0.07% BrACs, the overall specificity of the device can be calculated as follows:

Sober drivers are 98% of all drivers and will have a specificity of 99.99999%, therefore:

Specificity = 99.99999* # of Sober Drives = 97.89%

If we assume that the 1.25% of all drivers that are between 0.001 and 0.049 % BrAC have a specificity of 99.99%:

Specificity = 99.99 * # of all drives between 0.001 and 0.049% = 1.25%

and that the 0.31% of all drivers that are between 0.05 and 0.079 % BrAC have a specificity of 90%:

Specificity = 90 * # of all drives between 0.05 and 0.79% = 0.31%

Combining the results together:

(99.99999*97.89) + (99.99 * 1.25) + (90*0.31))/(97.89+1.25+0.31) = 99.97.

Thus, the device would result in an overall sensitivity of approximately 99.97%.

In reality, it would be expected that the specificity would be higher than this because the 1.25% of drivers who have BrACs of between 0.001% and 0.049 % will mostly be made up of drivers that have a BrAC that is significantly less than 0.05. However, since the exact distribution of drivers in this range is not available from the data, it is only possible to conservatively estimate the sensitivity as 99.99%. Similarly, many of the drivers in the 0.05% - 0.08% BrAC range likely will have specificity higher than 90% (the 0.07% BrAC specificity), because drivers with a BrAC of 0.05 will have a specificity of 99.99%. Since the actual distribution of BrACs of drivers in this range is unknown, the specificity could be estimated at 90%.

IMPLICATIONS OF THE RISK-BASED ANALYSIS OF OUTCOMES MODEL FOR DADSS SYSTEMS.

For an alcohol detection system to be effective it would need to be installed in a large number of vehicles. For example, policy makers could require their installation in all passenger cars. Based on this assumption, it is useful to look at the implication of the risk-based model for DADSS if it were to be implemented nationally. As can be seen from equation 1 on page 4, the overall value from the use of a device depends on the value of the benefit from a True Positive and the value of the harm from a False Positive. This depends on the intended use of the device, which is to say, what action is taken when a positive result is obtained. As described earlier, potential actions could include preventing motive power, a warning indicator, or a limp mode. For instance, an impaired driver could be prevented from putting the car into motion. This would maximize the value of the benefit because the impaired driver would be compelled to find alternative means of transportation and would be unlikely to get into crashes resulting in property damage, injuries, deaths, or other outcomes. However, this also means that the inconvenience to an unimpaired driver with a false positive would be greater, as the driver would be prevented from driving.

Since it is not possible to distinguish between a true positive and a false positive in terms of the device measurement, all positive results are treated the same. So drivers with false positives would also be prevented from driving, resulting in a major inconvenience. At the other extreme, an impaired driver could simply be given a warning. The overall value of the benefit from this system would be decreased because some fraction of impaired drivers would choose to drive in spite of the warning. However, the overall value of the harm would be significantly decreased because false positive drivers would not be inconvenienced to the same degree. It is possible to imagine other possibilities between these extremes, for example a car might be put into limp mode limiting its speed, or a system might send a message to a designated driver. The actual value of the benefit and the harm are subjective and will depend on the values of the person or group doing the evaluation, but in general when analyzing the value equation for a device, the question that should be asked is, does the benefit from using the system outweigh the harm? If the answer is yes, then the system should be put into use, and if no, other alternatives should be considered.

By way of an example, the net value can be calculated of an alcohol detection system using a risk-based model of outcomes assuming a device prevents impaired drivers from driving. A significant benefit from preventing impaired drivers from driving is deaths prevented. There are approximately 36,000 traffic fatalities in the U.S. annually, of which one third involve an alcohol impaired driver. Thus, if the device prevents impaired driving, the potential benefit if the device is implemented in all cars is 12,000 deaths prevented annually. The harm from this implementation comes from drivers inconvenienced (i.e., drivers inconvenienced per million drives.

Because of the high value of preventing deaths even a modest sensitivity can result in large benefits. For example, if an alcohol detection system prevented drivers that had a positive reading at or above a BrAC of 08% from driving, and that device only had a sensitivity of 50% it would still potentially prevent 6,000 deaths:

0.50 sensitivity * 12,000 deaths annually = 6,000 deaths prevented annually.

However, based on the enormous number of drives each year by sober drivers, a high specificity would still result in a large amount of harm. If the system prevented drivers with lower BrACs from driving if it detected a BrAC at or above 0.08% and the sensitivity was 99.9%, this would result in 1,000 drivers inconvenienced per million drives:

Inconvenienced Drivers per million Drives *(1-99.9%) * 1,000,000 drives = 1,000 drivers

In the U.S. there are an estimated 600,000,000 drives per day of which approximately 98% are made by sober drivers. With a rate of 1,000 inconvenienced drivers per million drives there would be 588,000 unimpaired drivers prevented from driving each day:

Prevented from driving * 1,000/1,000,000 * sober drives per day (98% * 600,000,000) = 588,000 drivers prevented from driving per day.

Thus, for the net value of the alcohol detection system to be positive, the specificity must be high enough to limit the inconvenience to drivers.

POTENTIAL WAYS TO IMPROVE SENSITIVITY OR SPECIFICITY BY TREATING TEST RESULTS DIFFERENTLY

Using the risk-based analysis of outcomes, it is possible to determine sensitivity and specificity thresholds at which the net benefit is positive. This in turn can be used by policy makers to set performance criteria for alcohol detection systems. Once those performance criteria are established, system manufacturers will need to ensure that their systems meet those thresholds. Understanding what these goals are, system manufactures can design devices that are optimized for the intended use. Obviously, manufacturers will want to make systems that are as accurate and precise as practical, but in addition, the way in which the devices are implemented can improve either the sensitivity or specificity of the device. For a device with a given accuracy and precision it would be expected that the specificity and sensitivity of that device would be affected for the population of all drivers by treating the test results differently. One option is that for a given definition of impaired, the threshold at which a positive result is reported can be modified. Another option is to combine more than one test result into a single outcome decision, which also can affect the sensitivity and specificity.

Increasing the BrAC threshold

The specificity of the device can be improved by increasing the threshold at which drivers are reported as positive. For example, setting the threshold for a positive report to 0.10% BrAC would dramatically increase the specificity. Considering drivers with a BrAC of 0.07% as shown in Figure 2, the device with the precision shown has a specificity of 90% with a threshold of 0.08%. However, moving the threshold to 0.10% changes the specificity for drivers with a 0.07% BrAC to 99.99, and drivers with a BrAC of 0.05% and 0.0% would have a specificity of > 99.999999. This would improve the specificity for the entire population of drivers from 99.97% to 99.9995%.

However, the change in the threshold comes at the cost of sensitivity. If the threshold for reporting a positive result is set at 0.10%, impaired drivers that had a BrAC between 0.08% and 0.10% would be reported positive at a very low rate. For example only 10% of drivers with a BrAC of 0.09% would be reported as positive. Drivers with a BrAC greater than 0.10% make up the majority of alcohol impaired drivers involved in fatal crashes so this might be an acceptable trade off. It can be estimated that the sensitivity of a device that has the accuracy and precision that have been used in our examples would decrease from 96% to 71% if the threshold for reporting a positive result is changed from 0.08% to 0.10%. Using risk-based analysis of outcomes the net benefit of these sorts of changes can be calculated. Using equation 1, our example device would drop from a benefit of 96% *12,000 deaths prevented, that is 11,520 deaths prevented, to only 71% *12,000 deaths prevented, that is 8,520 deaths prevented. However, the harm would be reduced from (1- 99.97%), that is, 300 drivers inconvenienced per million drives, to (1-99.9995%), or 5 drivers inconvenienced per million drives (see Figure 6).

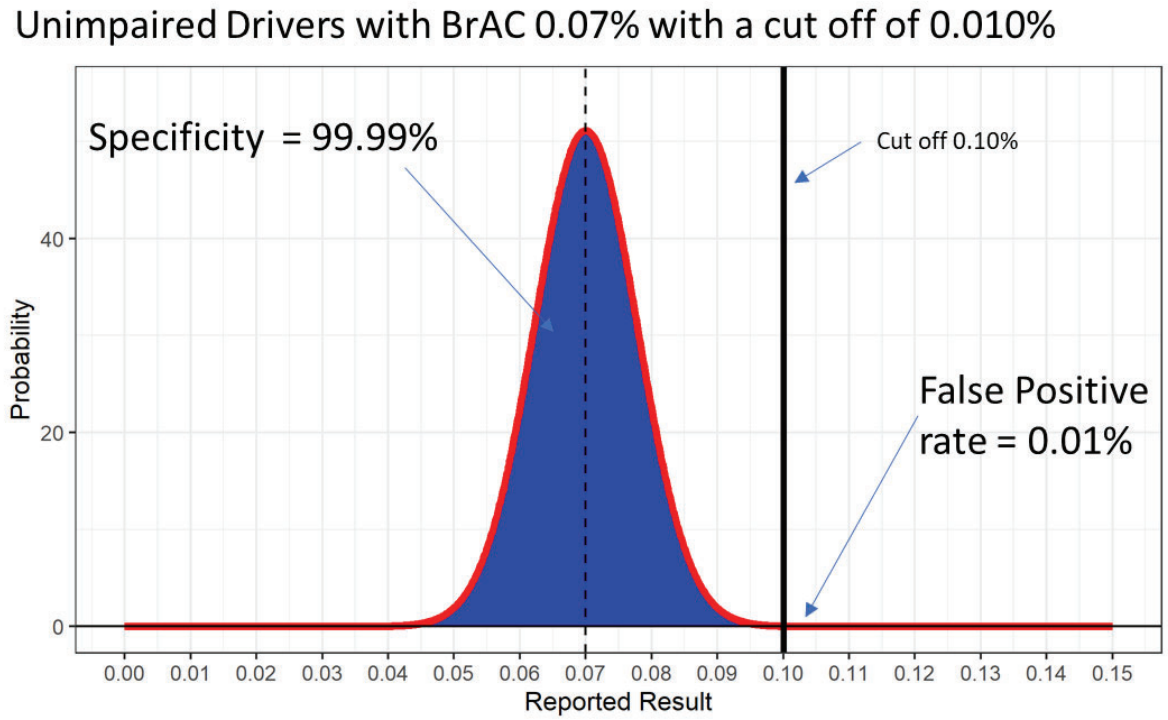

Figure 6. Unimpaired drivers with a BrAC 0.07% with a cut off of 0.01%

Decreasing the BrAC threshold

If sensitivity is more important than specificity, lowering the threshold has the effect of increasing sensitivity at the cost of specificity. For example, using the same process as described above, if the threshold is lowered from 0.08% BrAC to 0.06% BrAC the sensitivity of our example device increases from 96% to 99.98% increasing the benefit from 11,520 deaths prevented to 11,998 deaths prevented. However, specificity is lowered from 99.97% to 99.78% increasing inconvenienced drivers from 300 per million drives to 2134 per million drives.

Multiple breath tests

The second technique examined is the effect of combining more than one breath test into a single outcome. Averaging the results of multiple samples reduces the variability of the reported results. In general, the variability between outcomes (the standard deviation) is reduced by the square root of the number of samples averaged together. This reduction in the variability improves both the sensitivity and the specificity of a device but it comes at the cost of having to do multiple independent tests. In the case of a breath alcohol device, measuring multiple breaths could potentially increase inconvenience for the driver. At a minimum it would mean a somewhat longer time before a result could be delivered, potentially increasing the time before a car could be started. However, if the focus is on improving either sensitivity or specificity, retests could be undertaken only in certain conditions with the result that a much smaller number of drivers are inconvenienced. As we have demonstrated, because of the enormous number of drives every day it is very important to have very high specificity to avoid having the harm of the device outweigh the benefit. For example, if a device has a specificity of 99.9%, one in a thousand drivers would be inconvenienced. If a device asked for a retest sample for all drivers that were above 0.08% on initial analysis and only returned a positive result if both test results had a measured BrAC above 0.08%, then the sensitivity would be dramatically improved.

This effect is multiplicative, so the rate of harm (1-specificity) would improve from (1-99.9%) = 0.1% to (1-99.9%) * (1-99.9%) = 0.0001% or 99.9999% specificity (see Figure7). The rate of inconvenience would improve from 1 in 1000 drivers inconvenienced to 1 in 1,000,000 drivers inconvenienced. This does come at the cost of sensitivity. The impact on sensitivity is also multiplicative, so if the device had a 96% sensitivity the sensitivity would be reduced to

(96%) * (96%) = 92%. This is because some drivers who had a true BrAC greater than 0.08% could have a second measured result lower than 0.08% which would be considered a negative result.

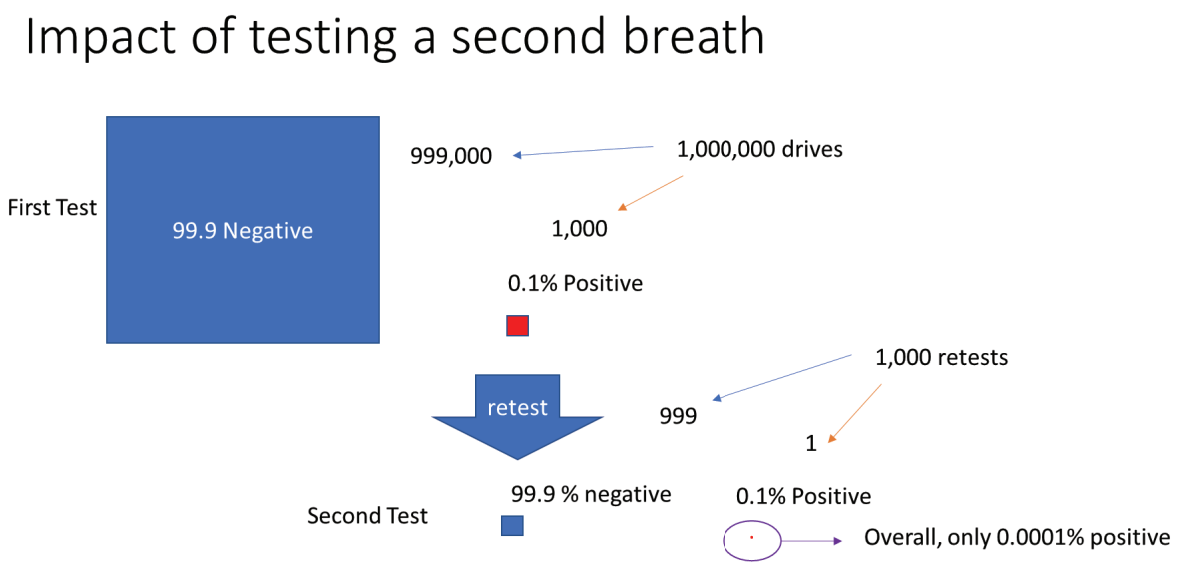

Figure 7. Impact of testing a second breath compared to the first.

ANALYSIS OF POTENTIAL OUTCOMES WITH THE CURRENT GENERATION 3.3 BREATH BASED DADSS SYSTEM

For the DADSS prototype generation 3.3 directed breath alcohol detection system, the sensitivity and specificity have been measured in multiple human subjects driving tests across a wide range of conditions and different alcohol levels. When using a limit of 0.08% as the definition of impaired, and weighting to account for the different levels of BrAC expected in the US population, the calculated sensitivity is 91.6% and the specificity is 99.93%. If we analyze the outcomes of using this device nationwide to prevent impaired driving we calculate a potential benefit of 10,992 deaths prevented (12,000 x .916 = 10,992) with a potential harm of 700 drivers inconvenienced per million drives ((1-0.9993) * 1,000,000). Using the method of retesting positive results as outlined above, there is a potential benefit of 10,069 deaths prevented (12,000 x 0.916 x 0.916 = 10,069) with potential harm reduced to only 0.5 drivers inconvenienced per million drives ((1-0.9993) * (1-0.9993) * 1,000,000).

CONCLUSIONS

The DADSS system, currently under development, has the potential to save thousands of lives a year by preventing drunk drivers from driving their vehicle. It is important to set device specifications that not only save as many lives as possible, but limit driver inconvenience to the extent possible. To calculate the appropriate specifications, a model for using a risk-based analysis of outcomes to determine the net value of implementation of a DADSS type system has been described. This should be a valuable tool that can be used by policy makers to guide their decision-making process in setting performance criteria for such devices. Specifically the models show that even modest sensitivity can have a significant positive benefit to the population because of the large number of deaths as a result of alcohol impaired driving. However, a very high specificity is required to prevent the negative impact from outweighing the benefit because of the enormous number of sober drives that take place every day.

That being said, there are many life-saving technologies currently in use in the vehicle fleet, such as frontal air bags, that also have potential negative effects on the driving population, and yet they are still acceptable to drivers. Initially, frontal air bags were developed to provide protection in frontal crashes for unbelted drivers and as a result were more powerful. However, these air bags were found to be overpowered resulting in injuries and deaths to occupants who were too close to them when they deployed. Frontal crash tests were modified to allow air bags to

deploy with less force, resulting in dramatic reductions in air bag injuries and deaths, but no loss of protection for front seat passengers (Ferguson and Schneider, 2008).

In summary, it is important to balance the positive and negative effects of new technology so that the maximum lives can be saved without potentially risking a backlash among the driving population. This paper has suggested ways in which these two can be balanced appropriately using a risk-based analysis of outcomes to determine the potential sensitivity and specificity, and hence accuracy and precision, of the device.

