FIGHTING DRIVER DISTRACTION – RECENT DEVELOPMENTS 2013 - 2015

Christian Heinrich
Daimler AG
Germany

Paper Number 15-0359

ABSTRACT

Research on driver distraction has a long history and attracts the attention of the scientific community, the public and the authorities. This resulted in a great number of activities. This article will summarize the main developments since the last ESV in 2013 from the perspective of an automotive manufacturer.

Guidelines and standards: In 2013 the National Highway Transport and Safety Agency (NHTSA) published a guideline for visual manual HMI . While basically following the structure of the European Statement of Principles (ESoP) [2] and the AAM guideline [1], the NHTSA guideline is much more restrictive, which means that more functions need to be blocked while driving. So the concern is that drivers will be inclined to use nomadic devices which have no restrictions at all (i.e. smartphones). Thus the overall impact of that guideline on safety may be negative.
The last version of the ESoP was published in 2008. In the framework of the iMobility Forum an HMI group was installed with the objective to check whether any changes, updates or additions are needed. The final report is expected for 2015. Some statements can be expected from the current draft which contains some recommendations and explicitly states to keep the ESoP as a design guideline, based on ISO standards, but not to include overall acceptance criteria. Due to the growing importance of applications that are being developed independently from hardware, another group was established within iMobility Forum, SafeAPP, in order to cover this specific topic.
On international level OICA (Organisation Internationale des Constructeurs d'Automobiles) published a white paper with recommendations for guidelines.

Naturalistic driving studies (NDS): Most experiments in simulators or on test tracks measure driving performance or glance behavior to determine mental workload. For using the telephone they usually report an odd ratio of four (The probability of a crash is four times higher compared to normal driving). It was quite a surprise when the 100 car study in 2013 presented odd ratios below one. Recent data from SHRP 2 give even lower values. NDS seem to be a powerful tool to identify actions and behavior that cause crashes. Detailed methods are under development and handling of great amounts of data is a challenging task.

Tethering: Nomadic devices can cause a problem since neither their displays nor their controls are developed for automotive use. Also they do not block functions that are not intended to be used while driving. One way to overcome this problem is to tether the telephone with the display and the controls of the car. This also makes it possible to apply existing guidelines (AAM, JAMA, ESoP). The Car Connectivity Consortium (CCC) has rephrased the existing guidelines so that they are better understandable by app developers. The CCC also established a process that will be run by certified labs to verify that the applications are in line with the guidelines.
Similar approaches are done be Google (Android auto) and Apple (Car play).

INTRODUCTION

Research on telematics applications started in the eighties. The experts realized already at that time the need for appropriate means to reduce driver distraction. Since then the awareness of the scientific community, the public and the policy maker has ever increased. Two facts are the main driving forces for recent developments. One is the

increased connectivity with smartphones whose computing power compares desktop PC. Drivers increasingly use these devices for some reasons: Some applications are relevant for driving like navigation or warning of speed cameras, some allow some work during commuting time and some allow the driver stay connected with friends and relatives also during the ride. Of course not all of these activities are compatible with driving. This raised great concern in the public and at authorities.

Another driving force are new research activities. In the past a number of measured variables has been used which are assumed to be an indicator of driving safety, like maintaining speed, keeping lane position, biometrical data etc. But the proof was still missing that these parameters describe driving safety, it was just a consensus in the scientific community. This changed with the naturalistic driving studies. This method allows observing the relation between driver behavior and crash risk in real traffic.

GUIDELINES

Based on the research done, guidelines have been developed in Europe, Japan and the US. A team of experts tasked by the European Commission developed the European Statement of Principles (ESoP) published in 2000 and revised in 2006. In Japan, the Japanese Automotive Manufacturers Association (JAMA) [4] published their guideline in 1990 with revisions in 2000 and 2004. In US the Alliance of Automotive Manufacturers (AAM) developed a guideline which was published in 2003 and revised in 2006. In 2013 the National Highway Traffic Safety Agency (NHTSA) in the US published a guideline. Since JAMA and AAM have not been updated in the last two years, the following focuses on ESoP and NHTSA Guideline.

NHTSA Guideline

In 2010 the NHTSA (National Highway Traffic Safety Administration) presented a project to fight driver distraction [5]. One objective was to develop guidelines for visual manual interactions. A draft for public discussion [6] was published in 2012. This draft based to great extent on the AAM Guideline, but introduced very detailed seven methods to assess driver workload [3].

EGDS	Eye glance testing
OCC	Occlusion testing
STEP	Step counting
DS-BM	Driving test protocol with benchmark
DS-FC	Driving test protocol with fixed acceptance criteria
DFD-BM	Dynamic following and detection protocol with benchmark
DFD-FC	Dynamic following and detection protocol with fixed acceptance criteria

After publication of the draft NHTSA faced strong opposition from the automotive industry. The main concerns were:

- NHTSA tightens the criteria very much without a basis of scientific data.
- If, as a consequence of these restrictions, functions of integrated devices are further restricted, users will be inclined to use handheld devices that do not have a user interface developed for use while driving and thus increase the amount of distraction and hence the probability of an accident.

Beside that there were a number of minor inconsistencies.

In 2013 the final version of the guideline was published. Some inconsistencies have been removed (e.g. 30 letter requirement which came from Japanese writing). The requirements for the test equipment give more

freedom now. However the basic restrictive approach persisted. The possible tests were limited to eye glance measurement and occlusion with slightly modified criteria.

By measuring eye glance time in a driving simulator each of the following criteria has to be met by at least 21 of 24 participants

1) Glance time (85. percentil): < 2sec
2) Mean glance: < 2.0 sec
3) Total glance time:< 12.0

For occlusion testing: 12 sec TSOT (Total shutter open time) for at least participants 21 of 24 has to be reached.

The consequences are demonstrated with the following example: Phoning will be impossible, because dialing will not pass the above criteria. For Total Off-Road Glance Time Mehler et al. reported mean values for selecting a contact from a directory between 12.12 and 16.86 seconds [F]. So this will never meet the above limit of 12 seconds for the 85 percentile.

This means that the above mentioned concerns still exist. If too many functions have to be blocked the driver may be inclined to use handheld devices that have no restrictions of complex functions and no displays and controls appropriate for automotive use. Up to now no automotive OEM is committed to the guideline.

European Statement of Principles

The last version of the ESoP was published in 2008. Due to the changes in HMI technologies and the rapid development of connectivity there was the need to review the guideline. In the framework of the iMobility Forum an HMI group was installed with the objective to check whether any changes, updates or additions are needed. This group started its work in August 2013 and the final report is expected for 2015. Some recommendations can be concluded from the current draft. Most parts of the ESoP have been confirmed after careful review. Proposed changes are:

- Revising the scope
- Taking into account the consequences of automated driving
- The grey area regarding time critical and not time critical warnings should be covered by examples

A special focus of the discussions was the question whether a pass/fail criterion for overall workload should be introduced. This was especially careful examined in the light of the recent NHTSA developments and the JAMA guideline. The conclusion of the group was, to keep the ESoP as a design guideline with reference to ISO standards.

The ESoP should be reviewed at least every three years and approaches in other countries should be monitored. Also there should be an ongoing process of dissemination and publicity since the ESoP addresses multiple stakeholders (e.g. service provider, radio stations) but only the OEM are aware of the ESoP.

Parallel to the HMI group another working group , Safe APP started a document that focusses on the adaption of the ESoP to application programs. It references the generic principles and the application design principles of the ESoP as far as they are relevant for application development. The paper also will deal with standards that work as an interface between different applications and applications and service provider.

OICA White Paper

End of 2013 OICA (Organisation Internationale des Constructeurs d'Automobiles) started a white paper with the title "Recommended OICA Worldwide Distraction Guideline Policy Position". The basic statements of this paper are:

- Vehicle manufacturers have long recognized the importance of supporting the driver's ability to maintain proper awareness of the driving situation.
- OICA members have worked to develop and adhere to regionally appropriate distraction guidelines for integrated systems.
- Since the introduction of these guidelines no significant increase of accidents caused by distraction from integtrated devices has been observed.
- The overly restrictive NHTSA guidelines for integrated vehicle systems are expected to push drivers toward the use of nomadic devices and thus reduce driving safety.
- OICA recommends not to develop additional guidelines, but that countries wishing to adopt distraction guidelines should follow one of the existing guidelines, namely, Japanese (JAMA)/ United States (Alliance)/European (ESoP) guidelines, in order to avoid unnecessary divergence among individual countries. The ESoP is a cultural independent approach.
- OICA supports efforts to develop and implement standards for automatic pairing of portable devices to vehicle integrated systems. This makes the use of in car input devices and displays possible and also allows to disable certain functions while driving.

NOMADIC DEVICES

Nomadic devices are personal electronic devices that can be brought into the car and have defined functionality including a user interface to control those functions. This definition includes smartphones, personal navigation devices (PND) or MP3 player. As pointed out in the introduction smart phones are the biggest concern because they are often used while driving in spite of the inappropriate user interface. Therefore this chapter focuses on smartphones. One possible way to reduce the distraction potential of smartphones is to connect them to the car: This has two advantages:

- The smartphone can use the large display and the automotive controls
- Certain functions can be disabled while driving

There are several realization of this concept by different organizations and companies.

OEM Integration

Nearly every car manufacturer has its own solution for integrating apps into the car e.g. Audi connect, BMW connected drive, Ford sync, Mercedes drive style, Volkswagen car net, Volvo Sensus Connected Touch. Besides the mirror technology (running the app on the smart phone while using input and output devices of the car) also other approaches are used. Some alternatives are: the car has its own communication device and does not need a phone. The software then can run on the head unit a server. In other cases the touchscreen of the smartphone is used and the HMI is adopted in a way that it fulfills the relevant guidelines even with a smaller display. From the view of an app developer this is a highly fragmented market place so most applications are developed by the car manufacturers themselves.

Apple CarPlay

Apple has defined a connection between the iPhone and the head unit of the cars with the brand name CarPlay. It uses different types of input devices like touchscreens, buttons or rotary knobs. A voice control (Siri) is implemented and can be activated by a button on the steering wheel. Of course all these features only work if the car makes these controls available to the iPhone. The applications within CarPlay are optimized for automotive use to reduce driver distraction.

Android Auto

Google has developed a similar interface for the connection of android phones to the car. They worked together with car manufacturers but also established their own HMI labs. Their guidelines regarding driver distraction followed the same general principles like ESoP, AAM and NHTSA guideline. Their interface to the app developer is template based. That means that templates are defined (e.g. selection from a list) that the app developer has to use. These templates can be tested regarding certain properties that have an influence on driver workload, especially font size and contrast. Together with rules about menue depth and list length the main factors regarding driver distraction are under control.

MirrorLink

The Car Connectivity Consortium (CCC) is a cross-industry collaboration to develop MirrorLink, an OS agnostic solution for smartphone and in-vehicle connectivity. The organization's more than 100 members represent more than 80 percent of the world's auto market, more than 70 percent of the global smartphone market and quite a number of aftermarket consumer electronics vendors.
The philosophy of MirrorLink was just to apply the existing guidelines (ESoP, JAMA and AAM) in the respective regions. But nevertheless it was necessary to compile new documents for the app developer for some reasons: The existing guidelines are written for experts of automotive HMI. For app developer with less experience of this topic some additional explanations are necessary. On the other hand the guidelines contain some parts (e.g. position of a display) that are important for automotive ergonomics but irrelevant for an app developer.
In addition to the guidelines CCC establishes two certification processes that cover both technical issues and driver distraction compliance. This means to authorize test labs and to select Subject Matter Experts. It is a challenging task to establish this process with many different stakeholders.

NATURALISTIC DRIVING STUDIES

During the last years there was significant progress in measurement technics. Storage media became smaller and cheaper. The same holds for cameras and sensors. Many physical data of the car are available via CAN (Car Area Network). This made a new type of experiment possible: Naturalistic Driving Studies (NDS). The basic idea is that the test persons drive a long time in their own cars which are equipped with cameras and sensors. The sensors typically include speed, acceleration yaw rate and distance to preceding cars. The cameras observe the traffic in all directions but also the face of the driver to determine glance behavior and the center stack to monitor the engagement in secondary tasks. While it has become easier to acquire all these data the processing is still a challenge especially since much of the work has to be done manually. If the objective is to analyze crash risk the statistic is still not very good since crashes and near crashes do not occur very often. So the results have to be taken with some caution. On the other hand the new method offers the opportunity to determine the relation between driver behavior including glance behavior and crash risk. All previous methods like determing driving quality or glance behavior just assumed based on the consensus of the scientific community that these are good indicators for crash risk.
Figure 1 shows the result for different secondary tasks. While previous measurements in simulators showed an Odd Ratio of around 4 for talking/listening on a phone we see here an Odd Ratio at or below 1. This means that the crash risk is lower than just driving. There are some hypothesis to explain that: 1) The influence of cognitive load on crash risk is overestimated 2) Glance behavior plays an important role and a person engaged in a phone call is less inclined to look around but always looks to the forward road. 3) The driver engages in a phone call only when the driving situation is less demanding. This means that the baseline risk is too high.