REFERENCES

American Automobile Association. 2022. Automatic Emergency Braking Performance in the Context of Common Crash Scenarios 2022, American Automobile Association Inc. https://newsroom.aaa.com/wp-content/uploads/2022/09/Research-Report-2022-AEB-Evaluation-FINAL-9-26-22.pdf

Cicchino J.B. 2017. Effectiveness of forward collision warning and autonomous emergency braking systems in reducing front-to-rear crash rates. Accident Analysis and Prevention, 99:142–152.

Federal Highway Administration. 2017. 2017 National Household Travel Survey, U.S. Department of Transportation, Washington, DC. Available online: https://nhts.ornl.gov.

Ferguson, S.A. 1996. Update on air bag performance in the United States: Benefits and problems. Airbag 2000+: Third Fourth International Symposium on Sophisticated Car Occupant Systems, 7:7-17, Fraunhofer- Institut Fur Chemische Technologie (ICT) Karlsruhe, Germany.

Ferguson, S.A. 1998. An update on the real-world experience of passenger air bags in the United States. Fourth International Symposium on Sophisticated Car Occupant Systems, 2-1:2-18, Fraunhofer- Institut Fur Chemische Technologie (. ICT) Karlsruhe, Germany.

Ferguson, S.A., Greene, M., Lund, A.K. 1995. Driver fatalities in 1985-1994 air bag cars. Insurance Institute for Highway Safety, Arlington, VA.

Ferguson S A, Traube E, Zaouk A, Strassburger R. 2011. Driver Alcohol Detection System For Safety (DADSS) – Phase I Prototype Testing And Finding. Paper Number 11-0230. Proceedings of the 22nd International Technical Conference on the Enhanced Safety of Vehicles.

Ferguson, S.A. & Schneider, L.W. 2008. An overview of frontal air bag performance with changes in frontal crash-test requirements: Findings of the Blue Ribbon Panel for the Evaluation of Advanced Technology Air Bags. Traffic Injury Prevention, 9:421-431.

Gross, A. 2022. Braking Bad: Automatic Emergency Braking Absent When You Need It Most. AAA. https://newsroom.aaa.com/2022/09/braking-bad-automatic-emergency-braking-absent-when-you-need-it-most/

Insurance Institute for Highway Safety. 2022. Alcohol and Drugs. https://www.iihs.org/topics/alcohol-and-drugs#by-the-numbers. Arlingon, VA.

Lund, A.K., Ferguson, S.A. 1995. Driver fatalities in 1985-1993 cars with airbags. Journal of Trauma, Vol. 38, 469-475.

National Center for Statistics and Analysis. 2018. Traffic Safety Facts, 2018 data. Occupant protection. DOT HS 812 967. National Highway Traffic Safety Administration. Washington DC.

National Highway Traffic Safety Administration. 2022. NHTSA Releases 2020 Traffic Crash Data. https://www.nhtsa.gov/press-releases/2020-traffic-crash-data-fatalities. Washington, DC

New York Times. 1974. Congress Clears Auto Safety Measure Eliminating Seat Belt Interlock System. https://www.nytimes.com/1974/10/16/archives/congress-clears-auto-safety-measure-eliminating-seat-belt-interlock.html

Ramirez A, Berning A, Kelley-Baker T, Lacey JH, Yao J, Tippetts AS … Compton R. 2016. 2013-2014 National Roadside Study of Alcohol and Drug Use by Drivers: Alcohol Results (Report No. DOT HS 812 362). Washington, DC: National Highway Traffic Safety Administration.

Zaouk A K, Willis M, Traube E, Strassburger R. 2019. Driver Alcohol Detection System for Safety (DADSS) – A non-regulatory approach in the research and development of vehicle safety technology to reduce alcohol-impaired driving - Status Update. Paper Number 19-0260. Proceedings of the 26th International Technical Conference on the Enhanced Safety of Vehicles.

CHARACTERISATION OF DROWSY DRIVER BEHAVIOUR AND DROWSINESS BASELINE DATA SET IN A DYNAMIC DRIVING SIMULATOR

Cristina Periago
James Jackson
Clara Cabutí
Fiona Azcarate
Elena Castro
Francesco Deiana
Adrià Roig

IDIADA Automotive Technology SA
Spain

Paper Number 23-0316

ABSTRACT

Drowsiness is one of the mean causes of road accidents, accounting for 1,200 fatalities and 76,000 injuries per year, according to several authors [1]. This transitional state between awake state and the sleep state behaves physiological symptoms such as yawning, loss of neck muscle tone, pupillary constriction, ptosis, decreased attention, psychomotor and cognitive performance [2]. The purpose of the present study is to observe the effects of monotonous driving on long journeys on driver behaviour in order to develop driver monitoring systems capable of detecting symptoms of drowsiness and thus be able to reduce its negative impact on the road. The experiment is conducted on a dynamic driving simulator, where conditions were configurated according to the aim of having a monotonous environment free from any distraction. Participants drive for 90 minutes and every 5 minutes the experimenter ask about their level of KSS, using the Karolinska Sleepiness Scale, a standardized instrument that measures the participant's subjective level of drowsiness. Moreover, participants are instrumented to collect physiological data (ECG, EEG, EDA and respiratory rate) and an eye-tracking system monitors other drowsiness behaviours such as blinking or yawning. The test finish when 90 minutes passed, or participants reached an advanced level of drowsiness on the Karolinska Sleepiness Scale (KSS). The study consists of two phases of testing. The first phase, with 10 participants, aims to validate the test method for both sleep induction and the integrated data collection setup. The second loop of testing, planned in January 2023, will involve 20 participants with different age and gender representation and aim to try to define the sleep behaviour patterns in relation to the different levels defined by KSS. In this paper we present the preliminary results of first phase of testing.

BACKGROUND

In recent years, it has been noticed that driving in a sleepy state poses a high risk to road safety. According to the DGT, drowsiness intervenes, directly or indirectly, in between 15-30% of traffic accidents in Spain [3]. Also, as reported by the recent statistics, drowsiness-related accidents account for 1,200 fatalities and 76,000 injuries per year [1] For this reason, there is growing interest in finding automatic systems capable of detecting the state of driver fatigue. In addition, as the implementation of this technology becomes more widespread, driven by the current regulations of entities whose objective is to reduce traffic accidents, such as the European Commission or EuroNCAP, the requirements for drowsiness detection systems validation tests are increasing. Validation tests can be conducted on test tracks (involving high cost and limitations by safety restrictions) and in a driving simulator (requiring a time and cost intensive integration process). Benefits of carrying it out in a laboratory-based driving simulator are the safety and the reproducibility of the experiments. [4]. To date, IDIADA has already developed specified methods for this type of testing, with a first successful application in proving ground since 2021.

Drowsiness definition

Drowsiness can be defined as the transition between the awake state and the sleep state where one's ability to observe and analyse are strongly reduced [5]. This transitional state usually goes hand in hand by physiological manifestations such as yawning, loss of neck muscle tone, pupillary constriction, ptosis, decreased attention, psychomotor and cognitive performance [2]. In addition, drowsiness mainly causes the following disruptions: increased reaction time, decreased concentration and more distractions, slower and more errors in decision making, motor disturbances and automatic behaviours, occurrence of micro dreams, sensory and perceptual disturbances, and changes in your behaviour [3].

Because of there are so many definitions of this concept, some authors disagree with each other. Even so, there are two concepts important to emphasise due to the contribution in developing different instruments that quantify drowsiness (instruments that we use in the present study): objective drowsiness and subjective drowsiness. The first refers to a person's tendency to fall asleep, and the second is considered as the subjective perception of the need to sleep associated with several subjective sensations and symptoms mentioned above [2]. Another way to measure drowsiness is gathering data from physiological parameters like electroencephalography (EEG), electrooculography (EOG), electrocardiography (ECG), respiratory rate and conductivity of the skin in where changes have been observed [6]. These changes include heart rate slowing, blinking, eyelid movement and breathing slowing, among others [7].

Purpose of the study

Considering the intention of increasing the benefits in terms of road safety for both driver and occupant [8] the aim of this study is to observe the effects of monotonous driving on long journeys on driver behaviour and to find patterns in variables for the development of driver monitoring systems.

OBJECTIVES

These testing activities have so-far involved representative inducements of sleepiness in naïve driver participants, with principal use of metrics for verification of sleepiness condition. Based upon discussions with existing and potential future automation industry needs, as identified two key areas of development for these types of sleepiness tests which form this paper objectives:

Technical objectives

The KSS (Karolinska sleepiness scale), regarded as the principal means of comparative evaluation, relies on participant subjective assessment. This additional objective measures have the potential to greatly improve the robustness of the assessment of participant condition. These are readily identified in literature, however there is a lack of a clear reference to critical KSS values.

Strategic objectives

Use of driving simulator and test tracks both involve prohibitive cost and timing implications for some validation activities. In response to this, a concept for a HiL (Human in the Loop) testing methodology has been developed, where relevant driver behaviour metrics would be fed into a client module for assessment of DMS sleepiness detection performance. Central to its potential implementation is the availability of a data set of relevant metrics for the detection of drowsiness / sleepiness in drivers with the KSS as a reference for subjective driver condition.

METHODOLOGY

Subjects

In this first phase of the study, 10 volunteers between 20 and 70 years were selected for the experiment, the proportion of which, between men and women, was almost equal: five and four respectively. Participants were separated into 4 groups during the course of the night: from 10 to 12pm, from 12 to 2am, from 2 to 4am and from 4 to 6am. This made it possible to englobe all the driving around the period of 2 pm to 4 pm, where it has been scientifically proven that the circadian rhythm renders a person more likely to get drowsy [3]. In addition, they were deprived of sleep by staying awake for the preceding 24 hours and were not allowed to drink coffee or any other type of stimulant either. All subjects signed a consent form, received a briefing (Annex 1), and did a previous questionnaire specifying driving characteristics and biometric measures as age, gender, height, skin tone among others (Annex 2).

Scenario's definition

Certain elements as the environmental stimulation, the time of the day, the hours of continuous wakefulness and driver's activity level can have an impact to the onset of drowsiness [3]. Taking this into consideration, we have implemented the following conditions for the test environment:

Table 1.
The conditions set for the KS-SLEEP tests

Time of the day	Participants did the test during the following interval of time: from 10pm to 6am. The scenario is configurated with to modes:

		- Day - Night
Wakefulness hours		Participants cannot sleep before the test to increase the hours that they are awake. They were not allowed to drink coffee or any other type of stimulant either.
Driver's activity level	Traffic	No random car traffic to avoid stimulating the driver's attention
	Luminosity of the environment	Dim light, that it is like at night and that the time also advances with the duration of the simulation
	Noises	No random noise. The scenario has been predefined with vehicle noise and ambient noise. To improve immersion, the sound of crossing the lane line has been reproduced.
	Speed	Sensation of speed at 80km/h. Annotation: We raise the speed in the simulator at 110 km/h to be able to reach this sensation.
Test track		A 20km highway loop, where you have a monotonous driving with a slightly changing environment in different places.
ACC ADAS system		The ADAS ACC system has been integrated into the vehicle model so that the participant can activate it during the test.

Measurement

Corresponding technologies and tools are used depending on the type of data. The data collected by the different data acquisition systems is integrated into the *iMotions* software. This software is an integrated analysis platform designed for human behavioural research and synchronizes all data obtained during testing.

Physiological sensors

The system used to collect EEG, ECG, EDA and respiratory data is *OpenSignals*. This *software* then shares the information with *iMotions*. The specifications of the sensors used are as follows:

a) Electroencephalography (EEG), one channel.
b) Electrocardiography (ECG), one channel.
c) Electrodermal Activity (EDA), one channel with two electrodes placed on the second and third finger of the hand.
d) Respiratory frequency with wearable chest-belt with an integrated localized sensing element.

Eye tracking

The system used for eye tracking is *SmartEyePro*. This system can determine the position of the head, the participant's features and iris and pupils' behaviour. Despite its multiple functionalities, the most relevant information for this project is eye opening and blinks. The instrumentation required to install this system consists of three cameras with their corresponding infrared light connected to the *SmartEyePro* computer where the software that processes the data is installed. This data is subsequently sent to *iMotions.*

The interest in this data collection lies in the measures of PERCLOS and blink measurement because they are one of the best indicators of drowsiness. Moreover, this data can be detected with non-intrusive, real-time detection systems, which is of benefit to users.

KSS

Karolinska Sleepiness Scale (KSS) is a 9-point scale able to measure the subjective level of sleepiness indicating which level is more in line with the psychophysical condition experienced [9]. It has been used in some studies to assess driving abilities and fatigue [10] [11]. Current regulations state that sleep monitoring systems must warn the driver when the driver is at KSS level 7 or higher.

This self-rated scale is assessed every 5 minutes by the experimenter asking to the volunteer "What is your perception of drowsiness in the last 5 minutes?". The participants were informed previously that was important to understand the scale for the proper functioning of the test. The experiment finalises when the participants reach the 8 level of drowsiness (where it is considered a certain effort to keep alert) or when 90 minutes have passed. Drowsiness subjective level is defined as follows:

Table 2.
Karolinska Sleepiness Scale (KSS)

Rating	Verbal descriptions
1	Extremely alert
2	Very alert
3	Alert
4	Fairly alert
5	Neither alert nor sleepy
6	Some signs of sleepiness
7	Sleepy, but no effort to keep alert
8	Sleepy, some effort to keep alert
9	Very sleep, great effort to keep alert, fighting sleep

Driving simulator
One of the main axes of this research is the Dynamic Driving Simulator: a cutting-edge tool with high added value that allows you to drive and experience dynamic driving sensations close to reality in a totally safe environment.

The simulator set-up consists of a cockpit based on a real VW Golf Variant 8 mounted on a dynamic platform with 9 degrees of freedom. The platform consists of a tripod for low frequencies and a hexapod for high frequencies which allows having movement and vibrations sensations while driving.

The cockpit interior is based on the actual vehicle, so all interior details are virtually the same. In addition, the cockpit also incorporates a parameterized active steering wheel and brake pedal, as well as a pneumatic seat and seat belt, which offer a more than correct response to the limitations of the platform movement when maintaining constant longitudinal and transversal accelerations. The driver's position inside the cockpit has no blind spots in terms of visibility and immersion within the virtual environment. The visual component is very important, so the driver will be surrounded by 5 fully merged and synchronized conical screens with a 240-degree field of view at frequency of 120Hz and 2k resolution.

At the software level, two different company tools are integrated. On the one hand we have the dynamic simulator software, provided by Vi-Grade, and on the other hand we have the virtual environment software, from AV Simulation. Both are being co-simulated and are fully integrated. The communication between the simulator computers and *SCANeR* is done entirely through the UDP protocol.

VI-Grade
VI-Grade is the company which provides the simulation tools and licenses. This includes different software such as SIMulation Workbench (for real-time execution and configuration of the processes and tests to be executed), VI-DriveSim (as a slightly more user-friendly interface of SIMulation Workbench), VI-CarRealTime (for the dynamics of the vehicle), VI-SimSound (for sound), etc.

AV Simulation
The AV SIMulation tool used in this study is comprehensive simulation platform called SCANeR Studio. SCANeR software takes care of everything related to the environment in which the car moves: the scenery. In addition, it also controls the interaction of the scenery with all the elements that appear on it. SCANeR is used to design the route and what the participant will see. The designed route runs along a 20km highway loop which is practically a linear road. In this way, a calm and smooth driving is achieved so that the participants have a high state of comfort and relaxation during all tests.

Procedure
The experiment has been distributed over 3 days, with 3 participants on the first day, 4 on the second day and 3 on the third day. Subjects were informed of the content of the study and signed the consent forms previously. They also completed a questionnaire beforehand (explained in Annex 1). Participants were transported by taxi from their home to our facilities to avoid driving before the test, which would have affected their alertness, and

on their return home, to avoid the risk of driving after having been induced to sleep and due to the recommendation not to drive 30 minutes after driving on the simulator.

The experiments were preceded by a few minutes of pre-driving to make the participant comfortable with the driving simulator. Subjects were instructed to drive safely, respecting traffic rules, and behaving as close as possible to a real situation. When the participant felt adapted, the driving simulation started and was extended for 90 min. Procedures were prepared to begin the driving session around 2 pm. This made it possible to englobe all the driving within the period of 2 pm to 4 pm, where it has been scientifically proven that the circadian rhythm renders a person more likely to get drowsy [3]. Aside from the driving task, participants were asked to rate their levels of drowsiness/alertness through the KSS, in 5-minute intervals.

PRELIMINARY RESULTS AND DISCUSSION

Expected results from this paper was obtaining objective measures to give robustness to the subjective Drowsiness Detection Assessment and to develop a drowsiness baseline data set. This also will be useful for the current development of ADAS, and ADS technologies related with driver drowsiness detection.

Objective measures of sleepiness are identified in the literature, but there is no clear correlation with KSS values, suggesting the complexity of identifying behavioural patterns easily generalisable to the population. In this first phase of the analysis, we have focused on validating the methodology to be applied in a second phase of the experiment and, secondly, on obtaining a data profile in relation to the results expected by the literature and exploring a possible relationship with the level of KSS. The study is still ongoing, and although we will not draw firm conclusions at this stage, it is a good opportunity to analyse objective data on drowsy behaviour collected under controlled test conditions. Thus, the following is a sample of some of the data obtained by one of the participants:

Physiological data

Regarding heart rate, the expected results are that it decreases as the level of drowsiness increases, being aware that the normal heart rate ranges between 60-100 beats per minute (bpm). Two further indicators that participants are approaching a state of drowsiness are increased yawn frequency and decreased respiratory rate. Normal breathing in adults is regarded as between 12 and 25 breaths per minute. In both cases the trend of the data is as expected. Figure 1, Figure 2 and Figure 3 show participant results in relation to the declared KSS level.