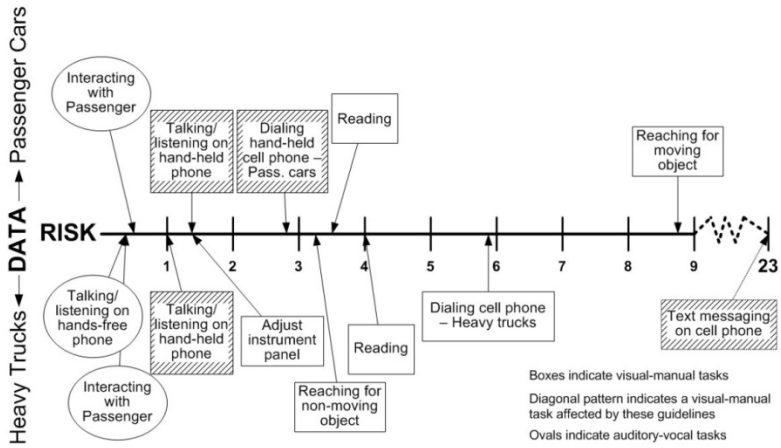

Figure 1 gives a graphical representation of some of the secondary task risk odds ratios determined by the 100-Car Naturalistic Driving Study and the Two Study FMCSA Analyses. [6]

Recently a evaluation of SHRP2 data was published [10]. This is a much bigger database than the 100 car study. It includes 3000 drivers and nearly 4000 vehicle years. Some results are shown in fig 2.

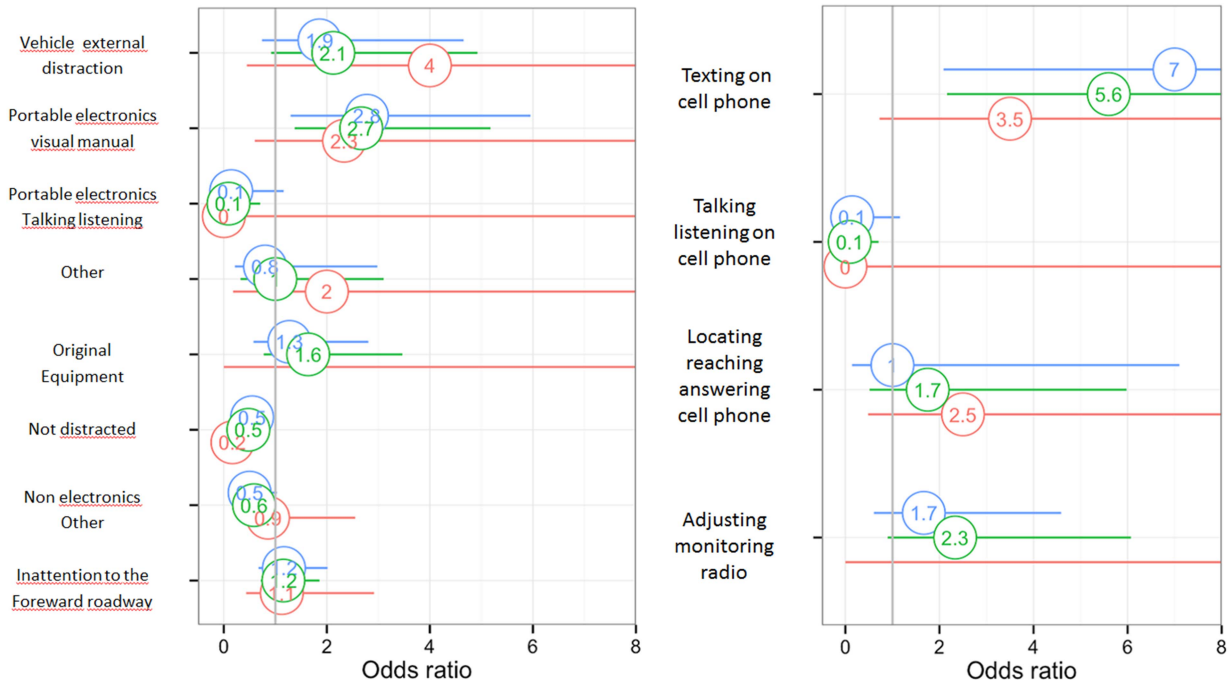

Figure 2. Odds Ratios (numbers inside circles) and confidence intervals (as horizontal lines) for classes of distracting activities in crashes (C) (red), near crashes (NC)(blue) and crashes and near crashes combined (CNC)(green). Presence of a distracting activity was coded between 5 seconds before and one second after the precipitating event.

Since not all data have been included in the evaluation statistic is still poor. Note that an odds ratio is significant only when the confidence intervals are fully above or below 1 (does not cross the vertical line at 1). Nevertheless the data give some indications what results can be expected in future evaluations with better statistics and some conclusions can be drawn for future standardization work.
Not surprisingly texting on cell phone and visual manual operation of portable electronic devices show high odd ratios. But also external distraction has comparably high values. It would be an interesting topic for future research

to have a more detailed description of these information and events. Outside information is never driver paced, i.e. the driver has only a short time slot to acquire that information If there is a chance to deliver part of that information inside the car, the driver can look at it when the driving situation is less demanding.
On the other hand it can be seen that talking and listening has only a very low odd ratio, even when it is done with a handheld device. The problems arise with locating, reaching and answering a cell phone.

VOICE INTERFACES

Since a long time there are also approaches develop guidelines for speech interfaces. Some give recommendations on dialog level. SAE J2988 [8] as an example claims to be "a technology-neutral approach to voice user interface principles and guidelines applicable in an automotive environment".

But there is also the objective to measure the effective workload of speech interfaces. Additional methods to measure cognitive load are under development. An example is the Driver Response Test, which is just standardized in the ISO committee (ISO DIS 17488).

But there is a specific problem to measure the overall quality and workload of a speech interface. This is the great influence of the quality of the voice recognition. If an utterance is misrecognized by the speech engine this will irritate the user and result in additional steps in the dialog. Whatever method is used to measure driver workload the results will become worse. On the other hand recognition quality cannot be considered a constant for a specific recognition engine. Some influences are quite obvious like ambient noise or reflection in the surrounding. Some are more difficult to determine. Details of the training material are generally unknown and so is the influence of the dialect coloration of the subject group. Also speech engines may have some adaption algorithms. The influence of those should be known to the experimenter. All effects of the speech recognition should be carefully monitored. The author is not aware of measurements in the automotive field that took care of this issue.

Another observation is the fact that speech control shows a much better glance behavior than manual operation [9]. From that can be concluded that guidelines with performance parameters are not appropriate for the time being. More research both regarding Naturalistic Driving Studies and detailed analysis of speech control is needed.

CONCLUSIONS

From the developments and research of the last two years the following conclusions can be drawn:

- The existing guidelines have proved to be efficient. For the time being the need for only slight corrections has been identified.
- Further improvement can only base on additional scientific findings. The Naturalistic Driving Studies seem to be a promising approach.
- Guidelines with too much restrictions will be contraproductive. The more the functionality of a well integrated OEM device is restricted the more the driver will be inclined to use a nomadic device. Thus reducing the functionality (i.e. making guidelines more stringent) will reduce the overall safety and should be avoided.
- Speech recognition seems to reduces the crash risk significantly. More and better speech systems should be offered to the driver
- To reduce the use of handheld devices the facility co pair the HMI of these devices to the car should be offered to a greater extend

ACKNOWLEDGEMENT

The author thanks all experts who work in the different organizations and committees in this field for their contribution to driving safety

REFERENCES

[1] Alliance of Automotive Manufacturers, "Statement of Principles, Criteria and Verification Procedures on Driver Interactions with Advanced In-Vehicle Information and Communication Systems" accessed February 9, 2012 http://www.autoalliance.org/files/DriverFocus.pdf

[2] European Commission "European Statement of Principles on the Design of Human Machine Interaction", 2006, accessed February 9, 2012. http://eur-lex.europa.eu/ Lex UriServ/ site/en/oj/ 2007/l_032/l_03220070206 en02000241.pdf

[3] Heinrich, C. "Fighting driver distraction: worldwide approaches", Proceedings of the 23rd International Technical Conference on the Enhanced Safety of Vehicles (ESV)

[4] Japanese Automotive Manufacturers Association , 2004 "JAMA Guideline" http://www.umich.edu/~driving/documents/ JAMA_guidelines_v30.pdf

[5] National Highway Traffic Safety Administration, "Driver Distraction Program", DOT HS 811 229, April 2012

[6] National Highway Traffic Safety Administration, "Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices" Docket No. NHTSA-2010-0053, Draft

[7] National Highway Traffic Safety Administration, "Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices" Docket No. NHTSA-2010-0053, Guideline

[8] SAE J2988 Society of Automotive Engineers, to be published

[9] Mehler, B, D. Kidd, B. Reimer, I. Reagan, J. Dobres and A. McCartt „Multi-Modal Assessment of On-Road Demand of Voice and Manual Phone Calling and Voice Navigation Entry across Two Embedded Vehicle Systems" Embedded Vehicle Voice Systems (2014-12-20 submission)

[10] Trent Victor, J. Bärgman, C.-N. Boda, M. Dozza J. Engström, C. Flannagan, J. D. Lee, G. Markkula. "Analysis of Naturalistic Driving Study Data: Safer Glances, Driver Inattention, and Crash Risk"

REPEATABILITY OF THE NHTSA VISUAL-MANUAL GUIDELINES ASSESSMENT PROCEDURE

Annie Rydström
Mikael Ljung Aust
Robert Broström
Trent Victor
Volvo Car Corporation
Sweden

Paper Number 15-0240

ABSTRACT

The Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices includes the Eye Glance Measurement Using Driving Simulator Testing (EGDS) task acceptance test. This paper describes the outcome of two EGDS tests, where both tests assessed the same set of 10 in-vehicle tasks, though each test employed a different group of 24 test participants, randomly sampled according to the proscribed EGDS procedure. Thus, in total 48 test participants drove on a simulated motorway in a lead vehicle following scenario while performing tasks such as changing temperature, setting destination in the navigation system and selecting and playing music using a prototype in-vehicle infotainment system located on the center panel. When comparing the test outcomes between the two groups, it was found that for 6 of the 10 tested tasks, pass/fail outcomes differed between the two groups on one or more of the proposed criteria. This high level of inconsistency in outcome between two identical tests using ten identical tasks raises questions regarding the repeatability of the proposed NHTSA EGDS procedure.

INTRODUCTION

Human-Machine Interface (HMI) guidelines and standards to promote safe design are important tools during the design of in-car interfaces. Several design guidelines and standards have been published, including the European Statement of Principles [1] the JAMA guidelines [2] and the Alliance guidelines [3] (see [4] for an extensive overview of guidelines and standards to use to enhance safety and ease of use, and reduce distraction). Lately authorities have expressed increased concern about a potential intensification in hazardous driver behavior following the functionality growth in infotainment systems in cars and smartphones. As a part of the Driver Distraction Program the National Highway Traffic Safety Administration (NHTSA) in 2012 issued additional guidelines, the Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices. The guidelines provide interface recommendations and performance-based acceptance criteria. After some criticism from both the car industry and the academia on first and foremost the task acceptance testing described in the guidelines, the final version of the guidelines was published April 26 2013 [5] (and some additional clarifications has been published after that [6]). The guidelines are valid for all new cars sold in the USA 3 years after the date of publication.

The guidelines include the task acceptance testing procedure Eye Glance Measurement Using Driving Simulator Testing (EGDS). Volvo Car Corporation (VCC) has for many years used test methods similar to EGDS to evaluate in-car HMIs during their development phase, and has in previous years performed numerous EGDS tests. However, when analyzing the test data according to the EGDS acceptance criteria, it has been found that a task which passes the test at one test occasion does not necessarily pass at another occasion.

A key property of any performance assessment procedure is reliability, i.e. for outcomes to be consistently similar when a test is repeated. If results are not repeatable, the whole idea behind the test fails. The lack of test repeatability when applying the NHTSA EGDS criteria at VCC has therefore raised concerns. However, since the testing performed at VCC has not strictly followed the prescribed EGDS procedure, it is possible that the lack of repeatability may be due to differences in test setups rather than an inherent property of the criteria formulation.

The objective of this study was to assess whether the lack of repeatable outcomes previously found at VCC is methodological or structural in character. To this end, two identical EGDS task acceptance tests were performed at a professional test facility external to VCC. The same 10 in-vehicle tasks were used for both tests, and the NHTSA

proposed EGDS test protocol was strictly followed in terms of recruiting and test setup, in order to have exactly the same trial conditions for the two test groups.

METHOD

The test was performed in the test facilities of Ergoneers GmbH in Manching, Germany.

Participants

A total of 48 participants, i.e. two test samples, were recruited in the area around Manching in Germany. Each test sample conformed to the test participant recommendations in the NHTSA guidelines in terms of general criteria, test participant impartiality, mix of ages (six participants in each of the age groups 18-24, 25-39, 40-54, and 55+) and an equal balance of men and women [5].

Equipment

The test was conducted in an Ergoneers Sim-Lab left-hand drive, fixed base and open cab driving simulator (Figure 1). The simulator is designed as a replica of an actual production vehicle and is adapted to the driving simulator recommendations in the NHTSA guidelines [5].