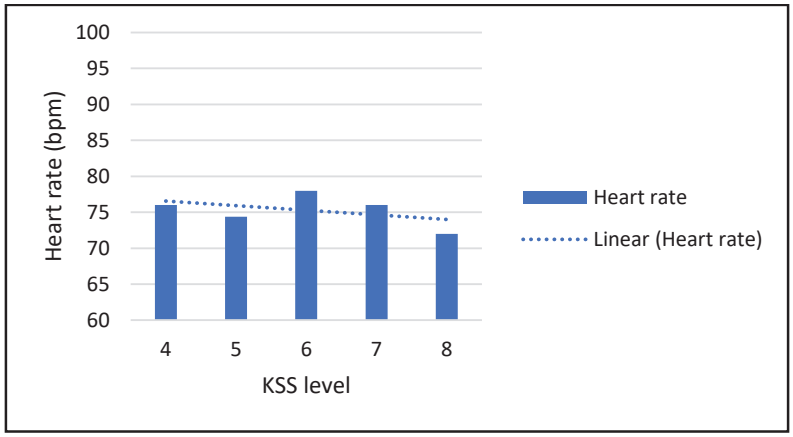

Figure 1. Heart rate-KSS level.

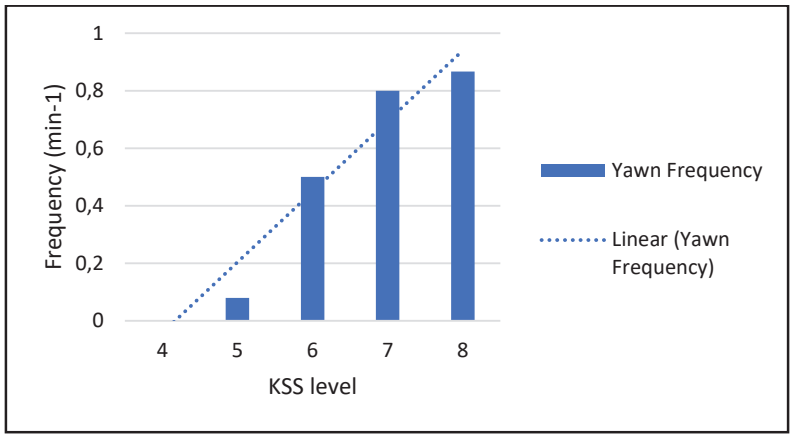

Figure 2. Yawn frequency-KSS level.

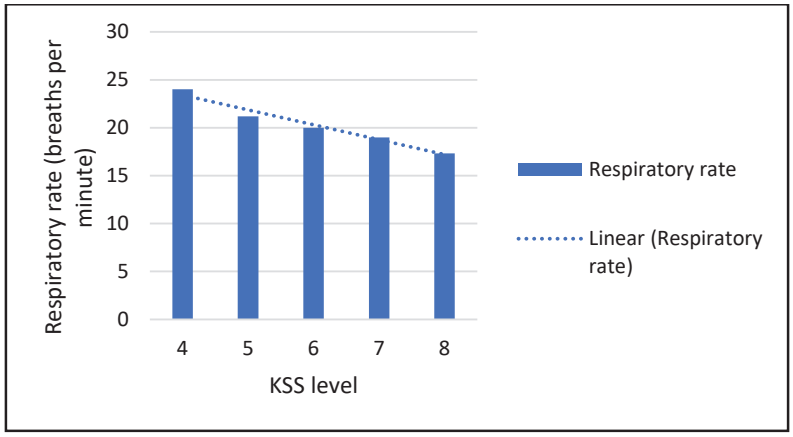

Figure 3. Respiratory rate-KSS level.

CONCLUSIONS AND NEXT STEPS

In this paper we present data that validate the method proposed in this first phase. This will be used to apply it in a second testing phase. The most relevant observed trends concerning one participant are: a) A relation between KSS level and yawning frequency: as the level of KSS increases, so does the frequency of yawning. This is not best explained by the time the participant resides in each KSS phase, as it refers to KSS=5 level for 25 minutes, while KSS=8 level only lasts for 15 min.; b) Regarding heart rate, it seems that decrease while KSS level increase. Even so, empirical assessment of the data will be necessary to draw definitive conclusions; and c) Respiratory rate decrease as expected while KSS level increase.

Once the methodology has been designed and validated, two parallel future lines of action emerge. On the one hand, consolidate the application of the methodology in the second testing phase in order to be able to reach definitive conclusions and, on the other hand, its improvement and extension.

- After a first analysis of this first loop of testing, the following aspects have been detected and should be considered for the second testing phase: Initially, the premise was to reach KSS level 8. However, as the simulator was assessed as a safe and risk-free space, it was considered to increase this level to score 9 to assess a further stage of sleepiness and its effects.
- Participants: it is important to EEG sensors signal that the volunteers come without products on hair. It has been observed that this interferes with the sensor signal.
- Extension of the instrumentation: it is proposed to extend the ECG sensors instrumentation to optimise their signal and to place some more cameras to obtain better visualisation of the participant's behaviour.
- Instrumentation variation: change EDA sensors (Electrodermal Activity) to obtain clearer results than with the present sensors.

As for the application of this methodology in future development, these are the main proposals for its implementation:

- Assessment of the effect of drowsiness on drivers.
- Comparative analysis of different drowsiness detection systems.
- Evaluation of the influence of distractions on the evolution of sleepiness.
- Validation of new systems in an autonomous vehicle.
- New users: the idea is to be able to apply the methodology developed to new groups of users.

REFERENCES

[1] Saini, V., & Saini, R. (2014). Driver drowsiness detection system and techniques: a review. International Journal of Computer Science and Information Technologies, 5(3), 4245-4249.

[2] Rosales Mayor, Edmundo, & Rey De Castro Mujica, Jorge. (2010). Somnolencia: Qué es, qué la causa y cómo se mide. *Acta Médica Peruana*, *27*(2), 137-143. Recuperado en 05 de diciembre de 2022, de http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1728-59172010000200010&lng=es&tlng=es.

[3] Pérez, P. S. (n.d.). Conducir con sueño o cansancio. Retrieved from DGT: https://www.dgt.es/muevete-con-seguridad/evita-conductas-de-riesgo/Conducir-con-sueno-o-cansancio

[4] Rimini-Doering, M., Manstetten, D., Altmueller, T., Ladstaetter, U., & Mahler, M. (2001, August). Monitoring driver drowsiness and stress in a driving simulator. *In Driving Assesment Conference* (Vol. 1, No. 2001). University of Iowa.

[5] Picot, A., Charbonnier, S., & Caplier, A. (2011). On-line detection of drowsiness using brain and visual information. IEEE Transactions on systems, man, and cybernetics-part A: systems and humans, 42(3), 764-775.

[6] García Daza, I. (2011). Detección de fatiga en conductores mediante fusión de sistemas ADAS.

[7] Bergasa, L. M., Nuevo, J., Sotelo, M. A., Barea, R., & Lopez, M. E. (2006). Real-time system for monitoring driver vigilance. IEEE Transactions on Intelligent Transportation Systems, 7(1), 63-77.

[8] Jannusch, T. S. (2021). Cars and distraction: How to address the limits of Driver Monitoring Systems and improve safety benefits using evidence from German young drivers. Technology in Society, 66, 101628.

[9] Shahid, A., Wilkinson, K., Marcu, S., & Shapiro, C. M. (2011). Karolinska sleepiness scale (KSS). In STOP, THAT and one hundred other sleep scales (pp. 209-210). Springer, New York, NY.

[10] Soares, S., Monteiro, T., Lobo, A., Couto, A., Cunha, L., & Ferreira, S. (2020). Analyzing driver drowsiness: From causes to effects. Sustainability, 12(5), 1971.

[11] Sommer, D., & Golz, M. (2010, August). Evaluation of PERCLOS based current fatigue monitoring technologies. In 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology (pp. 4456-4459). IEEE.

APPENDICES

Appendix A

Participants' briefing

Thank you for participating in this study. In this document we explain the details of the test that you will voluntarily take as part of a research project of the Human Factors team of the ADAS department at APPLUS+ IDIADA.

Before it is important that you understand why the research is being conducted and what it consists of. Please read this document carefully and ask the research staff any questions you may have. You will be asked to sign a consent form confirming that you have read and agreed to all the information contained in this document. In addition, you will be given a document requesting your consent to share the data recorded during the test with the client responsible for the study.

It is important that before the test you do not drink stimulating substances or beverages such as coffee, nor are you under the influence of any medication, alcohol or other drugs that could affect your alertness during the test. The research staff will inform you when the test is completed.

Description of the activity

The aim of this test is to observe the effects of monotonous driving on long journeys on driver behaviour. This type of study is important for the development of new driver monitoring systems capable of improving the safety of both driver and occupants.

The test will be conducted on the dynamic driving simulator. During the test, you will be asked to drive for 90 minutes in the centre of the lane at a constant speed.

Every 5 minutes, the researcher will ask you about your drowsiness, which will be assessed using the Karolinska Drowsiness Scale. It is important that you carefully read and understand the scale for the proper functioning of the test. It is defined in Table A1.

Table A1.
Karolinska Sleepiness Scale (KSS)

Rating	Verbal descriptions
1	Extremely alert
2	Very alert
3	Alert
4	Fairly alert
5	Neither alert nor sleepy
6	Some signs of sleepiness
7	Sleepy, but no effort to keep alert
8	Sleepy, some effort to keep alert
9	Very sleep, great effort to keep alert, fighting sleep

You will report your state of sleepiness by giving a value from 1 to 9 according to your actual state. In the course of the test, we will collect data from different sources: your driving behaviour, your interaction with the vehicle controls and your visual behaviour while driving will be recorded. In order to measure behavioural effects, we will instrument you with a total of 9 sensors (distributed on the head, abdomen and right hand) that will collect physiological signals such as heart rate, respiratory, skin conductivity and brain waves.

During the first 5 minutes of driving feel free to ask any questions you may have, but please remain silent for the rest of the test to avoid influencing your alertness through no fault of your own.

We remind you that participation is entirely voluntary. If for any reason you wish to stop driving, please inform us and you are free to leave the test.

Appendix B

Table B1.
Biometric data

1. Age
2. Gender
3. Driving experience (years)
4. Overall driving frequency (hours/week)
5. Commuting frequency (hours/week)
6. Skin tone (Fitzpatrick scale)
7. Standing stature (cm)
8. Maximum distance between eyes (mm)
9. Minimum distance between eyes (mm)
10. Nose length (mm)
11. Vertical relaxed left eye-opening aperture (mm)
12. Vertical max left eye opening aperture
13. Horizontal relaxed left eye-opening aperture
14. Vertical relaxed right eye-opening aperture (mm)
15. Vertical max right eye-opening aperture
16. Horizontal relaxed right eye-opening aperture

EXPLORING DRIVER DISTRACTION IN ADAPTATION TO LOWER LEVELS OF AUTOMATION: OLDER ADULT DRIVER COMPARISONS

Jon Antin, Charlie Klauer, and Shu Han
Virginia Tech Transportation Institute (VTTI)
USA

Thomas Fincannon
National Highway Traffic Safety Administration (NHTSA)
USA

Paper number 23-0318

ABSTRACT

This project examined how middle-aged and older drivers adapt to the use of Level 2 (L2) advanced driver assistance system (ADAS) features (i.e., the system controls lateral and longitudinal motion). Data were drawn from two naturalistic driving studies (NDS). In the *L2 NDS* study, 82 participants were recruited from the Washington, DC metro area and drove L2 vehicles for four weeks. A second NDS was conducted with 14 older adults (*Older Driver NDS*). In the Older Driver NDS, participants aged 70-79 drove L2 vehicles for six weeks. Speed setting above the speed limit was significantly more common when L2 was active than when it was available-but-inactive in the *Older Driver NDS* dataset. Older adults had shorter off-road glances than middle-aged drivers in the *L2 NDS* when L2 was available, regardless of L2 engagement status. Older drivers showed shorter glance durations overall. Older adult drivers had fewer glances away from the forward roadway and were significantly less likely to engage in secondary tasks when L2 was active. Evidence of older adult driver adaptation to L2 systems is seen most predominantly in the speed selection.

INTRODUCTION

Driving automation technology is rapidly proliferating into the U.S. vehicle fleet. Along with this trend, our society is aging. In light of this, questions remain regarding how older adult drivers adapt to novel technologies in the driving environment and how drivers respond to the introduction of new technology that serves specific needs. The primary research objective was to compare middle-aged and older adult driver safety behaviors and adaptation during initial exposure to SAE International (SAE) Level 2 driving automation (L2) advanced driver assistance systems (ADAS).

Vehicle automation control paradigms are becoming more novel across a variety of dimensions, specifically L2, where some degree of driving automation control is simultaneously exerted in the longitudinal as well as lateral dimensions. Typically, longitudinal control is manifested by adaptive cruise control (ACC) and lateral control via lane keep assistance (LKA) or lane centering assistance (LCA). Additionally, we are witnessing an aging of our society.. While there are theoretical approaches to conceptualizing driver adaptation (e.g., *risk homeostasis*, [1,2]; *risk allostasis*, [3]), questions remain about how older adult drivers specifically adapt to automation in the driving environment. The construct of adaptation is perhaps even more complex with older drivers in that cognitive decline may progress with age.

A few studies have examined this space. In a study focused on situation awareness in a simulated driving environment, researchers found that a group of older drivers (65–81) adapted to dynamic hazards with greater vehicle speed reduction than a group of younger drivers (18–25) [5]. While each study is unique, the naturalistic driving study (NDS) research paradigm typically involves the automatic recording of driver behaviors, vehicle kinematics (including speed and acceleration), and a GPS record of the vehicle's route driven. Liang and colleagues used the NDS paradigm to investigate older drivers' subjective adaptation to ADAS, including ACC over a 6-week period [6]. Weekly phone surveys found little change in the older drivers' trust of the ADAS features: they generally started high and remained at that level. However, focus group discussions conducted after the conclusion of the driving portion of the study did reveal attitudinal adaptation to the technologies across several dimensions, including perceived safety and functional benefits as well as confidence in the technology.

Objective

The objective of these analyses is to compare middle-aged and older adult driver safety behaviors during initial exposure to L2-equipped driving automation (i.e., driving automation to simultaneously control lateral and longitudinal motion, but where driver expected to maintain constant supervision of these support features and maintain responsibility for driving). This analysis provides a comparison that identifies how older drivers adapt to driving automation to discuss potential unique needs of that population.

METHOD

Two NDS databases were used to compare older adult drivers with a group of middle-aged drivers in the earliest phases of L2 technology use. The *Older Driver NDS* was conducted with 18 older drivers (70–79) [7]. Participants drove one of four L2-equipped vehicles for 6 weeks each. Vehicle makes included Audi, Infiniti, Mercedes, and Volvo from the 2015 – 2017 model years. Participants were eligible for the study if they met the age group criterion and had not driven L2-equipped vehicles. The *L2 NDS* study provided a database of middle-aged drivers, 25-54, for comparison with the older adults in the Older Driver NDS [8]. The same vehicles as were driven in the Older Driver NDS were also driven in the L2 NDS; however, the L2 NDS also included a Tesla. L2 NDS participants lived and commuted in the Washington, D.C. area. Participants in the Older Driver NDS were residents of the Blacksburg, Virginia, and surrounding areas. Data from both datasets were compared over the first 3 hours of driving exposure *with use of the L2 technologies*. While drivers in the L2 NDS database had more exposure to L2 technologies, older drivers did not use systems more than 3 hours, so older driver usage limited the amount of L2 exposure that could be used in this analysis.

Data were analyzed from 96 volunteer driver participants. This included 14 drivers from the Older Driver NDS and 82 drivers from the L2 NDS. There were 2,437 L2 activations, which included 130 activations from the Older Driver NDS and 2,307 activations from the L2 NDS. These observations were collected across 3,891

trips, including 370 trips from the Older Driver NDS and 3,521 trips from the L2 NDS. Table 1 provides a summary of these observations.

Table 1. Summary of Older Driver NDS and L2 NDS Data Sources

Data	Older Driver NDS (70+ years old)	L2 NDS (25 to 54 years old)
Drivers	14	82
L2 activations	130	2,307
Trips	370	3,521

The primary independent variable focused on L2 Activation Status, where the driver (i.e., systems were available but inactive) or the driving automation (i.e., the systems were both available and active) controlled both lateral and longitudinal motions of the vehicle. Additional independent variables include Time of Day, Day, Road Type, and Traffic Density. The main dependent measures included: speed selection, glance behaviors, and secondary task engagement. While this study intended to examine changes in driver behavior over time, older drivers did not use the systems enough (i.e., no more than 3 hours of experience with active L2 systems) to conduct this analysis.

RESULTS

Across both datasets, the duration of each individual L2 activation event was similar. In the Older Driver NDS, 90 percent of activations were shorter than ten minutes; in the L2 NDS, 99 percent of activations were shorter than ten minutes. This analysis evaluated multiple circumstances that included time of day, weekday versus weekend, and road type, but these factors were not found to influence any of the variables in this analysis.

Speed Selection

For each L2 system activation, vehicle speed, GPS coordinates, and road type were recorded. Each L2 system feature activation analyzed was required to be at least 120 seconds in duration to ensure the driver intentionally activated the L2 system features. Then, a random sample of matched controls was identified with the goal of a 1:1 match based upon driver identification number, Time of Day (± 4 hours), day of week (weekday versus weekend), and anytime the vehicle was traveling above 40 mph. The idea of *available-but-inactive* is important in ensuring comparisons are reasonable. That is, if comparisons were made between L2 usage periods and *all* non-L2 usage periods, any observed differences could readily be attributed to the different conditions, scenarios, and driving environments in which drivers tend to - or are permitted to - engage L2 systems. Using available-but-inactive driving epochs to provide control samples makes usage/non-usage comparisons more meaningful.

The speed profiles are shown in

Figure **1** as a histogram of the difference between speed limit and actual speed. Frequencies in each bin are plotted as a percentage of total events. Note that in these figures, the middle of the x-axis (zero) is representative of adhering to the speed limit (and is marked with blue vertical lines). Histogram bars to the right of zero indicate traveling faster than the speed limit, and bars to the left indicate traveling slower. Both L2 status, $F(1, 18{,}010) = 1{,}157.35$, $p < 0.001$, and driver groups, $F(1, 78) = 14.26$, $p < 0.001$, demonstrated statistical significance in speeding behaviors. When L2 systems were active, drivers tended to select speeds which were 4.4 mph faster than the average speed driven when the L2 system was available-but-inactive. In addition, middle-aged L2 NDS participants tended to drive 5.2 mph on average faster than the Older Driver NDS participants. In addition, the frequency with which older adults selected speeds over the speed limit was significantly more common when L2 was active than available-but-inactive ($F(1, 1{,}230) = 425.71$, $p < 0.001$).

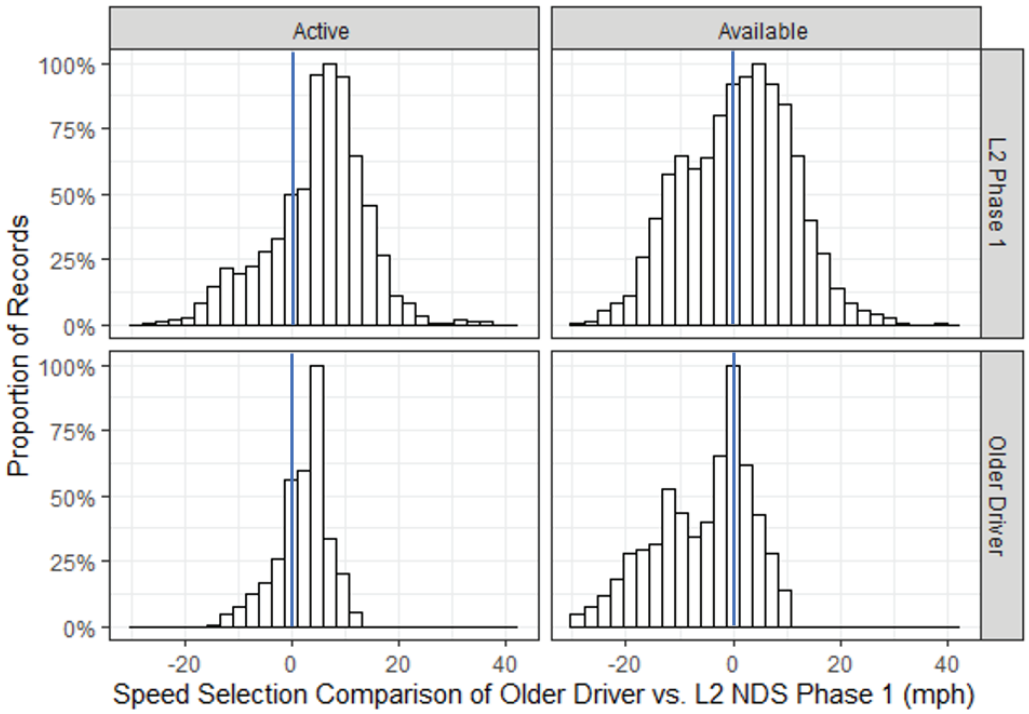

Figure 1. Histograms of speed selection profiles relation to speed limit comparison by L2 status and participant group (Older Driver vs. L2 NDS Phase 1) – speed limit represented by blue vertical lines.

Glance Behavior

Eyes-off-road variables for all eye glances away from the forward roadway are shown in

Figure 2. This figure displays four graphs of eye-glance data across both datasets for samples with L2 active and samples with L2 available-but-inactive for both the Older Driver NDS and the L2 NDS. There were several samples where eye-glances away from the forward roadway did not occur. This resulted in several values of zero in the dataset. The zeros were removed before analyzing these data to provide a clearer analysis of what eye-glance behavior looks like when it does occur. ANOVA tests were used to analyze all four metrics of eye-glance data.

Total off-road glance duration is plotted in the top left panel (**A**). Results showed that participants in the Older Driver NDS had shorter total off-road glance duration (i.e., per driver) than L2 NDS participants, both when L2 was active and when L2 was available but inactive. This was evidenced by a main effect of driver group, $F(1, 609) = 6.58$, $p = .01$. There was no significant main effect of whether L2 features were active versus available-but-inactive on total glance duration ($F(1, 609) = 3.02$, $p = 0.083$). Mean off-road glance duration is plotted in the top right panel (**B**), and showed main effects of both L2 active, $F(1, 609) = 9.00$, $p = .003$, and driver age group, $F(1, 609) = 7.36$, $p = .007$. Overall, L2 NDS drivers had longer mean glance durations compared to those in the Older Driver NDS. Single longest off-road glance (bottom-left panel, **C**) showed main effects of both L2 active, $F(1, 609) = 8.44$, $p = .003$, and driver age group, $F(1, 609) = 4.845$, $p = .028$. This indicated that longest glances were longer overall when L2 was active for the L2 NDS drivers, but older drivers' longest glances were shorter than those of the middle-aged L2 NDS participants. Finally, number of off-road glances showed a significant main effect of driver group (lower right panel, **D**), $F(1, 609) = 4.47$, $p = .035$, such that older drivers had fewer glances away from the forward roadway. There was no main effect of L2 active vs. available-but-inactive ($F(1, 609) = 0.043$, $p > 0.05$).

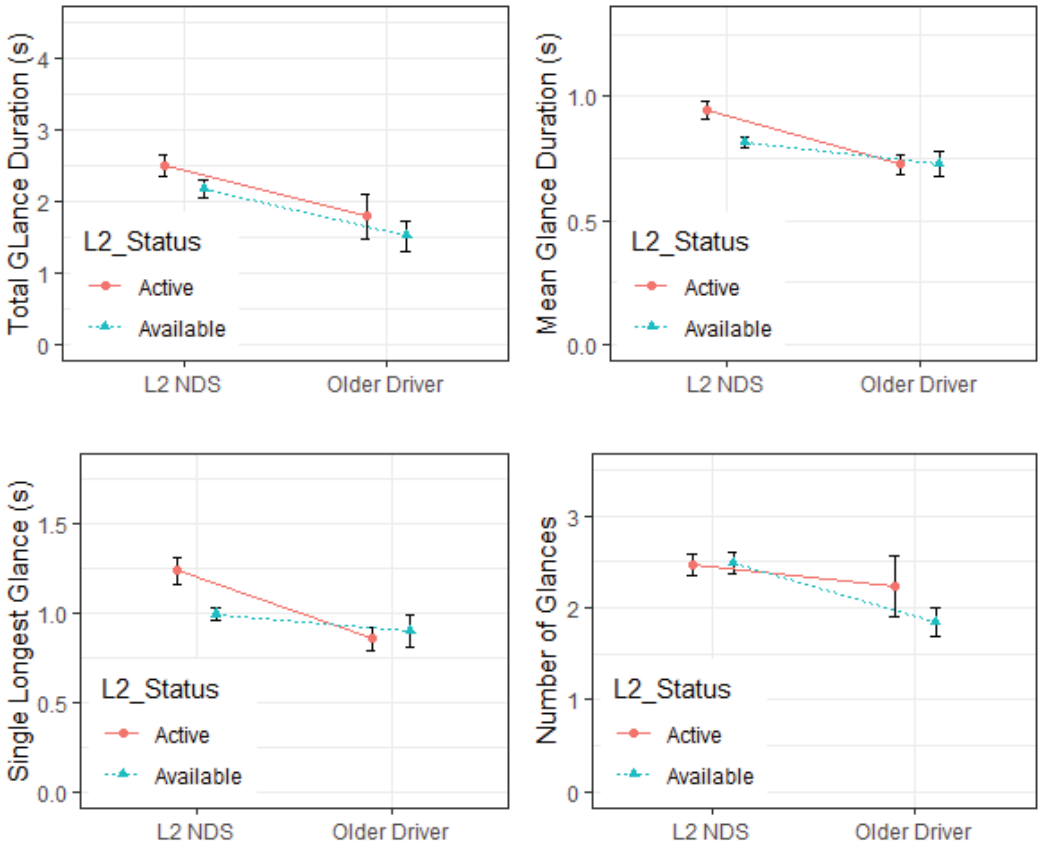

Figure 2. Eye glance metrics for Older Driver NDS vs. L2 NDS participants.

Secondary Task Engagement

Analyses for secondary task engagement were completed using the coded data where samples were randomly selected based upon whether L2 systems were active or available-but-inactive. Older adult drivers were significantly less likely to engage in secondary tasks when L2 was active compared to their middle-aged counterparts in the L2 NDS drivers, χ^2 (1, N = 792) = 4.22, p = 0.04. There were five categories of distraction compared between the two datasets: visual, manual, visual manual, cognitive, and no distraction. See Figure 3 for the percentage of each type of distraction observed in both datasets across L2 active and L2 available-but-inactive samples.

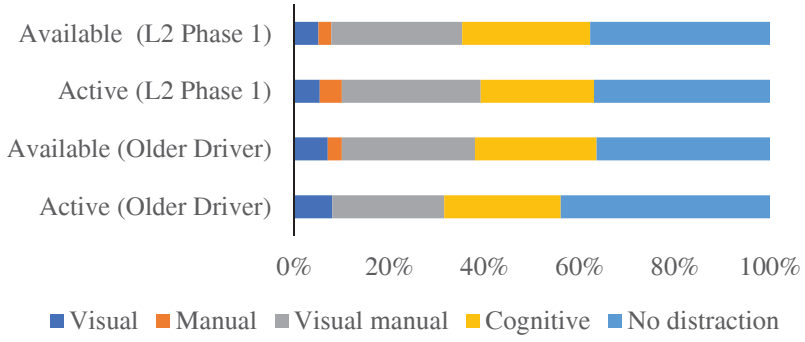

Figure 3. Secondary task distribution percentages comparison by L2 status and participant group (Older Driver vs. L2 NDS Phase 1).

Discussion

Speeding behavior results showed that older adults were more likely to speed with the L2 systems active compared with when they were available-but-inactive. This could demonstrate risk allostasis, wherein drivers are adjusting their behaviors to maintain a preferred level of risk when they perceive that active systems are safe and provide a protective effect. This may also represent drivers perceiving greater risk when systems were available but inactive, where they attend more to the roadway and drive more slowly or cautiously. Alternately, this could be a related to design (e.g., the default settings), where systems let ACC deviate above the posted speed and to set follow distances that may not match the driver's personal driving style. While a separate analysis from the L2 NDS study showed that middle-aged drivers selected increasingly higher speeds over time, older drivers did not use L2 ADAS feature more than 3 hours, which may reduce changes associated with time and allow for the impact of default settings in this demographic.

The glance analyses paint a clear picture. Older adult drivers demonstrated eyes-off-road glance patterns which were shorter in overall, mean, and longest-single glance duration, and the older drivers looked away from the forward roadway less frequently. This coincides with past research (prior to the L2 era) which has shown that older drivers scan less or have a narrower gaze dispersion in certain driving scenarios (e.g., intersection traversals, [9]). Following on this work, researchers examined several underlying conditions which might explain why older drivers demonstrate a more focused scan pattern at intersections. These included head movement limitations, memory-related issues, and distractibility. However, they found that none of these fully explained the observed behavior. Instead, they determined that the propensity was, in effect, older drivers' conscious (or unconscious) attempts to adapt to their own perceived functional decrements. The fact that this behavior also had potentially maladaptive consequences (i.e., missed emergency cues outside of the direct forward view) was unknown to the older drivers [10]. From a transportation safety perspective, these results and conceptualizations present a conundrum. On the one hand, we might interpret the glance behaviors observed in this study as older drivers demonstrating generally greater caution and less distraction, perhaps based on their greater maturity, experience, and very low risk tolerance. However, the research noted above by Romoser and his colleagues paints a different picture, wherein older adults demonstrate a glance pattern that may be too focused or narrow to effectively detect important emergency cues that may appear in the periphery. However, in the current study, an eyes-off-road glance was defined as one not directed to any of the following locations: forward, left window/mirror, left windshield, rearview mirror, right window/mirror, right windshield, or the instrument cluster. Thus, the phenomenon noted by Romoser and colleagues may be fundamentally different than the findings reported in the current study.

While the small sample size for older drivers can be problematic, the secondary task analysis from this study also indicates that older drivers behave more cautiously than their younger counterparts during L2 activation, demonstrating not only a lower percentage of visual and/or manual secondary task time, but also a greater percentage of time with no secondary task of any kind. When the secondary tasks were broken out by low versus high risk, the pattern was less clear, as the older drivers engaged in low-risk tasks during L2 activation at a lower percentage than their younger counterparts; both groups engaged in high-risk tasks at the same percentage. It should be noted that the designation of secondary tasks into low and high-risk categories is based on tasks and driving data observed or collected in the pre-L2 era. While L2 technology may improve safety, the risk levels of specific secondary tasks while L2 is active must be investigated empirically (i.e., considering risk allostasis).

LIMITATIONS

Participants in the L2 NDS were assumed to have little to no previous experience with L2 features. This is an assumption, in that researchers knew that participants had never driven the make/model of the instrumented vehicle assigned to them for data collection, and thus the specific L2 features were novel. While the drivers in these two studies drove a similar set of vehicle make/models, there were differences in how the various OEMs implemented the L2 technologies that were not directly tested or compared in these analyses.

The two datasets were collected in two different regions and driving environments. The L2 NDS drivers were commuters in the Washington, DC area, whereas participants in the Older Driver NDS were residents of a largely suburban and rural area of Blacksburg, Virginia. Thus, the driving environments that these two sets of drivers negotiated were different, and it is impossible to control for this difference in the analyses. While the vehicles in the L2 NDS and Older Driver NDS were similar, the L2 NDS had more variety of vehicles, which may have impacted

findings. Another limitation worthy of consideration is that the Older Driver NDS included only a pilot sample of 18 participants, which was smaller than the 82 participants from the L2NDS dataset.

CONCLUSIONS

Evidence of driver adaptation to L2 ADAS may be seen most predominantly in the speed selection analysis. When L2 systems were active, on average, drivers tended to select speeds which were 4.4 mph faster than the average speed driven when the L2 system was available but inactive. In addition, middle-aged drivers (L2 NDS) tended to drive 5.2 mph on average faster than older drivers (Older Driver NDS), but older drivers still selected speeds that were above the speed limit more when L2 ADAS features were active. Speed-selection is related to only one aspect of L2 control, which is often available for independent use as well (such as in the form of Adaptive Cruise Control). The analyzed datasets did not have sufficient instances where L2 was available but only ACC was engaged to be included in speed-selection analysis. Therefore, it is unknown if or how much of the observed effects in speed-selection may be due to the ACC feature use versus L2 use. The result may also be confounded by the possibility that drivers may be more likely to engage L2 features when conditions are generally supportive of speeds higher than the posted limits (e.g., free-flowing controlled-access roads). Thus, it is unclear whether any driver adaptation was observed in these analyses. It is possible that more complete or robust behavioral adaptation to these technologies would take several months, rather than weeks, especially as L2 features may be used infrequently (i.e., on less frequent, longer trips, as opposed to much shorter daily errands). Still, the results are useful in providing insight into how older drivers use L2 systems during the first 3 hours of cumulative use.