Figure 1: The Ergoneers Sim-Lab driving simulator

The driving simulation used in the study was a car following scenario on an undivided, four lane wide highway with a speed limit of 50 mph, designed in accordance with the driving simulator scenario recommendations [4]. The Ergoneers Dikablis head mounted eye tracking system was used to record the eye movements of the participants. A touch-screen tablet with a prototype infotainment HMI implemented on it was mounted at the center panel of the simulator. Tasks 1-9 were performed on this tablet. An additional, conventional Panasonic car stereo was installed in the simulator to be able to test the AAM radio manual tuning reference task [3].

Tasks

The prototype infotainment system was designed to replicate a typical in-car center stack interface. Nine tasks were included for the in-vehicle interface, and an additional radio manual tuning reference task was added for the conventional car stereo (Table 1). I line with the guideline recommendations, all tasks started from a home screen view and ended when the task had been executed correctly [5].

Table1: Tasks

Task no.	Task	Task description
1	Activate vehicle function	Access menu and activate a specific vehicle function (on/off option)
2	Activate USB	Access menu and activate USB
3	Set a new destination	Access navigation menu and activate a search field and tap in a destination of 10 characters and start guidance
4	Seta a favorite destination	Access navigation menu and a favorite menu, choose a specific destination and start guidance
5	Select and play music from USB	Access menu and activate USB and find a specified song in a list
6	Call contact from recent calls list	Access phone menu and recent calls list and choose a specified contact in the list to call
7	Dial a phone number manually	Access phone menu and keypad and tap in a 10 digit phone number and call
8	Change temperature	Access climate and increase temperature with 2 degrees
9	Set seat heat	Access seat heat and set it to a mid-level
10 – AAM reference task	Manual radio tuning	Access FM1 and adjust to 106.7 (starting position FM3, presetting 97.4)

Procedure

The test procedure, including test participant training, was arranged in alignment with the recommendations in the guidelines [5]. Upon arriving to the test facilities, the participants were given a brief description on the test procedure and they were asked to fill in a consent form and a demographic questionnaire. Then they were asked to sit down in the simulator and adjust the seat. They got to put on the eye tracking glasses and the eye tracking system was calibrated. The user interface and the tasks were demonstrated to the participants and they were able to practice each task as many times as they wanted to. The participants were then given instructions on the driving scenario and got to practice driving the simulator. When they felt comfortable driving the simulator they got to practice to perform the tasks as many times as needed while driving. After the training session the data trial was performed, i.e. the participants performed the tasks and data was collected.

Acceptance criteria

Eye movements were analyzed in accordance with the EGDS acceptance criteria [5, 6], which proscribes that no more than 3 out of 24 of the participants shall fail to meet each of the following criteria:

- **Acceptance criteria 1 (AC1):** For at least 21 of the 24 test participants, no more than 15 percent (rounded up to the next whole number) of each participant's total number of eye glances away from the forward road scene have durations of greater than 2.0 seconds while performing the testable task one time.
- **Acceptance criteria 2 (AC2):** For at least 21 of the 24 test participants, the mean duration of each participant's eye glances away from the forward road scene is less than or equal to 2.0 seconds while performing the testable task one time.
- **Acceptance criteria 3 (AC3**): For at least 21 of the 24 test participants, the sum of the durations of each individual participant's eye glances away from the forward road scene is less than or equal to 12.0 seconds while performing the testable task one time.

RESULTS

When comparing the acceptance test outcomes between the two groups of test participants, results show that for the 10 tested tasks, outcomes were the same in four instances and different in six instances.

Tasks 1, 2, 3 and 9 had the same outcome in both test groups. Tasks 1, 2 and 9 passed on all criteria, while task 3 failed on all criteria (Table 2). The other tasks had different outcomes. Task 8 failed on AC1 and AC2 with group 1, but passed on AC1 and AC2 with group 2. Tasks 4 and 6 passed on AC1 but failed on AC2 with group 1, while the inverse was true for group 2, i.e. failed on AC1 but passed on AC2. Tasks 5 and 7 failed on AC1 and AC2 with

group 1 but passed on AC2 with group 2. Task 10 failed on AC1 with group 1 but passed on AC1 with group 2 and failed on AC3 with both groups.

Table 2: The number of participants that failed to meet criteria, green = pass and red = fail

Task no.	Task	Group 1				Group 2		
		AC1	AC2	AC3		AC1	AC2	AC3
1	Activate vehicle function	2	2	0		3	3	0
2	Activate USB	3	3	0		1	1	0
3	Set a new destination	8	6	9		10	4	13
4	Seta a favorite destination	2	6	0		4	2	0
5	Select and play music from USB	8	6	3		6	3	1
6	Call contact from recents list	3	4	0		4	3	0
7	Dial a phone number manually	11	9	3		10	3	3
8	Change temperature	6	6	0		2	2	0
9	Set seat heat	1	1	0		0	0	0
10	AAM ref. task: Manual radio tuning	5	2	12		3	1	16

DISCUSSION AND CONCLUSIONS

The objective of this study was to assess whether the lack of repeatable outcomes previously found at VCC when applying the NHTSA acceptance criteria to EGDS test type methods was primarily methodological or structural in character. The results of the two identical EGDS acceptance tests reported above indicate that the lack of repeatability stems from how the criteria are designed, rather than how the test is performed. In this case, the two tests were set up following the proscribed protocol as strictly as possible, and the tested tasks were the same between the groups. Yet 6 out of 10 tasks came out different in terms of pass or fail on at least one of the proposed criteria.

From the above table, it mainly seems like acceptance criteria 1 and 2 have the most inconsistent outcomes, while acceptance criteria 3 (total eyes-off-road time) seems quite stable. Notable is the inconsistent outcome on task 10, which is a radio task that is supposed to work as a reference task, i.e. a task that has been prevalent throughout traffic through the ages and which drivers are thought to accept the risk of performing.

It is of course possible that some of the inconsistencies may still be due to methodological issues. In a perfect world, if two samples are recruited from the same population, and the same task is tested using the same test leaders, test method, equipment and logging apparatus, the outcome of a test should be consistent. In the real world on the other hand, differences may still occur due to different test leaders sometimes giving instructions in slightly different manners, and the environmental conditions and other factors may vary slightly. However, in this study, since the outmost care was taken to achieve an identical test setup for the two groups, it is unlikely that any unintended difference in the methodology would by itself lead to such a differentiated outcome. Moreover, since it would be a very large effort to reach a higher level of between test consistencies than the one in the current study, going for higher methodological precision does not seem to be an option. For an acceptance test to work in practice, it has to be robust enough to be applied at any certified test facility, without leading to inconsistent test outcomes.

Given that this study did have a high level of between test consistency, it thus seems like the problem with inconsistent outcomes have to do with how the acceptance criteria are formulated rather than with inconsistencies in

the EGDS testing. It is thus recommendable to investigate what it is in the criteria formulation that leads to these inconsistent outcomes, before applying the NHTSA EGDS method for assessment of production vehicles.

The results have some additional interesting properties. For one, the only tasks that passed all three criteria were the really simple ones that require only one or two button presses, such as activate a vehicle function, activate USB and set the seat heat. While that these simple tasks would pass was expected, it was not expected that a fourth task that presumably should belong to the same category of simple tasks, i.e. change the temperature in the vehicle, would fail (though it only failed in group 1, it passed in group 2). Looking at the detailed glace data from the two groups above, it can be seen that many drivers have long mean glance durations for these simple tasks. This suggests that drivers strive to complete these tasks in one go, rather than use multiple off-road glances.

This further highlights the need to understand what it is in the acceptance criteria formulation that leads to inconsistent outcomes. If future studies verify that mean glance durations indeed are longer for short and simple tasks, then this illustrates a deliberate driver strategy that has to be accommodated both in task development and testing. NHTSA has partially taken this into account by allowing a limited number of glances over 2.0 seconds for a task [6].

However, the results from this study shows that this topic still needs further investigation. For example, as the criteria are currently formulated, a person who uses a single glance of 3 seconds to complete a task will fail on both AC1 and AC2, but if that person uses a single glance of 3 seconds and then adds a check glance of 0.5 seconds (e.g. to verify that a setting has been updated), that person will pass both on AC1 and AC2. In the raw data from the two studies above, there are many examples of drivers who consistently fall into either of these two categories (i.e. they either use one long glance, or one long glance combined with a check glance). The behavior thus clearly is there empirically speaking, and therefore has to be considered in the acceptance criteria formulation.

In summary then, the difference in outcome between two identical tests using ten identical tasks raises questions regarding the repeatability of assessment outcome of the proposed NHTSA EGDS procedure and criteria. Since the need for an assessment method of this type is clear, and projected to lead to increased safety on the roads, further research into what exactly would need reformulation in criteria and method for the assessment procedure to fulfill its intended purpose is urgently needed.

REFERENCES

[1] European Commission. 2008. “Commission Recommendation of 26 May 2008 on Safe and Efficient In-Vehicle Information and Communication Systems; Update of the European Statement of Principles on Human-Machine Interfaces.”

[2] JAMA. 2004. “Guideline for In-Vehicle Display Systems, Version 3.0”. Japan Automobile Manufacturers Association, Tokyo, Japan, August 2004.

[3] Alliance of Automobile Manufacturers. 2006. “Statement of Principles, Criteria and Verification Procedures on Driver-Interactions with Advanced In-Vehicle Information and Communication Systems”, Driver Focus-Telematics Working Group, Washington, DC., June 26, 2006 update.

[4] Green, P. 2009. “Driver Interface Safety and Usability Standards: An Overview”. In K. Young, J. D. Lee and M. A. Regan (Eds.) Driver Distraction: Theory, Effects, and Mitigation. Boca Raton, FL: CRC Press.

[5] Visual-manual NHTSA Driver Distraction Guidelines for in-Vehicle Electronic Devices (2013). Docket no. NHTSA-2010-0053, National Highway Traffic Safety Administration, April 2013.

[6] Visual-manual NHTSA Driver Distraction Guidelines for in-Vehicle Electronic Devices (2014). Docket no. NHTSA-2014-0088, National Highway Traffic Safety Administration, September 2014.

Comparison between Occlusion Methods for Measuring Distraction Caused by Smartphone Interaction

Christian Jernberg
Pontus Larsson
Driver Environment & Human Factors, Advanced Technology & Research, Volvo Group Trucks Technology, SE-40508 Göteborg, Sweden (christian.jernberg@volvo.com)
Paper Number 15-0404

ABSTRACT

Research Question/Objective
The occlusion method is an established method for measuring the visual demand from in-vehicle interfaces. The method is usually employed by means of occlusion glasses which can be automatically made opaque or transparent. However, occlusion can also be achieved by turning on and off the screen of the interface being studied. It can be hypothesized though, that glasses-occlusion requires more visual re-orientation efforts after each occluded interval which may give longer Total Shutter Open Times (TSOTs) than the method where only the interface screen is occluded.

Methods and Data Sources
Two experiments were conducted, with 10 participants (5 female) in Experiment 1 and 14 participants (7 female) in Experiment 2. In both experiments, participants were seated in a truck mockup and performed tasks on a smartphone. Two occlusion conditions were employed: 1) Occlusion with glasses, and 2) Embedded occlusion (turning the screen of the smartphone on/off). Participants also performed a baseline condition with no occlusion. In Experiment 1, five different tasks were performed in each condition: destination selection, dialing a phone number, dialing a contact, changing radio frequency and setting the alarm clock. For each task, a note showing the data to be entered was posted next to the smartphone, e.g. the number to be dialed.
The four tasks used in Experiment 2 were: visual-manual text entry, make changes in a truck driving log, dial a phone number and select item from a map (using pinch/swipe operations). The participants did not have support from any note but had to remember the data to enter for each task.
For both experiments, TSOTs and Resumability (R) were submitted to separate ANOVAs to reveal differences between occlusion and tasks.

Results
The results of Experiment 1 showed a main effect of occlusion type on TSOT ($p < .05$), where Glasses resulted in longer TSOTs ($M = 8.1$s) than Embedded ($M = 7.4$s). However, the interaction between occlusion and task was also significant ($p < .01$), and suggested that mainly the phone dialing task caused the difference in TSOTs. An explanation to this could be that tasks requiring more visual-spatial reorientation are more difficult to perform with glasses occlusion. A simpler explanation could be that participants needed to look at the note while dialing – which can be done during the occluded intervals with the embedded method. The results from experiment 2 did not show any statistically significant effect of occlusion on TSOTs which suggests that the effects found in Experiment 1 were caused by the fact that participants had to look at the paper note while dialing.

Discussion and Limitations
The current experiments indicate that the embedded method gives TSOTs comparable to the glasses-occlusion method. However, one must be aware of the fact that there are practical differences between the methods which can lead to diverging TSOTs. The current experiments show that such divergence can occur when instructions are given visually.

Conclusions and Relevance to Session Submitted
The embedded occlusion method for evaluating distraction is nonetheless promising since it is easy to use and can be integrated in app developers' toolkit.

INTRODUCTION

The occlusion method is well adopted and established for measuring the visual demand of in-vehicle information systems (IVIS). The method dates back to the late 60s and was developed by Senders et al.[1] in order to find a suitable surrogate for the more expensive and time consuming eye-tracking method.