REFERENCES

1. Wilde, G. J. (1982). The theory of risk homeostasis: implications for safety and health. *Risk analysis*, *2*(4), 209-225.Wilde, 1982
2. Wilde, G. J. (1982). Critical issues in risk homeostasis theory. *Risk analysis*, *2*(4), 249-258.
3. Kinnear, N. A., & Helman, S. (2013). Updating risk allostasis theory to better understand behavioural adaptation. *Behavioural adaptation and road safety*, 87-110.
4. Antin, J. F., Guo, F., Fang, Y., Dingus, T. A., Hankey, J. M., and Perez, M. A. (2017). The influence of functional health on seniors' driving risk. Journal of Transport & Health, 6, 237-44. https://doi.org/10.1016/j.jth.2017.07.003
5. Kaber, D., Zhang, Y., Jin, S., Mosaly, P., & Garner, M. (2012). Effects of hazard exposure and roadway complexity on young and older driver situation awareness and performance. *Transportation research part F: traffic psychology and behaviour*, *15*(5), 600-611.
6. Liang, D., Lau, N., Baker, S., and Antin, J. F. (2020). Examining senior drivers' attitudes towards advanced driver assistance systems after naturalistic exposure. *Innovation in Aging*, *4*(3), 1-12. https://doi.org/10.1093/geroni/igaa017
7. Liang, D., Antin, J. F., and Lau, N. (2019). Examining Senior Drivers' Acceptance to Advanced Driver Assistance Systems. Presentation delivered at The 5th International Symposium on Future Active Safety Technology toward Zero Accidents (FAST-zero-19), Blacksburg, Virginia, 9/9-11/2019.
8. Russell, S.M., Blanco, M., Atwood, J., Schaudt, W.A., Fitchett, V.L., & Tidwell, S. (2018) Naturalistic study of Level 2 driving automation functions (Report No. DOT HS 812 642). Washington, DC: National Highway Traffic Safety Administration.
9. Romoser, M. R. E. and Fisher, D. L. (2009). The effect of active versus passive training strategies on improving older drivers' scanning in intersections. *Human Factors*, *51*(5), 652-68.
10. Romoser, M. R. E., Pollatsek, A., Fisher, D. L., and Williams, C. C. (2013). Comparing the glance patterns of older versus younger experienced drivers: Scanning for hazards while approaching and entering the intersection. Transportation Research Part F: Traffic Psychology and Behaviour, 16, 104-16.

EXPLORING DRIVER ADAPTATION TO LOWER LEVELS OF AUTOMATION (L2) USING EXISTING NATURALISTIC DRIVING DATA

Sheila G. Klauer
Shu Han
Feng Guo
Virginia Tech Transportation Institute
USA

Thomas Fincannon
National Highway Traffic Safety Administration
USA

Paper Number 23-0322

ABSTRACT

This project evaluated driver adaptation in the hours, days, and months after the introduction of level 2 (L2) advanced driver assistance system features (i.e., the system controls lateral and longitudinal motion) into the driving task. Two existing naturalistic driving study databases were analyzed: the L2 Naturalistic Driving Study and the Virginia Connected Corridor Elite Naturalistic Driving Study. To best assess driver adaptation, the analysis identified three phases of exposure time to L2 features: Phase 1 (immediate, under 3 hours), Phase 2 (short term, 3 to 8 hours), and Phase 3 (long term, over 8 hours). The results suggested that driver adaptation was present for high-risk secondary tasks, as significant increases in engagement were observed over the three phases, but only when L2 features were active. Additionally, drivers set their vehicle speed above the speed limit more frequently between Phases 1 and 2, with higher speeds set when L2 features were active as opposed to when they were inactive. While these results may be concerning, research efforts at a larger scale are needed to determine if there is increased crash risk associated with speeding and high-risk secondary task engagement with L2 features active. We also need to better understand the impact of traffic/roadway conditions on speed selection with L2 systems.

INTRODUCTION

In the driving domain, the concept of behavioral adaptation refers to how humans respond, either intentionally or unintentionally, to the introduction of a new technology that serves a specific driver need [1]. More general theories of behavioral adaptation focus on risk-based measures, beginning with a study of changes in galvanic skin response during driving [2], which led to further investigations of behavioral adaptation and integration of the theories of risk compensation and risk homeostasis [3]. Risk compensation and risk homeostasis theories function under the principles of a perception-evaluation process: (1) people have an idea of the level of risk that they are willing to tolerate; (2) people also have a "target" level of risk at which they are most comfortable; and (3) people have a reasonably good ability to perceive their current level of risk. With the notable exception of the zero-risk theory [4], risk compensation theories neglect to explicitly consider learning and time. However, the zero-risk theory posits that drivers' target level of risk can change with learning over time, which is where driver adaptation may occur [5].

Risk allostasis theory builds on the zero-risk theory to incorporate driver perception and decision-making with the constant changes that occur in the environment (i.e., learning over time). Kinnear and Helman [5] utilized risk allostasis theory to evaluate and potentially predict behavioral adaptation for drivers using driver assistance technologies [6]. Their assessment incorporates the driver's feelings of risk, task difficulty, and workload. They maintain that with sufficient sensitivity to risk, task difficulty, and workload, risk allostasis theory predicts that any alteration of the driving task (i.e., introduction of advanced driver assistance systems [ADAS]) will result in driver adaptation that will trend toward maintaining task demand within a preferred range. In other words, as the automated driving features simplify the driving task, the driver will feel free to increase task demand in a variety of ways that could include increased speed, increased secondary task engagement, and decreased following distance. Thus, risk allostasis theory predicts that the use of driver assistance technologies would result in increased trust and reliance on these technologies. This claim was substantiated by Llaneras et al. [7], who showed that, when given the opportunity to relinquish control of lateral and longitudinal operations to a simple but reliable ADAS, most drivers will engage in moderate to complex secondary tasks and will also exhibit increased eyes-off-the-forward-road time.

While behavioral adaptation and user reliance can occur because of changes to any aspect of the roadway system, the present study is concerned specifically with how drivers initially adapt their behaviors to level 2 (L2) ADAS features, as defined by SAE International [6]. L2 system features assist the driver through a combination of simultaneous adaptive cruise control (ACC) and lane centering assistance (LCA) for longitudinal and lateral control of the vehicle, respectively, while driver constantly supports this support feature and maintains responsibility for driving. When operated without other traffic in the driver's travel lane, ACC maintains vehicle speed in a manner like that of conventional cruise control. However, if the driver approaches a slower moving or decelerating lead vehicle in their travel lane, ACC can attempt to reduce the speed of the driver's vehicle to that of the lead vehicle. In many driving situations, this results in the driver's vehicle following the lead vehicle at a prescribed following distance, or headway. Some ACC systems can also follow the lead vehicle to a complete stop; however, a constant headway operates within a speed range and is not typically maintained at very low speeds (i.e., approaching zero). Lateral features such as LC provide sustained lateral assistance using cameras to determine the location of lane lines on either side of the vehicle and can attempt to keep the driver's vehicle in the center of the travel lane. Although some systems require lane lines on both sides of the driver's vehicle to remain operational, some newer generation systems may be able to use the contrast between the road and an unpaved shoulder to define a lane boundary if a line marking is not apparent.

Both longitudinal and lateral control features are currently available on a wide variety of vehicles, and a growing number of these types of features will continue to enter the market in the near future. Figure 1 describes SAE L2 ADAS features.

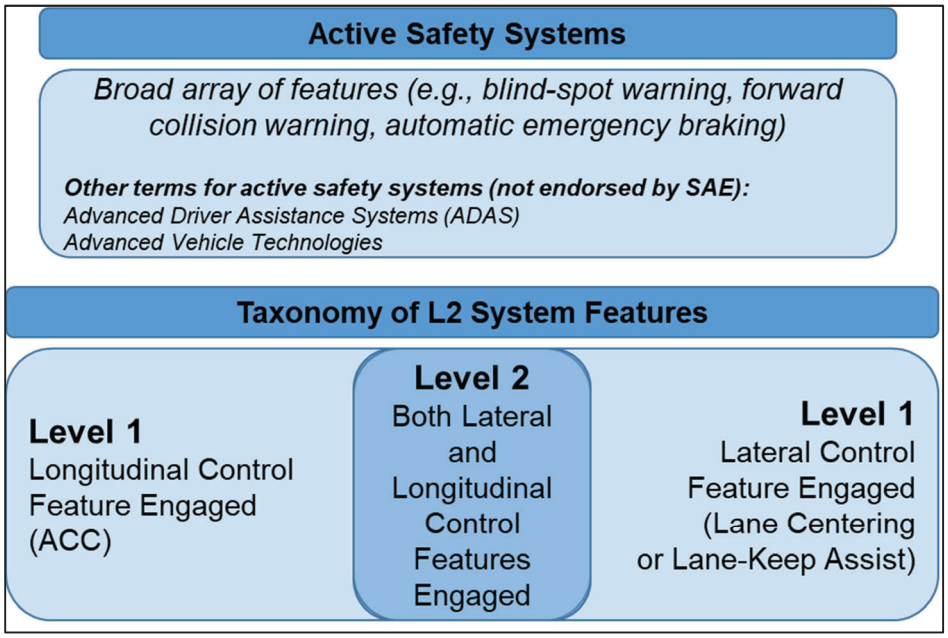

Figure 1. Definition of terminologies used to describe active safety and L2 driving system features.

Adaptation to Driving Automation Features

Although L2 system features are increasingly available in vehicles sold in the United States [8, 9], based on a 2016 survey [10], light vehicles driven in the United States were on average 12 years old. Therefore, it will likely take another decade before these systems reach substantial levels of market penetration in the United States. Given their growing availability, it is imperative that human factors researchers get ahead of this curve and develop a broad understanding of how drivers, over a range of ages and levels of driving experience, use and activate these L2 system features.

Learning theory suggests that improved performance comes with increased practice as a function of the power law of practice [11]. Figure 2 displays the pattern in which acquisition of a new skill occurs as the user gains experience. The greatest amount of new learning occurs during the early stages, when the user is initially gaining experience. Learning then follows an exponential curve, continuing but at a decreased rate [11].

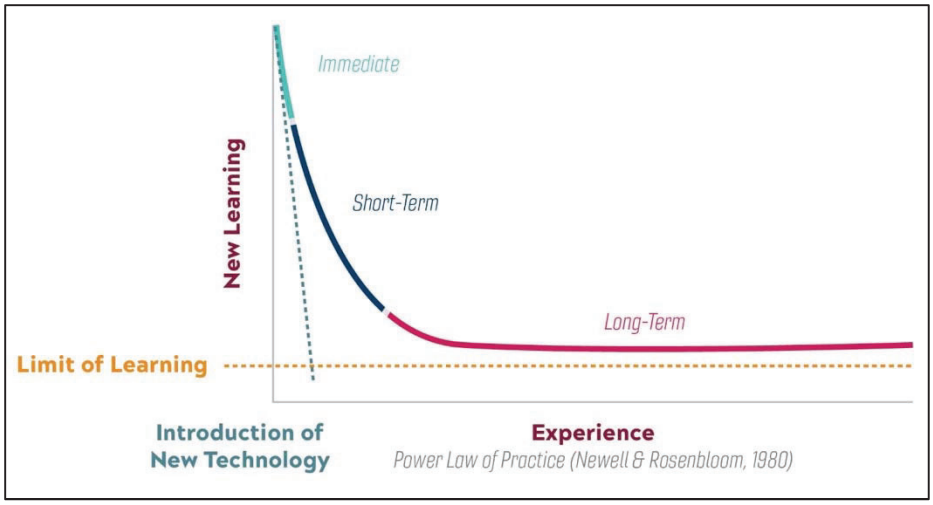

Figure 2. Depiction of immediate, short-term, and long-term phases of behavioral adaptation as plotted on a traditional learning curve (adapted from [11]).

By evaluating driver behavior for varying durations of experience with L2 system features, we can assess the moments when driver behavior may change most rapidly (immediate and short-term exposure) compared to when driver behavior may change much more slowly or when it is fairly stable after long-term use [12].

Considering this anticipated pattern of learning, we hypothesized that most behavioral adaptation to L2 system features would occur in the initial periods of use. Therefore, to gain insight into driver adaptation to L2 system features, it was necessary to observe and measure driver behavior while using these systems over time. The analysis defined three phases of exposure time to L2 features: Phase 1 (immediate, under 3 hours of L2 activation), Phase 2 (short term, 3 to 8 hours of L2 activation), and Phase 3 (long term, over 8 hours of L2 activation). There is an absence of NDS research using this approach with exposure to L2 features, so in order to select hours for the three phases, this study: (a) examined exposure data within the databases (see Method below) to observe how much drivers used L2 features and (b) sampled shorter time periods to test examine how quickly driver behavior might change.

METHOD

Overview of Naturalistic Driving Study Databases

This study evaluated driver adaptation to L2 systems using data from two naturalistic driving studies (NDSs): (1) the Naturalistic Study of L2 Driving Automation Functions (L2 NDS); and (2) the Virginia Connected Corridor 50 Elite Vehicle (VCC50 Elite) NDS.

The L2 NDS database [13] was used to assess driver adaptation to L2 system features over the course of each driver's initial 4 weeks of driving a vehicle with those features present. In the original study [13], 120 participants drove study-provided vehicles that were different from those they currently owned. Of the 120 participants, only 82 had sufficient driving time with L2 systems active (i.e., greater than 3 cumulative hours), so the analyses were based upon 82 participants. Observation from this dataset examined three phases of exposure time to L2 features: Phase 1 (immediate, under 3 hours of L2 activation), Phase 2 (short term, 3 to 8 hours of L2 activation), and Phase 3 (long term, over 8 hours of L2 activation).

The VCC50 Elite NDS database was used as a comparison group of experienced participants. In that study, 50 participants owned personal vehicles equipped with L2 system features and had driven them for several months to over a year. Of the 50 drivers in the VCC50 Elite dataset, only 33 drivers were included for sampling: drivers who owned vehicles equipped with both ACC and some form of LCA or lane keep assistance (LKA and for whom data to indicate L2 system state was available. Observations from this dataset corresponded to long-term exposure to L2 features.

L2 NDS and VCC50 Elite NDS Variables

To assess driver strategies and behaviors to maintain vehicle safety while using L2 system features, speed selection at moment of L2 feature engagement and driver engagement in secondary tasks were used for analysis. Speed

selection at moment of L2 feature engagement was determined using database coding algorithms. Every time the driver engaged L2 features, the speed at which the vehicle was traveling was coded and marked, as was the posted speed limit.

The primary independent variable focused on L2 activation status, where the driving automation (i.e., the systems were both available and active) or the driver (i.e., systems were available but inactive) controlled both lateral and longitudinal motions of the vehicle. The idea of available-but-inactive is important in ensuring comparisons are reasonable. If comparisons were made between L2 usage periods and all non-L2 usage periods, any observed differences could readily be attributed to the different conditions, scenarios, and driving environments in which drivers tend to - or are permitted to - engage L2 systems. Using available-but-inactive driving epochs to provide control samples makes usage/non-usage comparisons more meaningful.

To better assess the types of behaviors that drivers engage in when using L2 features, trained data coders reviewed randomly selected matched cases (when L2 features were active) and controls (when L2 features were available but inactive) to determine prevalence of secondary task engagement. Using available data, secondary task types were grouped into those types of tasks that are high risk (e.g., texting on cell phone) or low risk (e.g., adjusting radio). High-risk tasks were those tasks that were found to be associated with increased crash risk in an analysis using the Second Strategic Highway Research Program (SHRP 2) NDS database [14]. The low-risk tasks were not found to be associated with an increase in crash risk in the same analysis. Additionally, frame-by-frame eye glance locations were recorded by trained data coders, and duration of eyes-off-road time was also calculated when L2 features were active versus when they were available but inactive.

RESULTS

Speed-Selection Behavior

Both inexperienced drivers (L2 NDS) and experienced drivers (VCC50 Elite NDS) tended to select speeds above the speed limit more frequently when L2 systems were active than when L2 systems were available but inactive. The experienced drivers demonstrated more frequent selection of speeds above the speed limit, primarily for the categories of 10 to 20 mph over the speed limit and greater than 20 mph over the speed limit compared to the inexperienced drivers. On average, drivers tended to travel 4.56 mph higher for inexperienced L2 drivers [$F(1,42,405) = 2724.83$, $p < 0.001$] and 2.66 mph higher for experienced L2 drivers [$F(1,75,208) = 452.12$, $p<0.001$], when L2 systems were active than when L2 systems were available but inactive.

The speed distributions are plotted in Figure 3 and Figure 4, in which the *x*-axes indicate speed in relation to the speed limit. Thus, 0 in the center of the *x*-axis refers to the vehicle traveling at the same speed as the posted speed limit (no difference in GPS speed and the posted speed limit). The *y*-axis represents the proportion of activations that occurred within each speed bin. Data plotted to the left of 0 indicates that the vehicle speed was slower than the posted speed limit. Data plotted to the right of 0 indicates that the vehicle speed was faster than the posted speed limit.

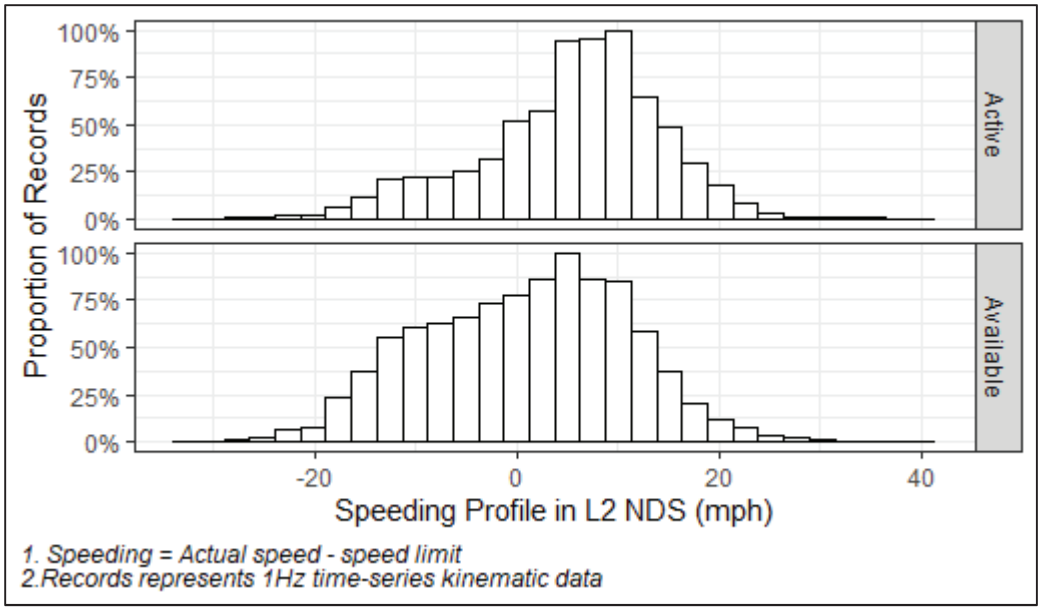

Figure 3. Inexperienced (L2 NDS of 82 drivers) driver speed selection profiles for when L2 systems were active (top) compared to when L2 systems were available but inactive (bottom).

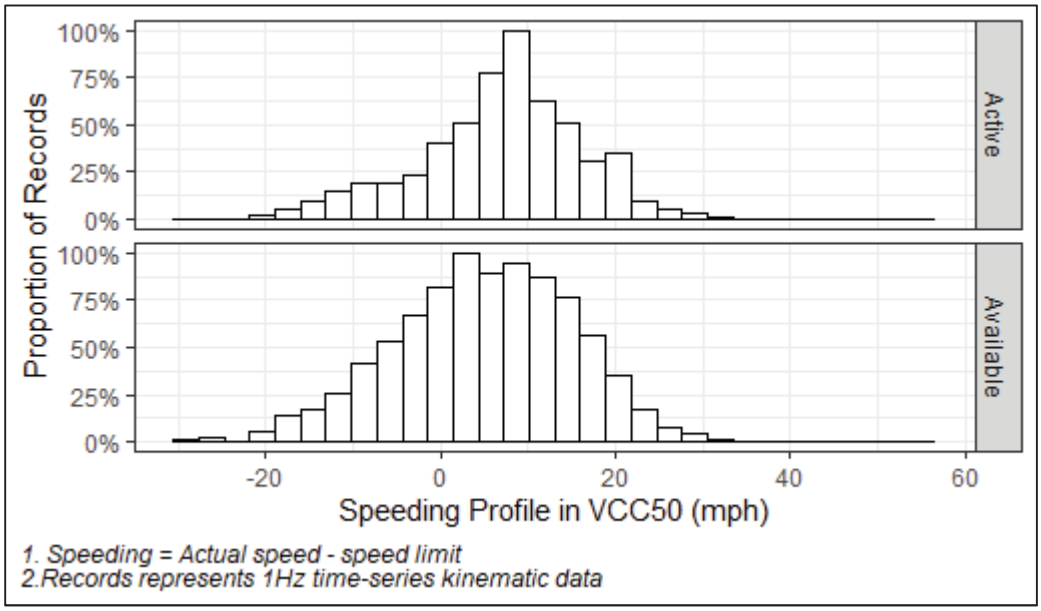

Figure 4. Experienced driver (VCC50 Elite of 33 drivers) speed selection profiles for when L2 systems were active (top) compared to when L2 systems were available but inactive (bottom).

Eye-Glance Behavior

Given that drivers were more likely to look off the forward roadway when L2 systems were active, the rest of the analyses will focus on those matched samples where drivers looked away from the forward roadway. An ANOVA was conducted to assess whether eyes-off-road glance metrics were significantly longer when L2 systems were active versus available but inactive and if eye-glance durations changed over time. Four glance metrics—the total eyes-off-road time, the mean duration of glances, the single longest glance, and the number of glances—were computed for eyes-off-road (Figure 5).

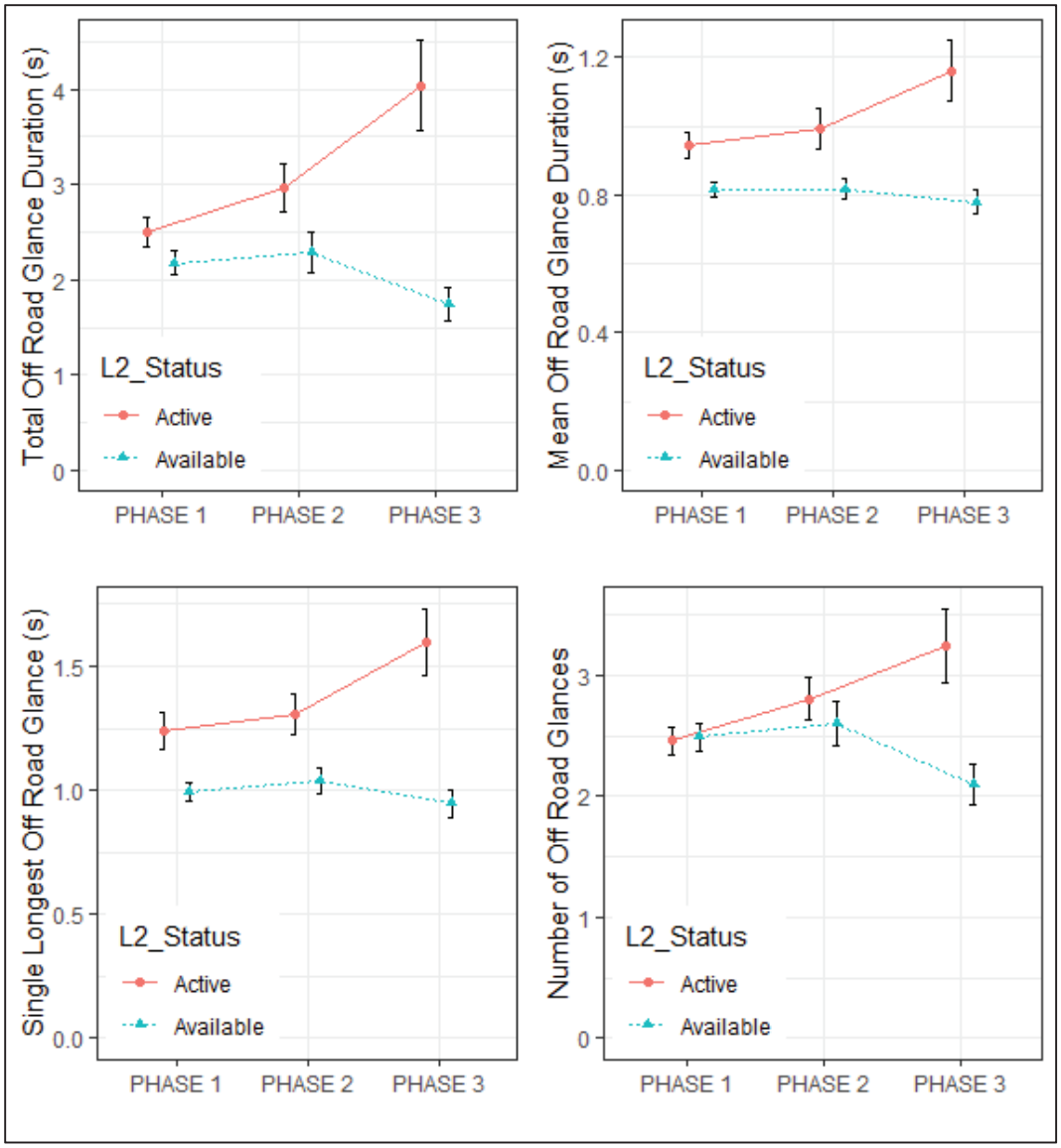

Figure 5. Total eyes-off-road time (top left), mean glance off-road duration (top right), single longest off-road glance (bottom left), and number of off-road glances (bottom right), for each exposure phase for the inexperienced drivers (L2 NDS).

Secondary Task Engagement

This study also examined the prevalence of secondary task engagement and eyes-off-road time. High-risk secondary task prevalence increased over time when L2 systems were active. High-risk secondary task prevalence decreased when L2 was available but not active. Analysis of the interaction between L2 system status and L2 exposure phase showed a statistically significant interaction (z value = -2.806, p = 0.005). As shown in Figure 6, high-risk secondary task prevalence increased over time when L2 systems were active. High-risk secondary task prevalence decreased when L2 was available but not active.

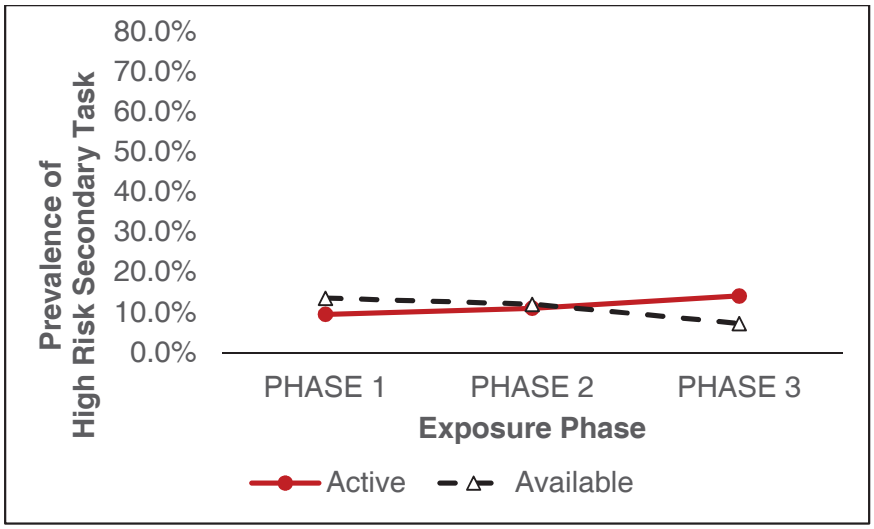

Figure 6. The interaction of L2 status by exposure phase for prevalence of engagement in high-risk secondary tasks.

CONCLUSIONS

Overall, the results from these analyses indicate that driver selection of higher speeds and high-risk secondary task engagement increased with the use of active L2 ADAS features. Eyes-off-road durations increased with use of L2 systems for both the L2 NDS drivers and VCC50 Elite NDS drivers. Regarding changes across three phases of exposure (i.e., less than 3 hours; 3 to 8 hours; 8+ hours), these findings illustrate how driver behavior changes when L2 ADAS features are used.

Regarding limitations, speed-selection is related to only one aspect of L2 control. Specifically, features are often available for independent use, such as in the form of ACC without LCA. The analyzed datasets did not have sufficient instances where L2 was available but only ACC was engaged to be included in speed-selection analysis. Therefore, it is unknown if or how much of the observed effects in speed-selection may be due to the ACC feature use versus L2 use.

This analysis also evaluated time of day, weekday versus weekend, and road type. This analysis was unable to identify a specific condition under which drivers were more likely to use L2 systems. A more nuanced, higher-level analysis involving additional data coding and/or algorithm development could identify effects of other factors.

REFERENCES

[1] Manser, M. P., Ward, N. J., Kuge, N., & Boer, E. (2005). Driver behavioral adaptation in response to a novel haptic driver support system. *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting, 49*(22). https://doi.org/10.1177/154193120504902214
[2] Taylor, D. H. (1964). Drivers' galvanic skin response and the risk of accident. *Ergonomics, 7*(4), 439-451.
[3] Kulmala, R., & Rämä, P. (2013). Definition of behavioural adaptation. In C. Rudin-Brown & S. Jamson (Eds.), *Behavioural adaptation and road safety* (pp. 18-22). CRC Press, Taylor and Francis Group.
[4] Summala, H. (1988). Risk control is not risk adjustment: The zero-risk theory of driver behaviour and its implications. *Ergonomics, 31*(4), 491-506. https://doi.org/10.1080/00140138808966694
[5] Kinnear, N.A.D., & Helman, S. (2013). Updating risk allostasis theory to better understand behavioral adaptation. In M. Rudin-Brown & S.L. Jamson (Eds.), *Behavioural adaptation and road safety* (pp. 87-110). CRC Press, Taylor and Francis Group.
[6] SAE International. (2018). J3016 – Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles.
[7] Llaneras, R. E., Salinger, J., & Green, C. A. (2013). Human factors issues associated with limited ability autonomous driving systems: Drivers' allocation of visual attention to the forward roadway. *Proceedings of the Seventh International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, 92–98. https://pdfs.semanticscholar.org/344d/7759e45e059b9b73a0de25e7d8b4af52f5bb.pdf

[8] American Automobile Association. (2019). *Advanced driver assistance technology names: AAA's recommendation for common naming of advanced safety systems*. https://www.aaa.com/AAA/common/AAR/files/ADAS-Technology-Names-Research-Report.pdf
[9] Consumer Reports. (2019). Cars with Advanced Safety Systems - Consumer Reports. Retrieved February 19, 2019, from Consumer Reports website: https://www.consumerreports.org/car-safety/cars-with-advanced-safety-systems/
[10] Bureau of Transportation Statistics. (2017). *Average age of automobiles and trucks in operation in the United States*. https://www.bts.gov/content/average-age-automobiles-and-trucks-operation-united-states
[11] Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Lawrence Erlbaum, Associates.
[12] Manser, M. P., Creaer, J., & Boyle, L. N. (2013). Behavioral adaptation: Methodological and measurement issues. In M. Rudin-Brown & S. L. Jamson (Eds.), *Behavioural adaptation and road safety* (pp. 339-358). CRC Press, Taylor and Francis Group.
[13] Russell, S. M., Blanco, M., Atwood, J., Schaudt, W. A., Fitchett, V. L., & Tidwell, S. (2018). *Naturalistic Study of Level 2 Driving Automation Functions*.
[14] Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(10), 2636–2641. https://doi.org/10.1073/pnas.1513271113

ROLE OF SYSTEM STATUS INFORMATION IN THE DEVELOPMENT OF TRUST AND MENTAL MODEL IN AUTOMATED DRIVING SYSTEMS

Michael Manser
Texas A&M Transportation Institute

John Campbell
Exponent

Thomas Fincannon
National Highway Traffic Safety Administration
United States

Audra Krake
Exponent

Liberty Hoekstra-Atwood
Exponent

Caroline Crump
Exponent

Lingtao Wu
Texas A&M Transportation Institute

Paper Number 23-0342

ABSTRACT

Research Question/Objective: In transportation, mental models are essential to mobility and safety because drivers rely on them to understand how to interact properly with their vehicles, the transportation infrastructure, and the environment. Poor performance and errors can occur when a driver acts in accordance with inaccurate mental models. Mismatches between mental models and actual experiences can also lead to reduced trust when, for example, the system with which they interact fails to perform to their expectations. The current study examined differing information types regarding Automated Driving System (ADS) capabilities and limitations on development of mental models and trust while using simulated Level 3 (L3) systems and a "dual model" use case of Level 4 (L4) systems (i.e., the vehicle can be both manually operated and can be controlled by ADS in certain ODDs).

Method and Data Sources: 48 females and males between the ages of 25 and 65 had four exposures to L3 and L4 systems in a driving simulator. Participants used either a basic human machine interface (HMI) that indicated the ADS was active, or they used an enhanced HMI that provided additional information indicating when the system was experiencing limitations (e.g., regarding detection of degraded lane lines). Participants used a simulated Level 3 system for two exposures and a simulated Level 4 system for two exposures. The acquisition and development of mental models and trust were assessed with standardized questionnaires.

Results: Regardless of exposure to each system over time, participants' mental models were more accurate for the simulated Level 4 system compared to the simulated Level 3 system and trust was greater for the simulated Level 4 system during the second exposure.

Discussion and Limitations: This paper summarizes research an ongoing project, and a final report will be published at a later date. Results of the current work suggests that the acquisition and development of mental models and trust can be differentially impacted by how well the ADS performs and the level of automation. However, because the study relied on simulated Level 3 and Level 4 systems, the results may not represent real world implementations of the technology.

INTRODUCTION

At their maturity, Automated Driving Systems (ADS) hold the potential to greatly decrease the number of crashes and save lives. However, there are many important and unanswered questions regarding Level 3 (L3) and Level 4 (L4) ADS [6], particularly around mental models. Mental models refer to a user's knowledge of an automated system's purpose, how it functions, and how it is likely to function in the future [1]. It is therefore important to consider the protentional relationship between a user's mental model of a system and safety. This may be particularly important for L3 vehicles that "cannot guarantee automated achievement of minimal risk condition in all cases within its ODD" and therefore, relies on a fallback-ready user [6]. As described in Campbell et. al. [2], users of L3 vehicles with a functionally accurate mental model are more likely to avoid errors based on incorrect assumptions about system operation and to use the automation appropriately. In contrast to L3, SAE discussion of L4 vehicles states that the system "must be capable of performing the DDT fallback and achieving a minimal risk condition," but also states that these systems "may allow a user to perform the DDT fallback, when circumstances allow this to be done safely" [6], so it may still be important to understand how a user's mental model and trust factors into operation of L4 vehicles. While some drivers may have existing mental models for common automation features such as cruise control, they will have vague or non-existent mental models for early implementations of L3 and L4 ADS [2].

The link between mental models and safety is mediated by trust. Specifically, a person's mental model of an ADS includes their understanding of what a system can and cannot do and it will influence their trust in what the system will do under specific conditions. Research in a variety of domains has identified that a functionally accurate understanding of automated systems is a central aspect to improving users' level of trust of the system [3, 4], where better understanding of the ADS should increase the likelihood that users will have the appropriate level of trust. Inaccurate mental models can lead to both under trust and over trust in an ADS. Research on trust in automation has shown that if a system is unreliable or causes a user to lose trust it will be underutilized and thus, not able to be effective [5]. While L3 and L4 ADS are not currently on the road, it is important to provide an early consideration of trust and its relationship with mental models.