The International Organization for Standardization (ISO) state in their guideline for occlusion that when using the occlusion method, the system that are being tested should work as intended and be fitted in a vehicle, simulator buck, or mock up in a design which duplicates the placement of the system in its intended environment. The viewing angle and the placement of the controls in relation to the driver should be the same as in the finished product[2]. Before testing the system the participant should be allowed to practice until he or she can adequately handle the system. A rule of thumb is mentioned as between two and five test trials per function, including two with occlusion. After the practice, a total of five test trials should be performed with similar end state of the system. Each function should be practiced and evaluated separately before moving on to the next.

The ISO standard specifies three possible ways to use occlusion, and these are as follows:

1 – Occlusion glasses
2 – Display blanking
3 – Mechanically or electronically operated shutters.

All these methods have in common that they are not allowed to interfere with the manual controls of the IVIS and the change from occluded to un-occluded state (opaque to transparent) or vice versa cannot take longer than 20 ms. The length of the interval during which the shutter is open or closed is 1.5s.

The two most relevant measurements used in this type of evaluation are the Total Shutter Open Time, TSOT, and the Resumability Ratio, R. The TSOT value represents an estimation of how long the driver would have to avert the gaze from the road to perform a task and R is a measurement of how easy it is for the driver to continue a dialogue after it is interrupted. R is calculated as the ratio between the TSOT and the total task time unoccluded (TTT_{unoccl}).

The occlusion method is accepted by ISO, the Society of Automotive Engineers (SAE)[3], the National Highway Traffic Safety Administration (NHTSA)[4], the Alliance of Automobile Manufacturers (AAM)[5] and the Japan Automobile Manufacturers Association (JAMA)[6] to determine if a system is safe enough to be used while driving. However, the criteria for how the method should be used might vary slightly. In the ISO standard the criteria is up to the person who is performing the evaluation or developing the interface. JAMA on the other hand say that the limit for total task time is 8 seconds, which make the maximum acceptable TSOT 7.5 seconds. For the NHTSA and AAM guidelines the maximum TSOT are twice as long, 15 seconds.

Occlusion should be optimal for evaluations in the early phases of the design process[7]; the main reason being that it is comparably cheap and simple to perform. Therefore, it is possible to have a larger number of participants for several design loops and iterations compared to when the eye-tracking method is used for example. So even though the occlusion method is not as precise in measuring where the most problematic areas of an IVIS are as the eye-tracking method, it can still give an indication of which concept that is the least distracting when comparing different ones[8]. Furthermore, if the display blanking variant of occlusion can be integrated in software development packages, it should even further simplify the process of iteratively developing and testing new in-vehicle interfaces.

In this study, the focus lies mainly with comparing the occlusion glasses to display blanking (hereafter referred to as “embedded occlusion”), and it is hypothesized that there are some differences between these which may lead to that the different methods result in different TSOT- and R values. For example, with the glasses there should be a short period of time where the eyes need to adjust to a new focal distance; something not needed when it is only the screen of the IVIS that is turned blank. Also, while the glasses do not interfere with the manual controls, it covers these as well as the hand of the operator. This is not the case with embedded occlusion and thus it should be easier to resume a task or even continue a task during the occluded interval (so

called ‘blind operating’) since there is still the ability to see where the fingers are positioned relative to the edge of the IVIS. This hypothesis is tested in two experiments presented in the following paragraphs in which a set of apps in a regular smartphone is evaluated using both the glasses-occlusion method and a smartphone-embedded method.

METHODS AND DATA SOURCES

Two experiments were conducted, between which the main differences were that Experiment 2 included tasks of higher difficulty and that participants did not receive any visual instructions during the tasks.

Experiment design

For both experiments, a within-group factorial design was used with Occlusion type (None/ Glasses / Embedded), Task (1-5 for Experiment 1 or 1-4 for Experiment 2) and Task repetition (5) as independent variables. The order of conditions was balanced between participants in order to avoid order effects.

Participants

In Experiment 1, 10 participants (5 female) took part. Their mean age was 39.5 years (*SD* = 9.25). Two participants were over 50 years old. In Experiment 2, 14 participants (7 female) took part. Their mean age was 36.7 years (*SD* = 6.98). All participants were recruited from within the company.

Apparatus and equipment

Apparatus and equipment was identical for the two experiments. A Samsung I9001 Galaxy S Plus smartphone was used as the main interface with which the participants would perform the task. The smartphone was positioned in a cradle mounted to the dashboard of a truck cab mockup. The truck cab mockup consisted of a truck seat, in which the participants were seated during the whole experiment, a steering wheel, and a complete dashboard (see Figure 1).

Occlusion was accomplished with PLATO visual occlusion spectacles for the glasses-occlusion conditions (see Figure 2) and with a custom app for the embedded occlusion conditions. Timing of the tasks was accomplished with a regular stop watch app in a smartphone.

Figure 1. Truck cab mockup used in both experiments

Figure 2. Person wearing occlusion glasses.

Tasks

In both experiments, five variations of each task were to be completed (which is according to the ISO standard), for example: call five different phone numbers or select five different destination addresses.

In Experiment 1, five different tasks were to be conducted in each occlusion condition. For each task in Experiment 1, a note showing the data to be entered was posted next to the smartphone, e.g. the number to be dialed. The tasks were:

1. Change radio frequency: In the smartphone's built-in radio app, the participants had to repeatedly tap either a left or right arrow until a frequency given by the paper note had been reached. In total, the arrow had to be tapped 31 times to complete each variation of the task.

2. Select destination: In the navigation app "Waze", the participants had to tap to open the menu, tap "navigate", scroll through a list and select one of five pre-defined destinations and then finally tap to start navigation.

3. Set the alarm clock: In the smartphone's built-in clock app, the participants had to first tap "create alarm", then tap the hour digits and through the appearing on-screen keyboard enter one out of five given times (20:00, 00.00, 17:00, 10:00, 18:00), tap "next", tap "repetition of alarm", tap "every day" and finally tap "Save" twice.

4. Dial a phone number: In the smartphone's regular phone interface, the participants had to manually enter one out of five phone numbers – all containing a long push "+" and 13 digits (e.g.: +49 157 3943 3782) and then tap the green phone symbol to call the number.

5. Dial a contact: In the smartphone's regular contact list, the participants had to scroll to one out of five names, tap the name, and finally tap the contact's phone number. All names started with the letter "M" and the participants had to scroll down approximately 30 contacts before the target contact became visible.

In Experiment 2, four different tasks were to be conducted in each occlusion condition. In this experiment, no note displaying the target data (i.e. phone number, destination address etc.) to be entered was used; the participants had to remember the data to be used in each task, and this was given verbally by the test leader before each task. The tasks were:

1. Enter destination: In the navigation app "Waze", the participants had to tap to open the menu, tap "navigate", tap in the search field and manually enter one out of five addresses given by the test leader with the on-screen keyboard, then tap the search icon, select the appropriate destination address from the list appearing and tap "drive" twice to start navigation. The addresses were Importgatan, Planetgatan, Teknikgatan,

Kvillegatan, and Londongatan (all selected from the vicinity of the test site so that no time lag would occur after the search/navigate icon had been tapped.)

2. Change time in log: In the app "TimeBook" (a tool for truck drivers to record their working and resting times), the participants had to tap "list" to open a list with logs, then tap one out of five specific logs (identified by the number of tasks they contained – between 21-25) , scroll to the last task in that log, tap the task, tap "end time" for the task and tap the "+" icon twice to extend the task by 2 hours and finally press "enter".

3. Dial a phone number: In the smartphone's regular phone interface, the participants had to manually enter one out of five phone numbers – all containing a long push "+", the numbers "495" and then 10 digits. Since the participants had to remember the number sequences by heart, logical number sequences were used (e.g.: +495 24 68 10 12 14). When the number had been dialed, the participants had to tap the green phone symbol to call the number.

4. Call dealer: In the app "DealerLocator" (a tool for truck drivers to find nearby truck dealers), the participants had to use multiple swipe / pinch-zoom operation to navigate a digital map, starting from Göteborg, Sweden and find one out of five dealers in the Barcelona, Spain region. Once the dealer had been located, the participant had to tap the dealer marker, and finally tap the dealer's phone number to make a call.

Procedure

The procedure was identical for both experiments. Participants arrived individually to the test lab. They were firstly introduced to the study, the smartphone, the occlusion glasses and the embedded occlusion app. Then, the tasks to be performed during the test were demonstrated.

The actual test then started. Before each task, participants practiced the task (using other data entries than the ones used in the experiment trials) until they could perform them without error. They then got to practice the task also with occlusion (either embedded or glasses depending on the condition) before an actual trial was conducted during which the task completion time was recorded by the test leader. The test leader instructed the participants verbally to start each task by saying "start… now" and the participant notified when he/she was finished with a task by saying "done". If the task was not completed without errors, they were asked to perform the task once more. After all tasks had been completed, participants were debriefed and thanked for their participation.

RESULTS

Using the acquired Total Task Times (TTT), the measures Total Shutter Open Time (TSOT) and Resumability ratio (R) were first calculated according to:

$$TSOT = 0.75 \cdot \left(INT\left(\frac{TTT}{1.5}\right)\right) + 0.5 \cdot MOD\left(INT\left(\frac{TTT}{1.5}\right), 2\right) + MOD\left(\frac{TTT}{1.5}, 1\right) \cdot \left(1 - MOD\left(INT\left(\frac{TTT}{1.5}\right), 2\right)\right) \quad (1)$$

$$R = \frac{TSOT}{TTT_{unoccluded}} \quad (2)$$

TSOT and R values were then submitted to separate 2x5x5 - occlusion (embedded/glasses) x (task number) x (task repetition) and 2x4x5 - occlusion (embedded/glasses) x (task number) x (task repetition) - repeated measures Analyses of Variance (ANOVAs) for Experiment 1 and Experiment 2 respectively.

Experiment 1 results

Considering TSOT, a statistically significant main effect of occlusion was found: $F(1,8) = 8.323$, $p = .02$. Post hoc test using Bonferroni's adjustment for multiple comparisons showed that the Glasses occlusion condition resulted in longer TSOTs than the Embedded occlusion ($M = 8.122$, $SE = 0.423$ compared to $M = 7.397$, $SE = 0.436$).

A main effect of task was also found: $F(4,32) = 61.843$, $p < .001$ (Sphericity assumed). Post hoc test showed that task 4 (dial phone number) resulted in statistically significantly longer TSOTs than all the other tasks (Bonferroni's method used to adjust for multiple comparisons). The effect of task on TSOT is shown in Figure 3 below.

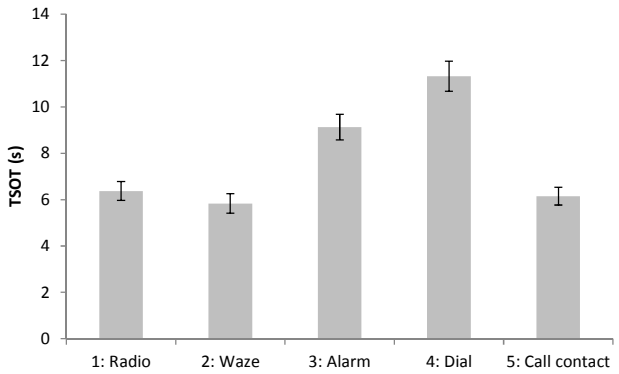

Figure 3. TSOT as function of Task for Experiment 1. Whiskers show standard error.

However, a statistically significant interaction effect between occlusion and task was also found: $F(4,32) = 6.143$, $p = .001$, Sphericity assumed. The effect is visualized in Figure 4 below, where it can be seen that it seems it is mainly Task 4 (Dial a phone number) for which the occlusion types differ.

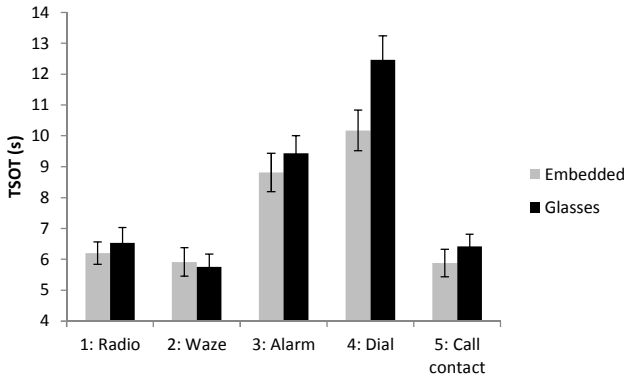

Figure 4. Interaction between occlusion and task for TSOT, experiment 1. Whiskers show standard error.

For Repetition, a main effect was found ($F(4,32) = 6.495$, $p = .001$), and from the result of a Bonferroni-adjusted post hoc test, one could see that this effect was due to repetition 4 resulting in a statistically significantly lower TSOT compared to repetition 1 ($M = 7.485$, $SE = 0.423$ vs $M = 8.313$, $SE = 0.516$, respectively, $p = .01$).

For R, no main effect of occlusion was found, although a trend could be noted indicating that Glasses occlusion could give higher R than embedded occlusion ($M = 0.849$ vs $M = 0.791$, $p = .067$). A similar trend was found for the interaction between task and occlusion. A statistically significant main effect was however found for task: $F(4,28) = 7.579$, $p < .001$. Figure 5 below shows that it was mainly the

difference between Task 1/ 2 (Radio / Waze) compared to Task 3 (Alarm) and the difference between Task 1 and 4 (Dial) which caused this effect (Bonferroni's method used to adjust for multiple comparisons). No other effects were found.

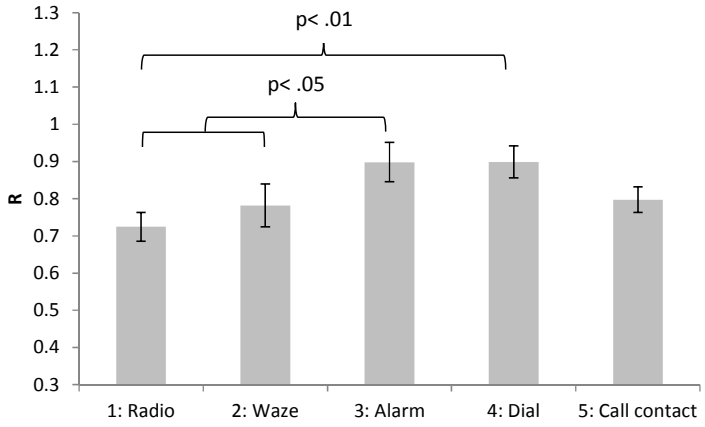

Figure 5. R as function of task for R, Experiment 1. Whiskers show standard error.

Taken together, the Dial and Alarm tasks seemed to require more glance time and were also more difficult to resume compared to the other tasks. The Dial task seemed to be the reason why a main effect of occlusion on TSOT was found. The reason for this effect may be that more difficult tasks, which require more visual reorientation than simpler tasks, are more difficult to perform with the glasses occlusion. A much simpler explanation could be that the Dial task required the participants to look at the paper note to be able to complete the task (the data to be entered for the other tasks were much simpler to remember). This could obviously not be done while the glasses were shut, compared to the Embedded occlusion where participants could read the number to be dialed also during blanked-sceen intervals.
These explanations were the reason for conducting Experiment 2, in which more difficult tasks were included and the participants did not have the paper note to read off.

<u>Experiment 2 results</u>

There was no effect of Occlusion on TSOT in experiment 2 ($F(1,13) = .160$, $p = .695$, $M = 15.163$ for embedded occlusion and $M = 15.708$ for glasses occlusion, $p = .695$). A main effect of Task on TSOT was however found: $F(1.637, 21.284) = 56.4$, $p < .001$ (Greenhouse-Geisser corrected). Post hoc test using Bonferroni's method of adjustment showed that all tasks differed from each other ($p < .01$); means and standard errors are visualized in Figure 6 below. Here it can also be seen that the Dial task, when performed with numbers in a logical sequence, resulted in lower mean TSOTs compared to Experiment 1, when the participants had to read on the paper note (of course not statistically tested here however). Interestingly, a similar effect of repetition as found in Experiment 1 was also found in Experiment 2: $F(4,52) = 4.401$, $p = .004$. Again, it seemed that Repetition 4 resulted in the lowest TSOTs, but also that the last repetition gave longer TSOTs than the others. The general trends for repetition for both Experiment 1 and 2 are shown in Figure 7.

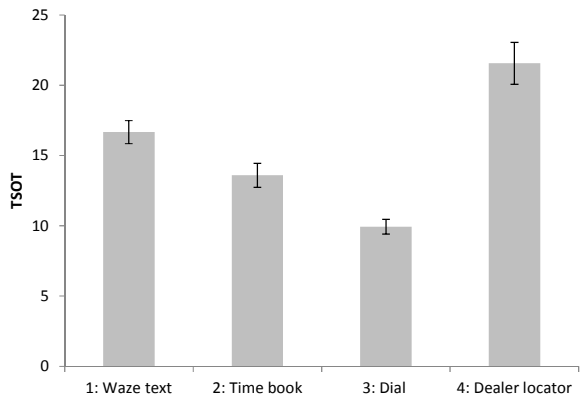

Figure 6. TSOT as function of task, experiment 2. Whiskers show standard error.

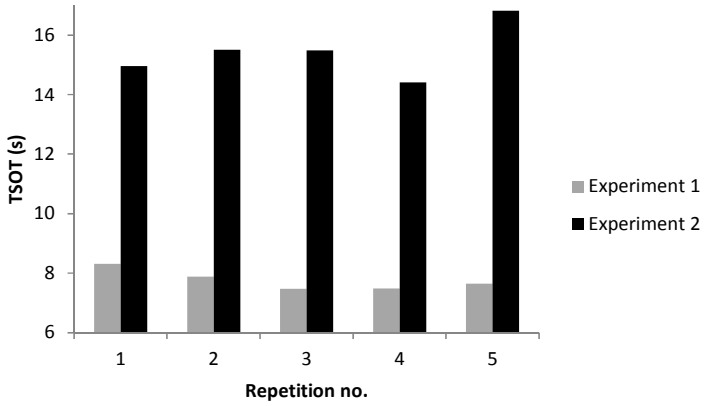

Figure 7. TSOT as function of repetition comparing both experiments.

No statistically significant effect of Occlusion on R was found: $F(1,12) = .013$, $p = .912$. Moreover, there were neither any statistically significant effects of Task ($F(3,36) = 2.348$, $p = .089$) or Repetition ($F(4,48) = 1.599$, $p = .190$) on R.

DISCUSSION AND LIMITATIONS

In Experiment 1, a statistically significant effect of occlusion on TSOT was found. However it was suspected that this effect was only due to one of the tasks, namely "Dial phone number". One possible explanation was that the use of visual instructions (a paper note with the number to be dialed) caused this effect. In case of the embedded occlusion, the phone number could obviously be read also during screen blanking, but when wearing the glasses the number can only be read during shutter-open intervals. Another explanation was that the occlusion types generate different results only for more difficult tasks which require visual reorientation, in turn which could be more difficult when the whole scene is blanked out as in the glasses-occlusion situation.

Experiment 2, in which there was no paper note to be read from, did not result in any differences in TSOT depending on occlusion type. Experiment 2 also included more difficult tasks, especially one involving interacting with a map which one would expect to be more difficult for the glasses occlusion - but even the measures obtained from this task did not seem to be affected by occlusion type. Thus, we can be fairly certain that the two occlusion methods will not generate different results, if visual instructions are avoided. It is important to note however, that the current set of experiments has not considered visual instructions and task difficult as factors in a factorial design (since this was not the purpose/hypotheses of the experiments from the beginning). To be able to draw firm conclusions,

another experiment should be conducted where visual instructions and task difficulty are independent variables.

CONCLUSIONS AND RECOMMENDATIONS

Our conclusion is that the occlusion method is a good way to early on determine if a system is too distracting for a driver or not. Even though it is not as accurate as eye-tracking, the simplicity with which it can be performed makes it ideal for app-developers to use while developing automotive applications. With the embedded occlusion method, they could test their systems before presenting them to a vehicle manufacturer and thus save time and resources for both parties.

However, a recommendation would be to add in the ISO guideline that the occlusion method should not allow visual instructions (apart from such that are presented in the actual interface under study) during the trials as this could interfere with the results of the evaluations.

ACKNOWLEDGEMENTS

The work presented in this paper was conducted as a part of the ongoing project SICS (Safe Interaction, Connectivity and State), co-financed by the Swedish Governmental Agency for Innovation Systems (VINNOVA).

REFERENCES

[1] Senders, J.W., Kristofferson, A.B., Levison, W.H., Dietrich, C.W. and Ward, J.L. 1967. "The attentional demand of automobile driving."
Highway Research Record, No 195: 15-33

[2] ISO 16673:2007. "Road vehicles – Ergonomic aspects of transport information and control systems – Occlusion method to assess visual demand due to the use of in vehicle systems."
Available for purchase at: http://www.iso.org/iso/catalogue_detail.htm?csnumber=38035. Accessed March 3, 2015.

[3] SAE J2364, 2004. "Navigation and Route Guidance Function Accessibility While Driving".
Available for purchase at: http://standards.sae.org/wip/j2364/. Accessed March 3, 2015.

[4] NHTSA: National Highway Traffic Safety Administration, 2013. "Guidelines for Reducing Visual-Manual Driver Distraction during Interactions with Integrated, In-Vehicle, Electronic Devices, Version 1.01 Docket No. NHTSA-2010-0053".
Available at: https://www.distraction.gov/download/11302b-Visual_Manual_Distraction_Guidelines_V1-0_010815_v1_tag.pdf. Accessed March 13, 2015.

[5] Driver Focus-Telematics Working Group, 2006. "Statement of principles, criteria and verification procedures on driver interactions with advanced in-vehicle information and communication systems."
Alliance of Automobile Manufacturers, Washington, DC.

[6] JAMA: Japan Automobile Manufacturers Association Inc., 2004 "Guidelines for In-Vehicle Display Systems - Version 3.0."
Available at: http://www.jama-english.jp/release/release/2005/jama_guidelines_v30_en.pdf. Accessed March 13, 2015

[7] Young, K, Lee, J.D. and Regan, M.A. 2008. "Driver Distraction: Theory, Effects and Mitigation"
CRC Press, Boca Raton, FL: 123-134

[8] Kroon, M. 2006. “Comparison Between Visual Occlusion and Eye-movement based visual demand assessment of in-vehicle information systems”
Departure of Computing Science, Umeå University, Sweden

RELIABILITY IMPROVEMENT NEEDED IN THE EYE GLANCE MEASUREMENT USING DRIVING SIMULATOR TEST

Mikael Ljung Aust
Annie Rydström
Robert Broström
Trent Victor
Volvo Car Corporation
Sweden

Paper Number 15-0398

ABSTRACT

To improve the interface design of in-vehicle infotainment systems, robust evaluation methods are required. The Eye Glance measurement using Driving Simulator test (EGDS) defined in the Visual-manual NHTSA Driver Distraction Guidelines for in-Vehicle Electronic Devices is a promising candidate. However, the present study indicates that EGDS needs further refinement to become sufficiently robust. When two randomly selected groups of 24 drivers tested the same ten in-vehicle tasks following the EGDS protocol, test outcomes were not the same for the two groups. The analysis showed this to be a consequence of how the EGDS pass/fail criteria are calculated. As currently formulated, they make test outcomes highly dependent on between-driver variability. To assess the problem magnitude with repeated EGDS testing, another eight virtual test groups were created by for each group randomly selecting 24 of the 48 participants' test scores. The analysis showed that EGDS outcomes were 60 % consistent between these ten groups. While six tasks consistently passed or failed, the outcome for the other four depended on which group had tested them. This EGDS reliability problem could possibly be overcome by matching the criteria calculation principles to the underlying population variability.

INTRODUCTION

Naturalistic driving studies indicate that looking away from the forward roadway for more than a certain period of time is a key crash contributing factor (Dingus et al, 2006; Klauer et al, 2006; Victor and Dozza, 2011; Liang et al, 2012). This fact, combined with the societal transition towards ubiquitous use of information technology, has put driver distraction on top of the traffic safety agenda.

From a vehicle design point of view, many potential sources of distraction such as eating, drinking and interacting with passengers lie outside the vehicle manufacturers' influence. However, they do control the design of all in-vehicle systems that involve visual-manual interaction while driving. To guide the design of these system, several design guidelines have been published. These include the European Statement of Principles (European Commission, 2008), the JAMA guidelines (JAMA, 2004) and the Alliance of Automobile Manufacturers guidelines (AAM, 2006).

These guidelines are sometimes difficult to apply, because while they may propose thresholds that in-vehicle tasks need to meet to be allowed while driving, they rarely propose methods for testing task compliance with those thresholds. For example, the AAM guidelines (AAM, 2006) suggest that each visual-manual task should require less than 2 seconds in mean glance duration and maximum 20 seconds total eyes off-road time, but there is no associated test method.

The latest addition, i.e. the Visual-manual NHTSA Driver Distraction Guidelines for in-Vehicle Electronic Devices (NHTSA, 2013), departs from the above tradition. It specifies both specific performance criteria against which in-vehicle tasks should be evaluated as well as specific test methods for testing criteria compliance. This paper is focused on the Eye Glance Measurement Using Driving Simulator Test Procedure, or EGDS (NHTSA, 2013). In EGDS, 24 randomly recruited test participants drive a lead vehicle following scenario on the highway while performing the in-vehicle tasks under evaluation. Their off-road glance durations during task performance are measured and assessed against the following Acceptance Criteria (AC), (NHTSA, 2013):

- *AC1: For at least 21 of the 24 test participants, no more than 15 percent (rounded up to the next whole number) of each participant's total number of eye glances away from the forward road scene have durations of greater than 2.0 seconds while performing the testable task one time.*

- *AC2: For at least 21 of the 24 test participants, the mean duration of all eye glances away from the forward road scene is less than or equal to 2.0 seconds while performing the testable task one time.*

- *AC3: For at least 21 of the 24 test participants, the sum of the durations of each individual participant's eye glances away from the forward road scene is less than or equal to 12.0 seconds while performing the testable task one time.*

If the tested task does not fulfil AC1-AC3, it should not be accessible while driving.

The publication of the NHTSA guidelines is very timely and will drive certainly drive vehicle safety forward. However, when looking closer into EGDS in its current formulation, there are indications that EGDS test outcomes may not be robust, i.e. that repeatability is low.

Two of the clearest indications come from two previous studies of off-road glance durations during interaction with in-vehicle tasks. Both Ljung Aust et al (2013) and Broström et al (2013) found that a significant portion of the test participants in their studies exhibited traits in their off-road glance behaviour that, if prevalent in the driver population at large, would lead to non-robust outcomes when testing in-vehicle tasks with the EGDS method. Both predicted that test outcomes would depend more on who the randomly selected test participants happened to be, rather than on the design of the task being tested.