Establishing appropriate levels of trust through functionally accurate mental models is a primary topic of concern in the development and deployment of ADS. When users do not fully understand the system, a mismatch between user expectations and vehicle actions may have a detrimental impact on trust, so it is important to consider how functionally accurate mental models are supported and shaped by an ADS's Human-Machine Interface (HMI). However, despite these established notions, there remain two critical areas in need of examination relative to the relationship between mental models and trust that can inform the HMI design of ADSs. The first area relates to the development of mental models and trust over time when using L3 and L4 ADSs. It is expected that people will begin using L3 and L4 ADSs with mental models and levels of trust based on prior knowledge, likely partly informed through media advertising, news reports, and discussions with peers. However, the continued development of mental models and trust would likely occur through direct interactions with ADSs. This study examines how HMIs can impact mental model and trust development over time.

The second area relates to the HMI implementation of L3 and L4 ADS features. Currently available implementations of HMI provide simple status information to users, typically using very simple binary information (e.g., telltales), such as system ready for activation yes/no and system activated yes/no. However, a relatively simple HMI for ADS may bely the complex nature of its operation and capabilities. It is easy to appreciate that in this case, mental models and subsequent trust in the ADS will be developed based on a limited perspective of the ADS which may not accurately reflect the true nature of the technology. A critical area to examine is how mental models and trust in ADS may be developed through the provision of richer system information that reflects a deeper understanding of the ADS. Specifically, will richer information result in improved/appropriate mental models and levels of trust or will this simply serve to overwhelm or distract users, thus having a negative impact on mental models and trust development. To address this issue, the second goal of the study was to explore the possible benefits of providing information to users about ADS limitations in addition to providing basic status information. In light of these two critical areas, the goals of the study were 1) to assess the impact of varied simulated HMIs on a user's ability to update mental models and appropriately trust L3 and L4 systems over time and 2) assess the possible benefits of providing information to users about the capabilities and limitations on an ADS in addition to providing basic status information.

METHOD

Participants

There were 48 participants in this study. 24 participants were females between 25 and 65 years of age (*M* = 47.8 years, *SD* = 12.6 years) and 24 were males between 25 and 65 years of age (*M* = 43.7 years, *SD* = 12.7 years). To control for potential differences due to demographics, for each HMI treatment level (i.e., Basic versus Information) our goals were to recruit participants so that the group means for age and years of driving experience for females and males would be approximately equal (Basic HMI: females *M* = 44.3 years, *SD* = 14.7 years, males *M* = 46.4 years, *SD* = 11.4 years; Informational HMI: females *M* = 51.3 years, *SD* = 9.4 years, males *M* = 40.9 years, *SD* = 13.9 years; Basic HMI: females mean driving experience of 27.5 years, *SD* = 15.8 years, males mean driving experience of 29.8 years, *SD* =12.2 years; Informational HMI: females mean driving experience of 35.3 years, *SD* = 9.44, males mean driving experience of 24.1 years, *SD* = 14.3 years). The actual mean differences in age and years of driving experience for females versus males were greater than we had hoped and could have affected the results. All participants possessed a valid United States (State of Texas) driver's license, self-reported normal (20/40) or corrected to normal visual acuity, and no color vision deficiencies which may have affected recognition of vehicle-based system icons or human-machine interface elements. To avoid possible experience bias, participants did not have any prior experience with L2 advanced driver assistance systems (ADAS) or L3 ADS technologies.

Apparatus

ADS The ADS was designed to be consistent with SAE J3016 representation of features and operational characteristics of specific implementations of highway automation systems that could be either L3 or L4 ADS [6]. Both the simulated L3 and L4 systems could operate on a four-lane highway with a median in good to moderate weather and could be activated when the vehicle was in "drive" and traveling at least 40 mph. Each simulated L3 and L4 system employed the functional equivalent of: (1) an adaptive cruise control system that would default to 45 mph or a 2s time-headway when a lead vehicle was present and (2) a lane centering system, while also performing the complete object and event detection and response (OEDR) [6]. It is important to note that this study used a short form term L4 to specially mean "dual-mode L4," where the vehicle could be both manually operated and controlled by ADS in certain ODDs. The systems were engaged by pressing a single button on the steering wheel while disengagement could occur when the same button was pressed, the brake or accelerator pedal was pressed, or the steering wheel was turned left or right more than five degrees. The visual HMI icon was positioned between the speedometer and tachometer and depicted a lead vehicle, lane lines, and the distance headway setting to a lead vehicle (see Figure 1).

Figure 1. Instrument panel with ADS information available in the center.

Within the Basic HMI condition, when the system became available, the system visual HMI icon appeared white to indicate the system was in standby mode and then, after activation, became green. The icon appeared continuously to inform users of the system status. Within the Informational HMI condition, the system presented the same information as the Basic HMI with the exception that the icon would change from green to yellow, flash at a 2 hz rate, and be accompanied by a two-beep tone when the roadway elements did not provide complete or clear information for the ADS to detect and use. This "limitation" message was triggered due to limitation scenarios (e.g., faded or degraded lane lines, a motorcycle as a lead vehicle, as examples) and the icon would remain yellow once the message was dismissed until the condition that triggered it ended. There was no response or action required

from the participants when the limitation message was presented; a key research question was whether or not such information affected their understanding (mental model) or trust in the simulated ADS. In all of the drive segments that included a limitation condition and message, vehicle behaviors in response to the limitation conditions varied depending on whether the L3 or L4 feature was active. Essentially, with the L3 feature active, the vehicle behavior was slightly and temporarily unstable; however, with the L4 feature active, there was no changes in vehicle behavior. For example, in the *Degraded Lane markings in right curve segment*, the L3 ADS exhibited lateral instability in the lane until the lane markings returned to normal, at which time the vehicle returned to stable and accurate lateral lane tracking. For the L4 ADS condition, the vehicle steered to stay in the curve and on the roadway throughout the duration of the limitation condition.

Driving Simulator and Driving World Data collection was conducted in the Texas A&M Transportation Institute's driving environment simulator (manufactured by Realtime Technologies, Inc.) which featured an original equipment manufacturers driver's seat, steering wheel, and accelerator and brake pedals. The visual display consisted of three high resolution monitors providing approximately 160º horizontal and 40º vertical fields of view. Road and ambient noises were provided through a multi-speaker audio system. The driving world simulated a typical highway environment that consisted of a four-lane divided highway with a grass median, Manual on Uniform Traffic Control Devices (MUTCD) compliant roadway markings, and 12 to 15 buildings placed to the right side of each road per mile. Vehicles traveled along each road to mimic light traffic conditions. The roadway environment and traffic were selected to mimic a real-life driving experience in rural areas in which vehicles with L3 or L4 highway automation features would be expected to operate. The driving world was approximately 11 miles in length and required approximately 15 minutes to drive at the posted speed limit of 45 mph.

Drives, Segments, and Scenarios Participants performed four counterbalanced drives (i.e., Drives A1, A2, B1, B2). Each drive consisted of eleven segments with each segment being approximately 1.1 mile long. The first segment, *Start*, allowed participants to accelerate to the posted speed limit and transition from manual to ADS control after being prompted. The final segment, *End*, allowed participants to transition from ADS control to manual driving to exit the highway as the vehicle left its operational design domain. There were four ADS "Normal Driving" segments that contained one scenario each and four ADS "Limitation Message" segments that contained one limitation scenario each. The penultimate segment contained no scenarios for L3 and for L4 drives with the exception that it contained a system automation failure scenario for Drive 2 of L3 only. The penultimate segment allowed for an examination of user responses to an L3 ADS failure. This approach was chosen to demonstrate higher functionality of L4 vehicles. The results of this examination will be presented in future publications. The segment/scenario order for each of the four drives is presented in Table 1, while general segment descriptions are provided in Table 2. It is noted the segments that provided participants with a limitation scenario and resulting limitation message are italicized and underlined, while all other segments were considered "normal driving" in which the ADS did not encounter any limitations within that scenario. The scenario descriptions in Table 2 also summarize the vehicle behaviors that distinguished L3 versus L4 functionality during a limitation condition.

Table 1: Order of Segments/Scenarios within each Drive.

A1 Segment Order	A2 Segment Order	B1 Segment Order	B2 Segment Order
1. Start	1. Start	1. Start	1. Start
2. LVLC	2. RC	2. LC	2. LVPV
3. DMLC	*3. GL*	*3. LVI*	*3. LVMC*
4. RC	4. LVPV	4. LVLC	4. LC
5. LVMC	*5. LVI*	*5. GL*	*5. DMRC*
6. LVPV	6. LC	6. RC	6. LVLC
7. GL	*7. DMRC*	*7. LVSV*	*7. LVI*
8. LC	8. LVLC	8. LVPV	8. RC
9. LVI	*9. LVSV*	*9. DMLC*	*9. GL*
10. ND or RORC	*10.* ND or RORC	*10.* ND or RORC	*10.* ND or RORC
11. End	11. End	11. End	11. End

Table 2: Segment Descriptions.

Abbreviation	Name	Description
	Start	Allowed participants to accelerate to the posted speed limit and transition from manual to ADS control.
ND	Normal Driving	Normal driving operation with no scenarios once ADS activated
DMLC	Degraded Lane Markings in Left Curve	A 90-degree sweeping left curve where the centerline and side lane markings were partially masked due to simulated dirt. In the Basic HMI condition, the HMI remained green with no audible warning. In the Informational HMI condition, an ADS limitation message was presented as the participant traveled next to the degraded markings. In both HMI conditions, L3 ADS exhibited lateral instability in the lane until the lane markings returned to normal. For the L4 ADS condition only, the vehicle was not unstable, and steered to stay in the curve and on the roadway.
DMRC	Degraded Lane Markings in Right Curve	Identical to DMLC with the exception that the curve swept right.
GL	"Ghost" lanes	A section of roadway where one set of faded lane lines appeared offset from brighter lane markings by approximately six to 12 inches to the right (e.g., new lane markings were applied while old lane markings are still visible). In the Basic HMI condition, the HMI remained green with no audible warning. In the Informational HMI condition, an ADS limitation message was presented when the participants' vehicle traveled next to degraded lane markings. In both HMI conditions, the L3 ADS exhibited lateral instability until the lane markings returned to normal, at which time the vehicle returned to stable and accurate lateral lane tracking. For the L4 ADS condition only, the vehicle followed the brighter set of lane lines and did not exhibit any lateral instability throughout.
LC	Left Curve	A 90-degree sweeping left curve where the centerline and side lane markings were fully visible.
RC	Right Curve	A 90-degree sweeping right curve where the centerline and side lane markings were fully visible.
LVLC	Lead Vehicle Lane Change	Participant traveled along road with surrounding traffic changing lanes ahead of the participant's vehicle and ADS engaged. No ADS limitations encountered.
LVMC	Lead vehicle: Motorcycle	Depicted a motorcycle in the left lane moving to the right lane in front of the participant's vehicle. The Basic HMI remained green with no audible warning. In the Informational HMI condition, an ADS limitation message was presented when the motorcycle's wheels crossed the center line. In both HMI conditions, the L3 ADS began to "tailgate" the motorcycle and continued to drive very closely to the motorcycle while remaining near, though below, the set maximum speed of 45 mph. For the L4 ADS condition, the participant's vehicle adjusted its speed to travel two car lengths behind the motorcycle.
LVSV	Lead vehicle: Small vehicle	Identical to LVMC with the exception that a small vehicle moved in front of the participants vehicle.
LVI	Lead Vehicle: Incursion	Scenario entailed a vehicle that pulled into a driver's travel lane and suddenly slowed. The Basic HMI remained green with no audible warning. In the Informational HMI only, an ADS limitation message was presented when the lead vehicle's passenger side wheels crossed the driving line. In both conditions, the L3 ADS decelerated the participant's vehicle in response to the slower lead vehicle. If a driver did not take over within five seconds, the participant's vehicle continued at 40 mph behind the lead vehicle. For the L4 ADS condition only, the vehicle adjusted speed to travel two car lengths behind the lead

		vehicle.
LVPV	Lead Vehicle: Passenger Vehicle	Participant traveled behind a lead vehicle with ADS engaged. No ADS limitations encountered.
RORC	Run Off Road in Curve	The RORC scenario used the DMLC and DMRC scenario. In the Basic HMI, the HMI remained green with no audible warning. In the Basic HMI, the HMI remained green with no audible warning. In the Informational HMI, an ADS takeover request was presented as the participant passed the first degraded markings and continued for 5 seconds. For the L3 ADS Drive 2 only, the vehicle continued to drive straight as the roadway curved until the driver took over. For all the L4 ADS condition exposures, the vehicle steered to stay in the curve and on the roadway.
	End	On-screen message indicates need to resume driving and bring the vehicle to a stop.

Sign Detection Task Participants were asked to engage in a sign detection task to assess attention allocation between the HMI and external roadway. Services road signs were placed on the right and left-hand sides of the roadway at approximately 30s intervals. Each sign presented logos for gas, food, and beverage, lodging, or attractions (see Figure 2) and were different across the four exposures. Participants were given a specific "target" logo to search for at the beginning of each exposure and they then indicated when they detected the target logo by pressing a button on the steering wheel that corresponded to the side of the roadway that the logo was seen (i.e., right or left). There were 24 signs per exposure with approximately one sign with the target logo and two signs without the target logo per segment. Results of the sign detection task will be reported in future publications.

Figure 2. Example of sign detection task road sign.

Questionnaires

A Mental Model and Trust Questionnaire (MMTQ) was developed to understand changes in different components of mental models and trust over time and in response to the two different HMIs. The mental model items included questions about system operation (e.g., how to turn system on and off), participants' understanding of the limitations of the automation features and operator commands (e.g., "what would be your first command in an automated vehicle, such as that used in this study, in the following situation") with multiple choice answers (e.g., brake, stop, change lanes, no response needed, etc.), and questions about participants' understanding of system behavior and operation (e.g., "how would you expect the vehicle used in the study to respond to the following driving situation") with multiple choice answers corresponding to appropriate vehicle behavior (e.g., with L1, L2, L3, or L4/L5 automation). To measure trust, participants were presented with a series of statements (e.g., "Highly automated vehicles can handle unexpected roadway situations" or "Highly automated vehicles are generally safer than human-operated vehicles") and asked to indicate their level of agreement with each statement using a 7-point Likert-scale.) All surveys were administered via SurveyMonkey.

Procedures

Human Subjects Consent, Instructions, Practice Participants read and then signed the human subjects consent form and answered preliminary questions. Participants then completed the MMTQ to collect baseline measures and a simulator sickness questionnaire to screen participants who had a greater propensity to get ill when exposed to the simulator scene. All participants receive the same training and instruction prior to entering the driving simulator. A training video reviewed the interface and operation of the simulated L3 and L4 ADS. Participants then experienced a practice exposure (no ADS and no sign detection task) to become familiar with the operational characteristics of the driving simulator, the task of driving, and to identify participants who exhibited signs of simulator sickness. Participants completed the sign detection task for three minutes while seated in the driving environment simulator (but not driving) to become familiar with the task, completed the second presentation of the MMTQ to assess changes in mental models and trust after instruction, and completed the first administration of the HMI questionnaire.

Drives and Debriefing Participants then completed Exposures 1 through 4 and took, after each drive, the MMTQ and then the HMI questionnaires. The MMTQ administrations after each experimental drive provided an indication of how mental models and trust further matured due to experience with the ADS. The HMI questionnaire administrations after each experimental drive provided the opportunity to assess each interface's usability and effectiveness on the general utility of the ADS information. The order of HMI condition levels (i.e., Basic, Informational) and automation levels (i.e., L3 ADS, L4 ADS) were counterbalanced across exposures (see Table 3). Participants completed an exit questionnaire and received a debrief of the study The study lasted approximately two hours for each participant.

Table 3: HMI and Exposure Counterbalancing by Participant.

Participant Numbers	HMI Group	Order of ADS Levels	Order of Driving Segments
1-3	1 (Basic)	L3, L4	A1, A2 with ROR, B1, B2
4-6	1 (Basic)	L3, L4	B1, B2 with ROR, A1, A2
7-9	1 (Basic)	L3, L4	A2, A1 with ROR, B2, B1
11-12	1 (Basic)	L3, L4	B2, B1 with ROR, A2, A1
13-15	1 (Basic)	L4, L3	A1, A2, B1, B2 with ROR
16-18	1 (Basic)	L4, L3	B1, B2, A1, A2 with ROR
19-21	1 (Basic)	L4, L3	A2, A1, B2, B1 with ROR
22-24	1 (Basic)	L4, L3	B2, B1, A2, A1 with ROR
25-27	2 (Informational)	L3, L4	A1, A2 with ROR, B1, B2
28-30	2 (Informational)	L3, L4	B1, B2 with ROR, A1, A2
31-33	2 (Informational)	L3, L4	A2, A1 with ROR, B2, B1
34-36	2 (Informational)	L3, L4	B2, B1 with ROR, A2, A1
37-39	2 (Informational)	L4, L3	A1, A2, B1, B2 with ROR
40-42	2 (Informational)	L4, L3	B1, B2, A1, A2 with ROR
43-45	2 (Informational)	L4, L3	A2, A1, B2, B1 with ROR
46-48	2 (Informational)	L4, L3	B2, B1, A2, A1 with ROR

Independent Variables

Two types of an HMI were tested that included a "Basic" HMI focusing on on/off status and an "Information" HMI that provided a limitation message, reflecting situations where the system was uncertain about some aspect of lateral or longitudinal control. Two levels of automation were tested, a simulated L3 system and a more capable and better performing simulated L4 system, referred hereafter as "ADS Level". The HMI message served as a limitation message (i.e., for L3 only, this corresponds to a request to intervene) for the conditional driving automation associated with L3 ADS (see page 31 of J3016 [6] for a more detailed description of the conditional nature of L3 ADS) and leaving the decision to respond up to the user. Under L4 operation, there was a limitation message that served as a notification only, not as a request to intervene or a need for fallback performance.