These study results can be explained in two ways. One is methodological differences. Since both studies differed in setup and sampling procedure from the protocol in EGDS, it is possible that methodological differences are driving the conclusions rather than problems with the EGDS criteria.

Another possible explanation is that it is the way in which the acceptance criteria are calculated that leads to low robustness. Support for this latter explanation comes from a peculiar property of how the criteria are set up. In a general human factors study context, EGDS can be characterized as a between-group methodology, i.e. a new group of test participants are recruited for each test condition (e.g. each new set of task to be tested).

In terms of statistics, between-group designs are typically analysed on a group level. The data points for each test group are averaged or treated in some other way before comparisons between groups are made. Moreover, the important step, where sophisticated statistical methods such as ANOVA are used, is to assess if a difference in mean group values is significant. This step normally includes assessing whether the variability in each groups' test data is small enough to conclude that the observed mean value difference makes the groups significantly different. However, the EGDS acceptance criteria neither make the assessment on the group level, nor judge the outcome dependent on sample variability. Criteria fulfillment is instead determined by counting the frequency of individual outcomes.

In itself, a frequency counting approach does not lead to low test repeatability, as long as the variability in the tested group(s) does not overlap with acceptance criteria thresholds. If the thresholds however do overlap with population variability, it becomes a very different story.

To illustrate the problem, assume you want to performance test the height of manufactured door frames, and that the acceptance criteria state that anybody up to 10 ft. tall should be able to pass through. With this threshold, any randomly selected group of test participants is suitable for the assessment of frame heights, because even though people's normal height variability is several ft., nobody comes near the threshold limit.

However, if the threshold is lowered to 6'2", the situation changes. Now, the tested door frame will fail if any person in the test group is taller than 6'2", and pass if nobody is. Under this threshold, who comes to the lab becomes crucially important for test outcome. Given random participant selection and peoples' normal height

variability, the same 6'2" frame would pass the acceptance criteria on some days and fail on others, despite being 6'2" all the time.

For off-road glance duration measurements, there seems to exist a conceptual equivalent to tall people in the door frame example above. These drivers, aptly named "long glancers" (Broström et al, 2013), are ones who in general take their time in each step of an in-vehicle task interaction, quite independently of which task they are performing. Since long glancers were found to often look away for more than 2.0 seconds when doing in-vehicle tasks, and the current EGDS thresholds are set at 2.0 seconds, Broström et al (2013) predicted that the relative prevalence of such "long glancers" in an EGDS test group would have a large influence on the outcome. Essentially, if there are 4-5 long glancers in a test group of 24, any tested task tested would likely fail, regardless of how well designed it was.

Other studies also indicate that long glancers exist, and that their typical off-road glance durations overlap with EGDS criteria. One on-road study found that older drivers often glanced at the task display for more than 2 seconds at a time when doing in-vehicle tasks (Wikman and Summala, 2005). Donmez et al (2010) found in a simulator study that drivers could be divided into three visual strategy clusters where the "high risk cluster" looked away from the road for the longest periods, with mean off-road glance duration values over 1.6 seconds. Recently, Rydström et al (Rydström, 2015) found that when comparing two different groups of 24 randomly selected test participants performing the same ten tasks following the EGDS protocol in detail, the between group outcome was similar for less than half the tested tasks.

In light of the above, the first aim of the current study was to assess whether there exists an overlap between EGDS criteria thresholds and normal off-road glance duration variability when EGDS tests are performed fully according to the specified protocol (NHTSA, 2013). In addition, if an overlap was found, a second aim was to examine what level of robustness could be expected if the EGDS method would be used to assess 10 identical vehicles.

METHODOLOGY

Two full EGDS tests were performed at the test facilities of Ergoneers GmbH in Manching, Germany. A total of 48 participants, i.e. two test groups of 24 persons each, were recruited in the area around Manching in Germany. Each sample conformed to the EGDS recommendations (NHSTA, 2013) regarding general criteria, participant impartiality, age mix and gender balance.

Equipment

The test was conducted in an Ergoneers Sim-Lab left-hand drive, fixed base and open cab driving simulator (Figure 1). The simulator is designed as a replica of an actual production vehicle and was adapted to the EGDS driving simulator recommendations.

Figure 1: The Ergoneers Sim-Lab driving simulator

Participants drove a lead vehicle following scenario on an undivided, four lane, highway with a speed limit of 50 mph. The Ergoneers Dikablis head mounted eye tracking system was used to record participants' eye

movements. A touch-screen tablet containing a prototype infotainment system was mounted at the center panel of the simulator. Tasks 1-9 (see below) were performed on the tablet. An additional, conventional Panasonic car stereo was also installed to test the AAM radio manual tuning reference task (AAM, 2006).

Tasks performed

The prototype infotainment system was designed to replicate a typical in-vehicle center stack interface. Nine basic tasks were selected for assessment, and a manual radio tuning reference task was added to test a conventional car stereo (see Table 1).

Table1: Descriptions of the ten tasks used in the two EGDS tests

Task no.	Task	Task description
1	Activate vehicle function	Access menu and activate a specific vehicle function (on/off option)
2	Activate USB	Access menu and activate USB
3	Set new destination	Access navigation menu and activate a search field and tap in a destination of 10 characters and start guidance
4	Set favorite destination	Access navigation menu and a favorite menu, choose a specific destination and start guidance
5	Select and play music from USB	Access menu and activate USB and find a specified song in a list
6	Call contact from recent calls list	Access phone menu and recent calls list and choose a specified contact in the list to call
7	Dial known phone number manually	Access phone menu and keypad and tap in a well known 10 digit phone number and call
8	Change temperature	Access climate and increase temperature with 2 degrees
9	Set seat heat	Access seat heat and set it to a mid-level
10	AAM reference task - Manual radio tuning	Access FM1 and adjust to 106.7 (starting position FM3, preset 97.4)

Procedure

The test procedure, including test participant training, followed the EGDS protocol (NHTSA, 2013). Upon arriving to the test facilities, participants were given a brief description on the test procedure and asked to fill in a consent form, as well as a demographic questionnaire. Then they were asked to sit down in the simulator and adjust the seat. Next they put on the eye tracking glasses, and the eye tracking system was calibrated. The user interface and tasks were demonstrated to the participants and they were then instructed to practice each task as many times as they wanted to. Next, participants were given instructions on how to drive in the driving scenario, and got to practice driving it. When they felt comfortable driving, they got to practice performing the ten tasks while driving, as many times as they felt they needed. After training, data collection began. All tasks started from a home screen view, and ended when the task had been correctly executed.

Glance data analysis

After data collection was completed, the durations of all off-road glances for each test person during each task were calculated. This data set was then analysed to assess its basic properties and variability.

Acceptance test robustness analysis

To contextually frame the robustness analysis, it was assumed that NHTSA may test 10 vehicles in 2017, when the guidelines come into effect. Thus, a procedure was employed where 24 out of the 48 test participants from the current study were randomly selected using Matlab's random function to form a "virtual test group". The virtual test group essentially represents what the test outcome would have been if another combination of the 48 participants than the actual two is study had happened to been selected. The random selection of 24 participants was repeated 8 times, thus generating 8 virtual test groups in addition to the two real test groups.

Outcomes on the acceptance criteria were then evaluated for all 10 groups (the two real plus the eight virtual). Since the tested tasks were identical for all groups, this corresponds to a situation where 10 different test groups are brought in to evaluate the same vehicle 10 times over. Following the above example, this would correspond to NHTSA testing 10 vehicles with identical infotainment systems installed. For a method with high repeatability, one would expect very similar outcomes between all groups under these conditions.

RESULTS

The off-road glance duration distributions for the two tested groups are quite similar, as can be seen in Figure 2 below.

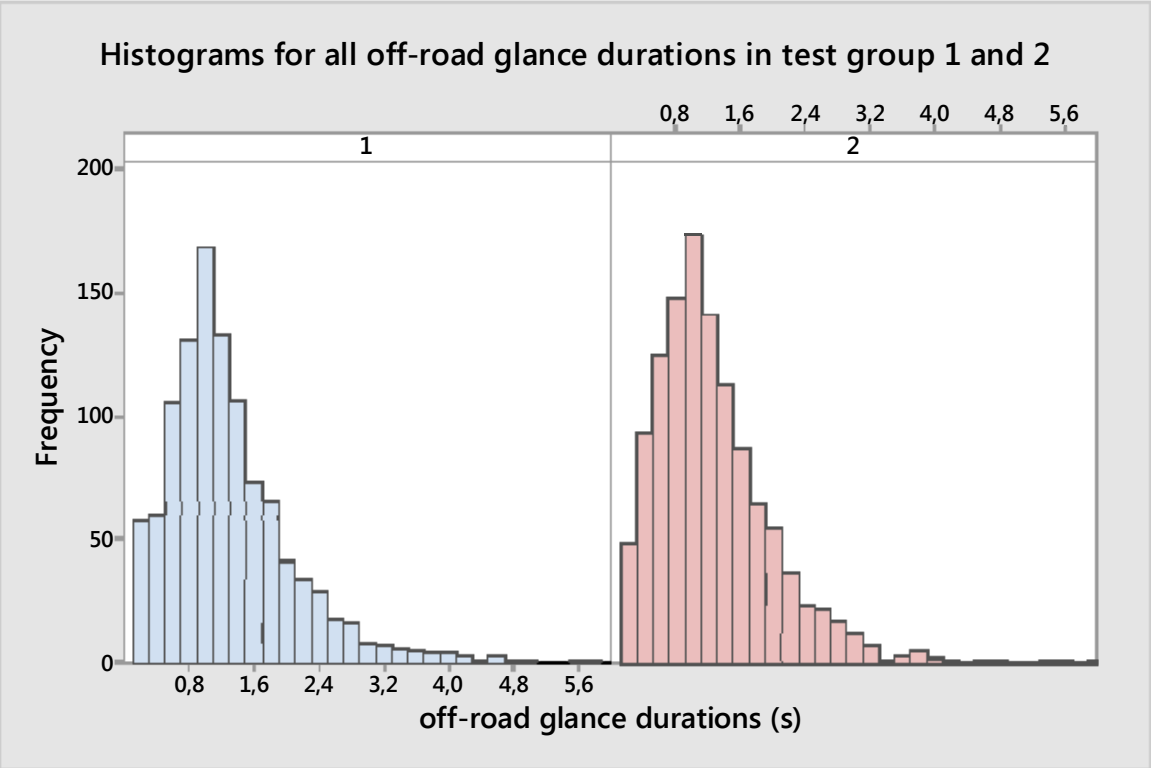

Figure 2: The histograms for the two groups show that they have similar off-road glance duration distributions

The glance duration data is also similar for the two groups in terms of means and standard deviations (Table 2).

Table 2: Means and standard deviations for off-road glance durations in the two groups

Group	Nr of glances	Mean	StDev
1	1085	1,3211	0,9023
2	1187	1,2605	0,7833

From Table 2, it is clear that the variability in the dataset overlap with the EGDS 2.0 second limits, since that lies within one standard deviation for both test groups.

The standard deviation indicates that the variability has a quite wide span. This is captured in Figure 2, which shows the average off-road glance duration value across all tasks for each test participant. As can be seen, the drivers with the shortest mean glance durations have values well below 1.0 second, while those with the longest are above 2.0 seconds.

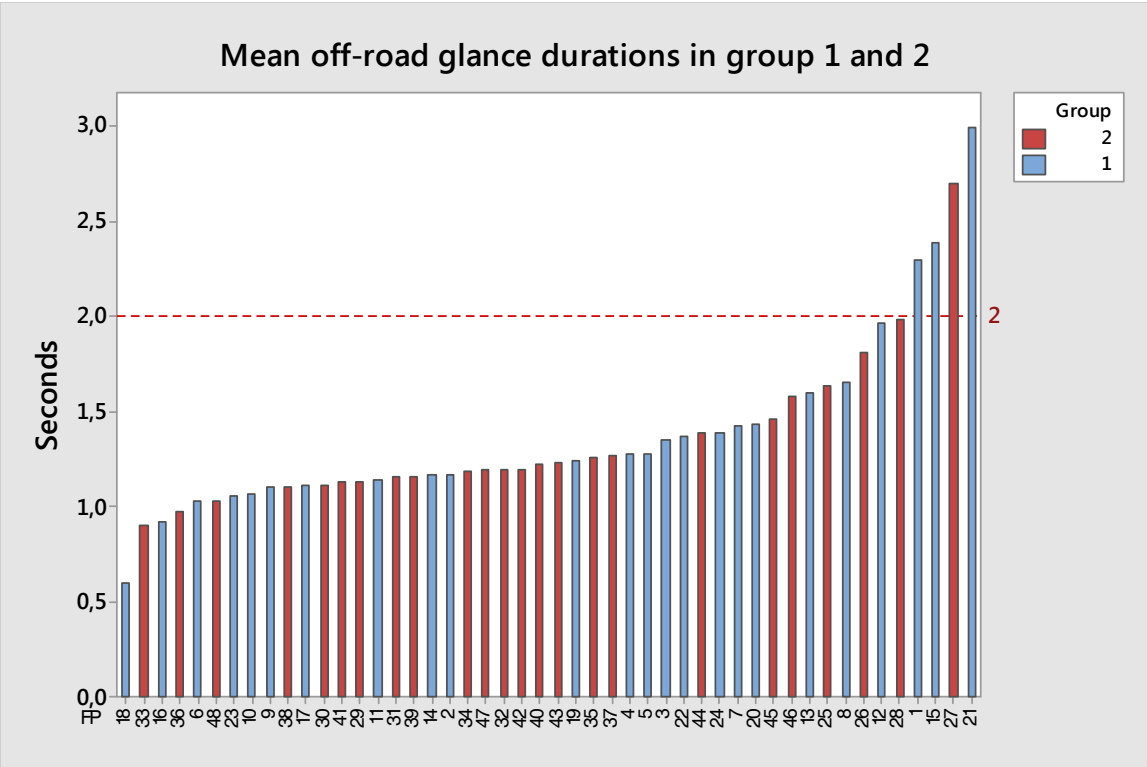

Figure 3: Mean off-road glance durations across all tasks for each test person, ranked according to duration. The reference line indicates the 2.0 s threshold in EGDS.

The performance outcome on the acceptance criteria for the two real test groups (RG1 and RG2) and the eight re-sampled groups (VG1-8) are shown in Table 3.

The results in Table 3 illustrate that the 10 tasks used in the study spans the whole range from difficult to simple. Setting a new destination fails on every criteria in almost every test group, while setting the seat heat and activating USB pass on all criteria in all test groups. The task set selected for this study thus seem to cover a relevant span for a reliability analysis.

Table 3 also indicates that testing with the EGDS method as currently formulated is not robust. For AC1, the outcome is inconsistent for 5 of 10 tasks. For AC2, the outcome is inconsistent for 7 of 10 tasks. For AC3, the outcome is inconsistent for 2 of 10 tasks.

When all three acceptance criteria are jointly considered on a per task basis, the inconsistencies remain. If the results of Table 3 were applied to the hypothetical test of 10 vehicles with identical infotainment systems in 2017 conjectured above, six tasks would consistently be allowed or prohibited while driving, while for the other four, it would depend on the group of test participants that tested them. Thus, the EGDS criteria yield 60 % consistency between groups in testing.

		Activate vehicle function	Activate USB	Set a new destination	Set a favorite destination	Select and play music from USB	Call contact from recents list	Dial known phone number manually	Change temp.	Set seat heat	AAM ref task: Manual radio tuning
AC1 For each driver, max 15% glances > 2.0 s	RG1	2	3	8	2	8	3	11	6	1	5
	RG2	3	1	10	4	6	4	10	2	0	3
	VG1	2	0	8	3	7	2	6	3	0	3
	VG2	2	2	7	3	7	5	10	4	1	4
	VG3	2	3	7	0	5	4	9	4	0	3
	VG4	2	1	7	4	4	2	7	3	1	4
	VG5	4	2	7	4	10	5	10	5	1	5
	VG6	2	2	8	2	5	2	9	3	1	2
	VG7	1	3	7	3	9	3	9	4	1	4
	VG8	4	2	7	4	8	3	9	5	1	7
AC2 For each driver, mean glance duration < 2.0 s	RG1	2	3	6	6	6	4	9	6	1	2
	RG2	3	1	4	2	3	3	3	2	0	1
	VG1	2	0	4	1	4	1	2	1	0	0
	VG2	2	2	4	3	5	6	7	4	0	0
	VG3	2	2	4	3	2	3	5	4	0	2
	VG4	2	1	2	1	2	3	4	3	1	2
	VG5	3	3	6	5	6	4	7	5	1	2
	VG6	2	2	2	2	4	2	5	3	1	0
	VG7	1	3	5	5	5	4	6	4	1	1
	VG8	4	3	5	3	6	4	7	5	1	2
AC3 For each driver, total glance time < 12.0 s	RG1	0	0	9	0	3	0	3	0	0	12
	RG2	0	0	13	0	1	0	3	0	0	16
	VG1	0	0	11	0	1	0	5	0	0	10
	VG2	0	0	7	0	2	0	3	0	0	11
	VG3	0	0	13	0	1	0	1	0	0	17
	VG4	0	0	13	0	3	0	4	0	0	15
	VG5	0	0	8	0	3	0	3	0	0	13
	VG6	0	0	11	0	4	0	3	0	0	13
	VG7	0	0	12	0	3	0	4	0	0	14
	VG8	0	0	9	0	3	0	1	0	0	14

Table 1: Outcome on the EGDS acceptance criteria for the two real and eight virtual test groups

DISCUSSION

The first aim of the current study was to assess whether an overlap between EGDS thresholds and normal off-road glance duration variability is present when EGDS tests are performed fully according to protocol (NHTSA, 2013). The results show that an overlap does exist.

Both Broström et al (2013) and Ljung Aust et al (2013) found that off-road glance lengths when performing in-vehicle tasks while driving were quite individual and task independent, i.e. the between driver variability on the

glance metrics was much larger than the within driver variability. The results in Figure 3 show this to be the case also in the current study. While the drivers with the shortest mean values are well below 1.0 seconds, those with the longest are well above 2.0 seconds.

Since the present study followed the proscribed EGDS procedure in detail, this indicates that this is not an artefact of the test set up, which one possible explanation for the Broström et al (2013) and Ljung Aust et al (2013) results. Rather it seems to be attributable to properties of the tested population. In other words, drivers seem to have their own, personal, pace by which they performed in-vehicle tasks while driving, and this is reflected in how long their off-road glances are.

As Figure 3 illustrates, for some drivers this individual pace result in off-road glance durations that make them prone to fail on acceptance criteria 1 and 2 regardless of which task is being tested. The likelihood of a task passing the EGDS test thus becomes highly dependent on how many long glancers happen to be included in the participant sample.

The second aim of this study was to examine what the effect of the overlap would be on a series of EGDS tests. Here, it was found that the effect was substantial. On a per task level, EGDS was only 60 % consistent between test groups, which seems low for a test meant to determine which in-vehicle tasks should be allowed while driving.

However, reliability levels could probably be improved with only minor alterations to EGDS. The current study illustrates that there is a mismatch between how the EGDS criteria are currently calculated and the underlying population variability. This mismatch could potentially be resolved by using other approaches than a frequency count to calculate criteria fulfilment. For example, one could study what reliability level is reached if thresholds are calculated on a group mean level, perhaps with additional thresholds determined by standard deviations. This is an urgent topic for future research.

A limitation of the current study is that it only involves two test groups, so more studies are naturally needed. Another limitation is that the re-sampling used to construct the 8 virtual test groups may lead to underestimation of observed variability. However, since this would lead to an over- rather than underestimation of robustness, the above method reliability analysis is likely a conservative one.

REFERENCES

Alliance of Automobile Manufacturers. 2006. “Statement of Principles, Criteria and Verification Procedures on Driver-Interactions with Advanced In-Vehicle Information and Communication Systems”, Driver Focus-Telematics Working Group, Washington, DC., June 26, 2006 update.

Broström, R., Bengtsson, P., & Axelsson, J. (2011). Correlation between safety assessments in the driver–car interaction design process. *Applied Ergonomics, 42(4)*, 575-582.

Broström, R., Ljung Aust, M., Wahlberg, L., & Källgren, L. (2013). What drives off-road glance durations during multitasking – capacity, practice or strategy? In *3rd International Conference on Driver Distraction and Inattention*. Gothenburg, Sweden.

Dingus, T.A., Klauer, S.G., Neale, V. L., Petersen, A. et al. 2006. “The 100-Car Naturalistic Driving Study, Phase II – Results of the 100-Car Field Experiment.” Cambridge: U.S. Department of Transportation, John A. Volpe National Transportation Systems Center.

Donmez, B., Boyle, L. N., & Lee, J. D. (2010). Differences in Off-Road Glances: Effects on Young Drivers’ Performance. *Journal of Transportation Engineering*, *136*(5), 403–409. doi:10.1061/(ASCE)TE.1943-5436.0000068

European Commission. 2008. “Commission Recommendation of 26 May 2008 on Safe and Efficient In-Vehicle Information and Communication Systems; Update of the European Statement of Principles on Human-Machine Interfaces.”

JAMA. 2004. “Guideline for In-Vehicle Display Systems, Version 3.0”. Japan Automobile Manufacturers Association, Tokyo, Japan, August 2004.

Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweeks, J., & Ramsey, D. J. 2006. “The impact of driver inattention on near-crash/crash risk: An analysis using the 100-Car Naturalistic Driving Study data.” Blacksburg, VA: Virginia Tech Transportation Institute, 2006.

Liang, Y., Lee, J.D., Yekhshatyan, L. 2012. “How dangerous is looking away from the road? Algorithms predict crash risk from glance patterns in naturalistic driving”. Published online before print, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, June 15, 2012.

Ljung Aust, M., Dombrovskis, S., Kovaceva, J., Svanberg, B. & Ivarsson, J. (2013) “An Empirically Based Suggestion for Reformulating the Glance Duration Criteria in NHTSA's Visual-Manual Interaction Guidelines”, in the proceedings of SAE 2013 World Congress, 15-18 April, Detroit, USA. Paper no 2013-01-0444.

Rydström, A., Ljung Aust, M., Broström, .R., & Victor, T. (2015) Repeatability of the NHTSA Visual-Manual Guidelines Assessment Procedure, paper nr 15-0240, ESV 2015.

Victor, T. and Dozza, M. “Timing matters: Visual behavior and crash risk in the 100-Car online data”. Presentation given at the 2nd International Conference on Driver Distraction and Inattention, Gothenburg, Sweden, Sept 5-7, 2011. Accessed 120125 at http://www.chalmers.se/safer/ddi2011-en/

Wikman, A., & Summala, S. (2005) Aging and Time-Sharing in Highway Driving. Optometry and Vision Science, 82, 716-723.

Visual-manual NHTSA Driver Distraction Guidelines for in-Vehicle Electronic Devices (2013). Docket no. NHTSA-2010-0053, National Highway Traffic Safety Administration, April 2013.

DRIVER BEHAVIOR DURING LANE CHANGE FROM THE 100-CAR NATURALISTIC DRIVING STUDY

Rong Chen
Virginia Tech
United States

Kristofer D. Kusano
Virginia Tech
United States

Hampton C. Gabler
Virginia Tech
United States

Paper Number 15-0423

ABSTRACT

Lane changes with the intention to overtake the vehicle in front are especially challenging scenarios for forward collision warning (FCW) designs. These overtaking maneuvers can occur at high relative vehicle speeds and often involve no brake and/or turn signal application. Therefore, overtaking presents the potential of erroneously triggering the FCW. A better understanding of lane change events can improve designs of human-machine interface and increase driver acceptance of FCW. The objective of this study was to characterize driver behavior during lane change events using naturalistic driving data.

The analysis was based on data from the 100-Car naturalistic driving study, collected by the Virginia Tech Transportation Institute. The 100-Car study contains approximately 1.2 million vehicle miles of driving and 43,000 hours of data collected from 108 primary drivers. In order to identify overtaking maneuvers from a large sample of driving data, our study developed and validated an algorithm to automatically identify overtaking events. The lead vehicle and minimum time to collision (TTC) at the start of lane change events was identified using radar processing techniques developed in a previous study. The lane change identification algorithm was validated against video analysis which manually identified 1,425 lane change events from approximately 126 full trips.

Forty-five (45) drivers with valid time series data was selected from the 100-Car study. From the sample of drivers, our algorithm identified 326,238 lane change events. Lane change events were evenly distributed between left side and right side lane change. The characterization of lane change frequency and minimum TTC was divided into 10 mph speed bins for vehicle travel speeds between 10 mph to 90 mph. A total of 90,639 lane change events were found to involve a closing lead vehicle. For all lane change events with a closing lead vehicle, the results showed that drivers change lanes most frequently in the 50-60 mph speed range. Minimum TTC was found to increase with travel speed, and the variability in minimum TTC between drivers also increased with travel speed.

INTRODUCTION

One of the most frequent crash modes on the roadway is rear end collisions. The National Highway Traffic Safety Administration (NHTSA) reported 1.8 million rear end crashes occurred in the United States in 2012 [1]. Forward collision warning (FCW) systems have great potential to reduce rear end collisions on the roadway. Studies have estimated that FCW can reduce the number of fatally injured drivers by as much as 29% [2]. However, driver acceptance of the systems is paramount to the effectiveness of a FCW system. If a system delivers the warning too early the driver may be annoyed and turn off the system.

Lane changes with the intention to pass the vehicle in front are especially challenging scenarios for FCW designs because these overtaking maneuvers can occur at high relative vehicle speeds and often involve no brake and/or turn

signal application. Therefore, overtaking presents the potential of prematurely triggering the FCW. A better understanding of lane change events can increase driver acceptance of FCW and improve the effectiveness of FCW systems. The objective of this study is to characterize driver behavior during lane change events using the 100-Car Naturalistic Driving Study.

METHOD

The 100-Car Naturalistic Driving Study

The 100-Car study was a landmark large-scale Naturalistic Driving Study (NDS) conducted by the Virginia Tech Transportation Institute (VTTI) from 2001 to 2004. The study contains approximately 1.2 million vehicle miles of driving and 43,000 hours of data collected from 108 primary drivers and 299 secondary drivers [3]. Drivers were recruited from the Washington D.C. metropolitan area. No restrictions were used to select subjects, e.g. excluding those with traffic violations. Younger drivers, i.e. under 25 years, and self-reported high mileage drivers were, however, sought and oversampled [4].