Statistical Approach
The experimenters scored answers to the target questions for each part of the MMTQ separately (i.e., questions about: System Knowledge, System Capabilities/Limitations, and perceived Automation level). The experimenters also created a composite mental models accuracy score (out of 100%) to evaluate overall effects. The composite mental model's accuracy score was derived from 11 questions: 3 system use questions, 2 vehicle behavior questions, and 6 operator command questions.

To test the differences between HMI evaluations and trust ratings (MMTQ) between conditions, the participant's mean ratings for each survey were used as a response variable. The HMI evaluation survey had 5 items and the trust survey had 13 items. In the initial models, these scores were treated as interval data, since the scales were made up of over four Likert-type items that are combined into a composite score [7]. It is important to acknowledge that the Likert scale results are limited: they do not allow for further inferences about the differences in the underlying characteristics reflected in these values (e.g., the meaning of a 0.37 difference in the Trust score).

The data were subjected to linear mixed models. The independent variables used in the models that included all survey administrations were HMI (Basic vs. Information) and Survey Administration (Baseline, Post-Instruction, After Exposure 1, After Exposure 2, After Exposure 3, After Exposure 4). The independent variables used in the models that only included post-exposure survey administrations were HMI, ADS Level (L3 vs. L4), and ADS Exposure (First vs. Second), that is whether the exposure was the first or second time the participant had encountered that ADS level. Participant was treated as a random factor in all models. Non-significant interactions were removed from the final models. The data were analyzed using linear mixed models, built using the lme4 package in R version 4.1.3 (2022-03-10). If the data in a particular model failed to meet all normality assumptions, an alternative model was used. Significant interaction effects were examined with planned post-hoc contrasts using R's multcomp package. Significance levels were set at $p < .05$ where statistical analyses were performed.

RESULTS

Stage 1 Analysis
The results are presented according to the two stages of model testing. The first stage entailed examining what response variables were best predicted by either HMI or Survey Administration (i.e., Baseline, Post-Instruction, After Exposure 1, After Exposure 2, After Exposure 3, After Exposure 4). Results are summarized in Table 4 and indicated that accuracy of mental models questions (i.e., percent questions correct) average trust rating (i.e., average "agreement" scores on the 7-point Likert scale) were each predicted by Survey Administration and that HMI was not a predictor of any response variables. The mental model accuracy analysis significant effect for Survey Administration (95% CI: 11.8, 22.7) indicated that participants scored 17.2% higher on post-training (M = 60.7, SD = 20.3) compared to baseline (M = 43.5, SD = 14.2). The mental model accuracy for system use significant effect for Survey Administration (95% CI: 27.4, 40.7) indicated that participants scored 34.0% higher on post-training (M = 64.9 , SD = 30.4) compared to baseline (M = 30.4, SD = 25.3). The trust analysis significant effect for Survey Administration indicated that trust ratings (-0.32; 95% CI: -0.47, -0.17) decreased between the first (baseline) administration (M = 3.73 , SD = 0.74) and last (exposure 4) administration (M = 3.41 , SD = 0.71).

Table 4: F-Statistics for Final Models Using All MMTQ Survey Administrations

Response Variables	HMI Condition F-value; *p*	Survey Administration F-value; *p*
Mental Model Accuracy	F(1, 46) = 0.007; $p = .93$	F(5, 235) = 21.4; $p < .0001$ (17.2; 95% CI = 11.8, 22.7)
Mental Model Accuracy: System Use	F(1, 46) = 0.07; $p = .79$	F(5, 235) = 46.3; $p < .0001$ (34.0; 95% CI: 27.4, 40.7)
Average Trust Rating	F(1, 46) = 0.10; $p = .75$	F(5, 235) = 6.42; $p < .0001$ (-0.32; 95% CI: -0.47, -0.17)

Stage 2 Analysis

The purpose of the second stage was to conduct follow-up linear mixed models contrast analyses to better identify what response variables are best predicted by HMI, ADS Level, and ADS Experience or an interaction between ADS Level and ADS Experience. Table 5 summarizes the results of the contrast analyses. Results indicated that mental model accuracy, mental model accuracy-operator commands, and average trust were each predicted by ADS Level. Results further indicated that each of the response variables were also predicted by the interaction between ADS Level and ADS Experience. Due to the accepted practice that interactions take priority over main effects, the remainder of this discussion will focus on the significant interactions.

Table 5: Results of the Final Models Using All Post-Exposure Survey Administrations.

Response Variable	**HMI Condition F-value; *p***	**ADS Level F-value; *p***	**ADS Experience F-value; *p***	**ADS Level by ADS Experience Interaction F-value; *p***
Mental Model Accuracy	F(1, 46) = 0.003; $p = .96$	F(1, 141) = 47.5; $p < .0001$	F(1, 141) = 0.03; $p = .86$	F(1, 141) = 8.83; $p < .01$
Mental Model Accuracy – Operator Command	F(1, 46) = 0.06 ; $p = .81$	F(1, 141) = 29.6 ; $p < .0001$	F(1, 141) = 0.01; $p = .94$	F(1, 141) = 8.93; $p < .01$
Average Trust	F(1, 46) = 0.002; $p = .97$	F(1, 141) = 23.2; $p < .0001$	F(1, 141) = 2.36; $p = .13$	F(1, 141) = 10.8; $p <.01$

The interaction between ADS Level and ADS Experience for mental model accuracy indicated that the average mental model accuracy scores for L3 and L4 diverged from the first to the second exposure (see Figure 3). There were no significant differences between mental model accuracy between the first and second L3 exposures and between the first and second L4 exposures. However, it is noted that participants exhibited more accurate mental models in the L4 exposures than in the L3 exposures for both the first (L4: M = 69.3, SD = 21.5; L3: M = 63.1, SD = 18.9) (6.25% higher for L4; 95% CI: 0.72, 11.8; $p < .05$) and second (L4: M = 73.8, SD = 20.2; L3: M = 58.0, SD = 19.4) (15.7% higher for L4; 95% CI: 10.2, 21.2; $p < .001$) survey administrations.

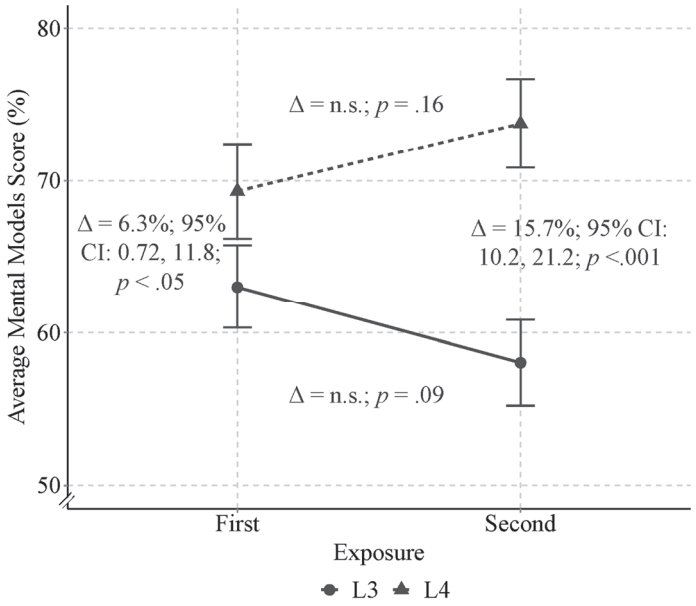

Figure 3. Depiction of the interaction between ADS Level and ADS Exposure on mental model accuracy. The error bars represent standard error.

Results indicated a significant interaction between ADS Level and ADS Exposure for the mental model accuracy –

operator command questions. The interaction was due to more accurate mental models for operator command in the second L4 exposure (M = 72.2, SD = 25.3) compared to the second L3 exposure (M = 53.1, SD = 25.6). Participants' accuracy was 19.1% greater (95% CI: 11.2, 27.0; $p < .001$) after the second exposure to L4 automation than after the second exposure to L3 automation (see Figure 4). There were no significant differences between the first and second L3 exposures, between the first and second L4 exposures, and between the first L4 exposure and first L3 exposure.

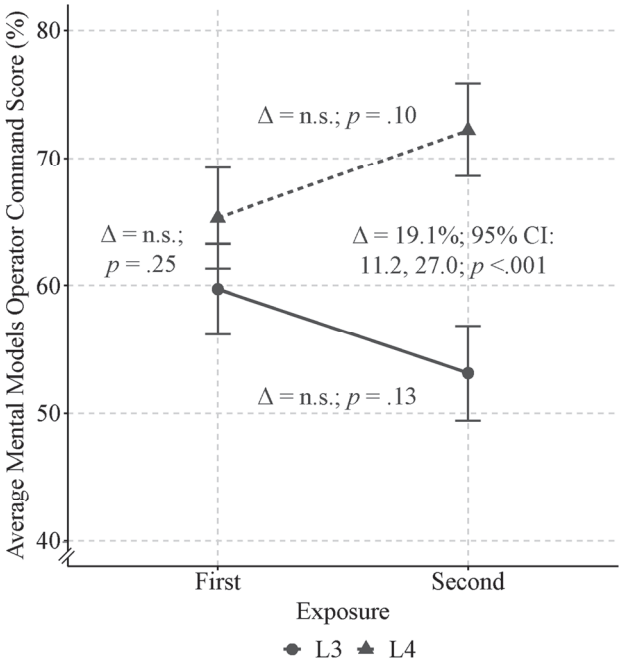

Figure 4. Depiction of the interaction between ADS Level and Survey Administration for mental model accuracy for the operator command questions. The error bars represent standard error.

The analysis indicated that average trust was best predicted by an interaction between ADS Level and ADS Exposure. Trust scores decreased significantly between the first and second L3 exposures (-0.22; 95% CI: -0.38, -0.06; $p < .01$) but not between the first and second L4 exposures. Trust scores were significantly higher after the second L4 exposure (M = 3.65, SD = 0.68) compared to the second L3 exposure (M = 3.28, SD = 0.74) (0.37; 95% CI: 0.21, 0.53; $p < .001$). There was no significant difference between the initial L3 and L4 exposure trust scores.

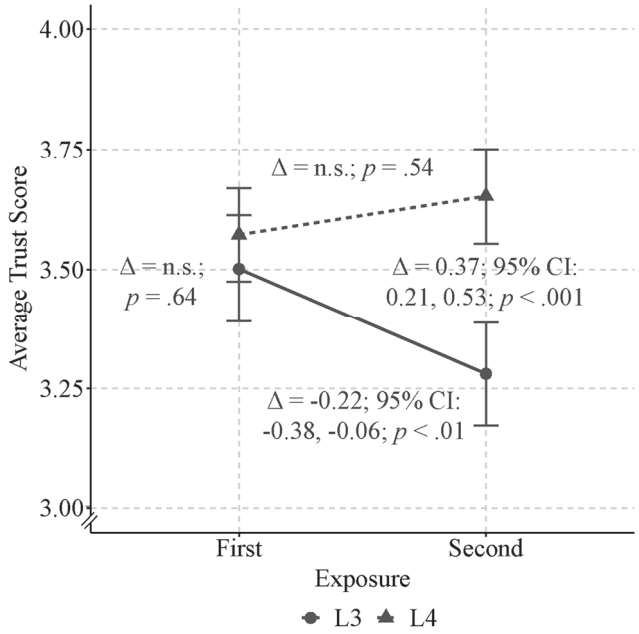

Figure 5. Depiction of the interaction between ADS Level and Survey Administration for trust. The error bars represent standard error.

DISCUSSION/CONCLUSIONS

Mental Models

Previous research indicated that an occupant's mental model of an ADS includes their understanding of what an ADS can and cannot do and will influence their trust in what the system will do under specific conditions. This notion places a critical emphasis on the importance of mental models because they can have a significant impact on not only occupant understanding of ADS but also on the establishment and management of trust over time. The current study sought to assess the impact of HMIs on an occupant's ability to update mental models for both L3 and L4 ADS over time and sought to assess the possible benefits of providing information about the limitations on an ADS in addition to providing basic status information on mental models.

Results of the current study indicated that, in general, mental model accuracy can be impacted by ADS competency and by continued ADS exposure. Specifically, mental model accuracy was generally higher for a better performing simulated L4 system compared to a more limited simulated L3 ADS and that there was a trend for mental model accuracy to increase with greater exposure to the L4 system and to decrease with greater exposure to the L3 system. This same pattern of findings was observed when mental model accuracy relative to operator commands was examined. This pattern of findings suggests that improved ADS understanding was associated with more exposure and with better stability and performance in the L4 compared to the L3 ADS.

It is noteworthy that overall mental model accuracy and mental model accuracy relative to operator commands were not predicted by HMI. Specifically, no differences were found regardless of whether the HMI provided basic "on/off" information or whether the HMI provided information relative to its uncertainty in detecting a potentially hazardous scenario. There were some advantages to the HMI providing limitation information relative at least one scenario (i.e., motorcycle scenario) but the findings were not pervasive across all scenarios. This finding may suggest that additional information about ADS operation may not positively or negatively impact mental model development.

Trust

In light of the notion that trust is mitigated by mental models, it was expected that if there were changes in mental models due to HMI, ADS Level, or ADS exposure, that there would also be changes in trust. Results of the current

work indicated that trust indeed changed based on ADS Level and ADS exposure. Specifically, ratings of trust at the first exposure to the simulated L3 or L4 were very similar, but trust in the L3 ADS decreased significantly by the second exposure. In contrast, although not significant, when users were exposure to L4 ADS their trust ratings increased from the first to the second exposure. Overall, the results suggest users' lower understanding of L3 ADS translated into lower levels of trust over time while the better-performing L4 ADS led to slight increases in trust over time.

Limitations

This paper presents an early overview of a study from an ongoing project. Due to the nature of conducting research with ADS vehicles that are not currently on the road, this research effort represents a projection of how these vehicles may operate, and these projections may differ from future implementations of ADS vehicles. For example, future implementations of L3 vehicles may have higher functionality than what appears in this study, and that may impact the development of trust and mental models. Therefore, there may be limitations regarding the degree to which these findings generalize to future research.

The study was presented with several situations that may have served as limitations. First, the study was conducted during the COVID-19 peak which may have impacted the type of people that were willing to participate. The particular concern is that participants willing to engage in socially interactive studies during a pandemic may have inherently different perceptions of trust than participants not willing to engage in studies. The potentially biased participant sample may not accurately reflect responses of the larger population. Second, this study used a group of participants that had limited exposure to ADAS. While this lower level of exposure allowed for the assessment of changes over time, future users may have more relevant exposure to capabilities that provide a better understanding of ADS-equipped vehicles. Third, there were no production-level vehicles available on the market for L3 or L4 features when this study was developed. While the descriptions, characteristics, and behaviors of the L3 and L4 features were consistent with specific implementations discussed in the SAE literature and with the discussions conducted by the research team with industry representatives in a different phase of the overall project, it is certainly the case that real-world implementations of L3 and L4 ADS-equipped vehicle may be different from those implemented or described in this research. For example, L3 vehicles may have better functionality that improves system performance, and these differences may impact the development of mental models and trust. Fourth, mental models and trust are inherently dynamic constructs that can remain stable or change over time and may do so due to a variety of factors. The relatively short study duration may not adequately capture changes in mental models and trust over time. The existence of these limitations suggests the need for future research and caution when extrapolating the results to real world applications.

REFERENCES

1. Goodrich, M. A., & Boer, E. R. (2003). Model-based human-centered task automation: A case study in ACC system design. *IEEE Transactions on Systems, Man, and Cybernetics, Part A, Systems and Humans, 33*(3), 325-336.

2. Campbell, J. L., Brown, J. L., Graving, J. S., Richard, C. M., Lichty, M. G., Bacon, L. P., Morgan, J.F., Li, H., Williams, D.N. and Sanquist, T. (2018). Human factors design guidance for level 2 and level 3 automated driving concepts (Report No. DOT HS 812 555). Washington, DC: National Highway Traffic Safety Administration. https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13494_812555_l2l3automationhfguidance.pdf

3. Lee, J.D., and See, K.A. (2004). Trust in automation; Designing for appropriate reliance. Human Factors, 46(1), pp.50-80.

4. Cassidy, A.M. (2009). Mental models, trust, and reliance: Exploring the effect of human perceptions on automation use (Master's Thesis). Naval Postgraduate School, Monterey, CA.

5. Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 30(3), 286-297.

6. SAE International, (2021). J3016-Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. Warrendale, PA; SAE International.

7. Boone, H. N. J., & Boone, D. A. (2012). Analyzing Likert Data. Journal of Extension, 50(2), 30.