Vehicles were instrumented with cameras and inertial measurement devices and equipped with a PC-based computer to collect and store the data. The data collection box housed a yaw rate sensor, dual axis accelerometers, and a GPS navigation unit. In addition, radar sensors were mounted on the front and rear of the vehicle that were able to track other vehicles. The data collection box was usually installed on the roof of the trunk of the vehicle in order to be unobtrusive. All data were sampled at a rate of 10 samples per second. Some of the sensors had lower sample rates. These data were still sampled at 10 samples per second but would have multiple samples with equal magnitude.

There were five (5) cameras that offered continuous views in and around the vehicle. An exemplar view is shown in Figure 1. The upper left frame shows a view of the driver's face and upper body, blurred to protect the identity of the driver in Figure 1. The lower left pane is an over-the-shoulder view of the driver, the upper right pane is a forward view out the front of the vehicle, and the lower right pane is split between a view out the passenger side of the vehicles and out the rear of the vehicle. The video was useful in the study in interpreting vehicle instrumentation data.

Figure 1. Combined Video Views from the 100-Car Naturalistic Study [4]**.**

Overall Study Approach

Our approach for the current study was broken down into three subtasks, as shown in Figure 2. The first subtask of the study was to find the lane change maneuvers and determine the start and finish of the lane change events. Since not all lane change events involve braking and/or turn signal use, we have developed an algorithm which used the lane tracking instrumentation data to identify lane change events. The second subtask of the study was to identify whether a lead vehicle was present during lane change events. Lastly, minimum Time to Collision (TTC) was computed for each lane change maneuver.

Figure 2. Overall Project Approach

Automated Search Algorithm Development

The identification of lane change events was based on the lane change signal recorded by the on-board lane tracking system. The lane change signal was triggered when the vehicle center line met the lane edge as the vehicle moved across the lane. In addition to the distance of vehicle centerline relative to the lane boundary, the lane tracking system also records the confidence level, from 0 to 100%, that the lane tracking system was providing correct distance evaluation. The following conditions must have been met in order to be accepted as a valid lane change: the duration of the lane change maneuver was greater than 0.3 seconds, no aborted lane change signals, vehicle speed above 10 mph, and average lane tracking confidence greater than 30% during the duration of lane change.

Figure 3 shows the definition of lane change duration used in this study. The start of lane change was defined as the time when the vehicle leading edge met the lane edge, and the end of lane change event was defined as the time when the vehicle was completely within the boundaries of the adjacent lane.

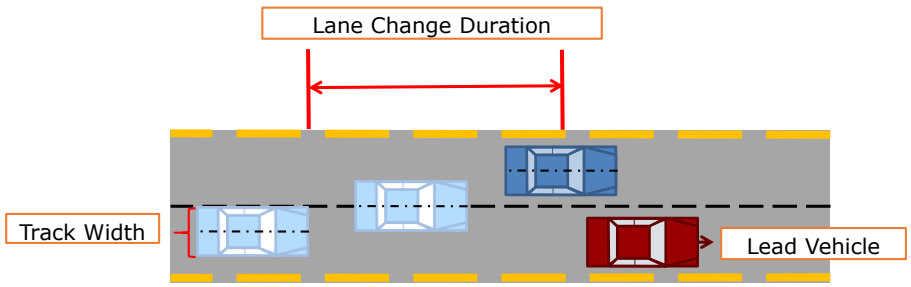

Figure 3 Definition of Lane Change Duration.

The lead vehicle identification algorithm developed in this study was based on a previously methodology [5]. The lead vehicle detection algorithm utilizes the on-board instrumentation, such as radar, vehicle speed, vehicle azimuth, and vehicle acceleration, to identify the correct lead vehicle at the start of a lane change event.

For each lane change with a lead vehicle present, the TTC was calculated as the range divided by range rate at the start of the lane change event, as shown in Equation (1).

$$TTC = \frac{x}{\dot{x}} \quad (1)$$

Algorithm Validation

Both the lane change identification and lead vehicle detection algorithm were validated against a sample of 1,425 manually identified lane change events. The sample of validation lane change events was examined from inspection of 126 trips, containing nearly 50 hours of video footage.

Table 1 shows the distribution of lane change maneuvers and the presence of lead vehicles in the validation sample. Merging lane change maneuvers include entering and exiting the highway. Lane changes involving drivers changing lanes to pass a lead vehicle were considered overtaking events. All other events were labeled “Lane

Change – Other". Examples of "Lane Change – Other" include events when drivers change lanes but do not overtake any vehicle. As shown by Table 1, 94% of the passing lane change events included the presences of a lead vehicle, as identified by reviewer. However, only 42% of merging events and 37% of lane change events had lead vehicle presence. By our definition, overtaking events should always have a lead vehicle. However, during the validation process, if the reviewer could identify a lead vehicle in the video but the on-board radar did not identify a lead vehicle, the particular lane change event was marked as not having a lead vehicle. Based on the large percentage of merging events and the low incidence of lead vehicle presence during merging events, this sample of validation data suggest that in most lane change event there will not be a lead vehicle present.

Table 1.
Lane Change Maneuver and Presence of Lead Vehicle

Maneuver	Total Number of Lane Changes	Lead Vehicle Present (Count)	Lead Vehicle Present (% of Lane Changes)
Merging	439	184	42%
Overtaking	420	395	94%
Lane Change - Other	566	209	37%
Total	1425	788	55%

RESULTS

Algorithm Validation

The performance of the lane change detection algorithm was evaluated based on contingency matrix shown in Figure 4. The sensitivity (% of correctly detected true positives) and specificity (% of correctly detected true negatives) of the algorithm was calculated as the ratio of true positive and true negative to the corresponding lane change maneuver.

Table 2 summarizes the comparison of the video and lane tracking indicated lane change events. For all valid lane changes with sufficient lane tracking information (i.e. lane tracking confidence > 30%, vehicle speed > 10 mph, and headway < 3 s). The lane change detection algorithm performed very well. The sensitivity of the algorithm was 0.87 and the specificity of the algorithm was 0.98.

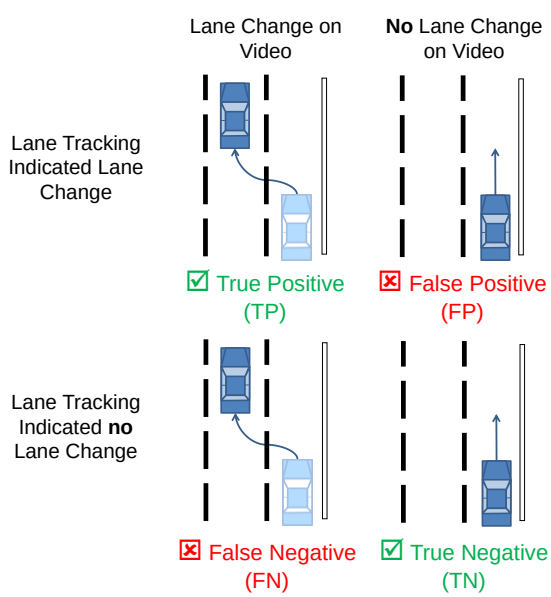

Figure 4: True Positive and False Negative Matrix

Table 2.
Comparison Results of Video and Lane Tracking Indicated Lane Changes

Lane changes with sufficient lane tracking information	645
$Sensitivity = \frac{TP}{TP + FN}$	0.87
$Specificity = \frac{TN}{TN + FP}$	0.98

The lead vehicle detection algorithm correctly identified the car following situation in 84% of the lane changes in the validation sample. Figure 5 shows the lead vehicle identification algorithm performance in five different car following scenarios. Scenario a) is the correct identification of a lead vehicle when there is lead vehicle present, b) is the correct determination of no lead vehicle, c) is the identification of a lead vehicle when no lead vehicle is present, d) is no lead vehicle is identified when a lead vehicle is present, and e) is identifying the wrong vehicle as the lead vehicle.

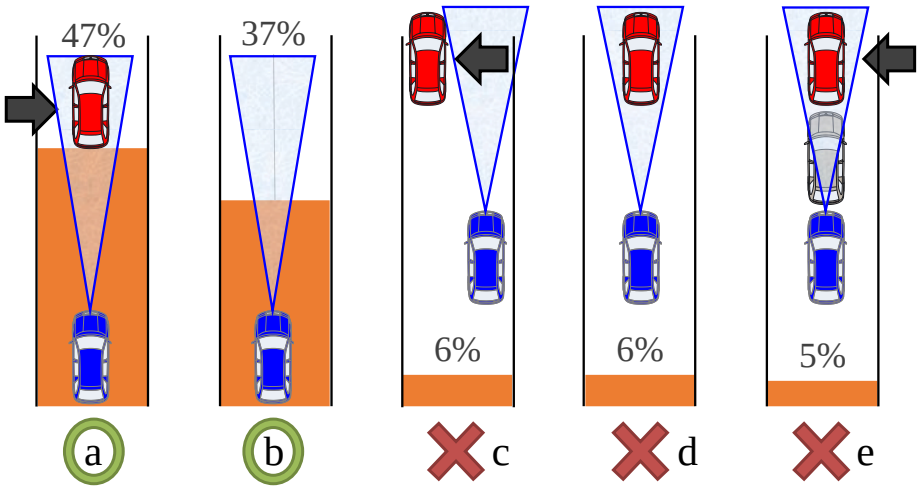

Figure 5: Lead Vehicle Detection Scenarios

Driver Selection

A total of 108 primary drivers and 299 secondary drivers were included in the 100-Car study period in which all driving in an instrumented vehicle was recorded [3]. Primary drivers were the primary owners or leasers of the instrumented vehicles. Secondary drivers occasionally drove the vehicles. Primary drivers accounted for 89% of all miles driven during the study period. The entire 100-Car database contains approximately 1.2 million miles of driving, 1,119,202 miles of which were driven by primary drivers in 139,367 trips [3]. Some primary drivers drove in multiple vehicles. For the current study only trips where a primary driver was driving in the vehicle that he or she most frequently drove during the study, i.e. their primary vehicle, were selected.

Prior to the analysis, the status of all time-series data was inspected. Instrumentation data, including front facing radar, vehicle speed, brake switch status, yaw rate signals, and lane tracking, were checked for missing or invalid data. The current study only included vehicle which had valid data in at least 60% of all trips and 60% of all distance traveled. Table 3 summarizes the sample of trips with valid sensor data. A total of 46,250 trips from 45 drivers were used in this study.

Table 3.
Driver Selection with Valid Sensor Data

Number of Drivers	45
Number of Trips	46,250
Total Miles Driven	406,606.7
Median Trip Distance (miles)	4.6
Median Trip Duration (minutes)	11.4

Population Distribution of TTC at Lane Change

Table 4 shows the result of the lane change detection algorithm for all drivers with valid sensor data, organized by lane change scenarios. A total of 326,238 lane changes were found in the 46,250 trips. The distribution of left side and right side lane changes was essentially even. A total of 90,639 lane changes involved a closing lead vehicle. Closing lead vehicles was defined to be lane change events where the driver was moving closer to the lead vehicle. As shown in the table, lane changes with closing lead vehicle accounted for approximately 28% of all lane change events. Similar to the distribution of all lane change events, left side lane change accounted for approximately 52% of all lane changes with closing lead vehicle, and right side lane changes was approximately 48% of all lane changes with closing lead vehicle.

Table 4.
Lane Change Scenario Distribution

	All Lane Changes (% of All Lane Changes)	**Lane Change with Closing Lead Vehicle (% of Lane Change with Closing Lead Vehicle)**
Left Side Lane Change	171,519 (52.6%)	47,420 (52.3%)
Right Side Lane Change	154,719 (47.4%)	43,219 (47.7%)
Total Lane Change	326,238 (100%)	90,639 (100%)

Table 5 summarizes the distribution of lane change frequency. The zero 10th percentile values in each of the measurements in the table resulted from the fact that no lane changes were detected in approximately 35% of the trips.

Table 5.
Lane Change Frequency Distributions

Measurement	**10th Percentile**	**Median**	**90th Percentile**
Lane change per trip	0	3	19
Lane change per mile	0	0.8	2.3
Lane change per hour	0	19.7	58.4

Figure 6 shows the distribution of total number of lane changes with a closing lead vehicle within each speed bin. The label above the figure shows the number of driver who made lane changes within the speed bin, e.g. n = 45 drivers in the 10-20 mph speed bin. Not all drivers made lane changes within each speed bins, therefore not all speed bins had 45 drivers. The figure shows a skewed left distribution, suggesting that lane change events with closing lead vehicles are more likely to occur at higher travel speeds, such as suburban roads with speed limits

ranging from 35-45 mph, and highways with speed limits ranging from 55-65 mph. As shown in the figure, the speed bin with the most number of lane changes is the 60-70 mph range. This speed range corresponds to the highest highway speed limit surrounding the Washington D.C. metropolitan area, and could be the result of frequent driver movement through traffic gaps in congested area. However, this total is skewed by one outlier driver, who had more than 4,000 lane changes in the 60-70 mph speed range.

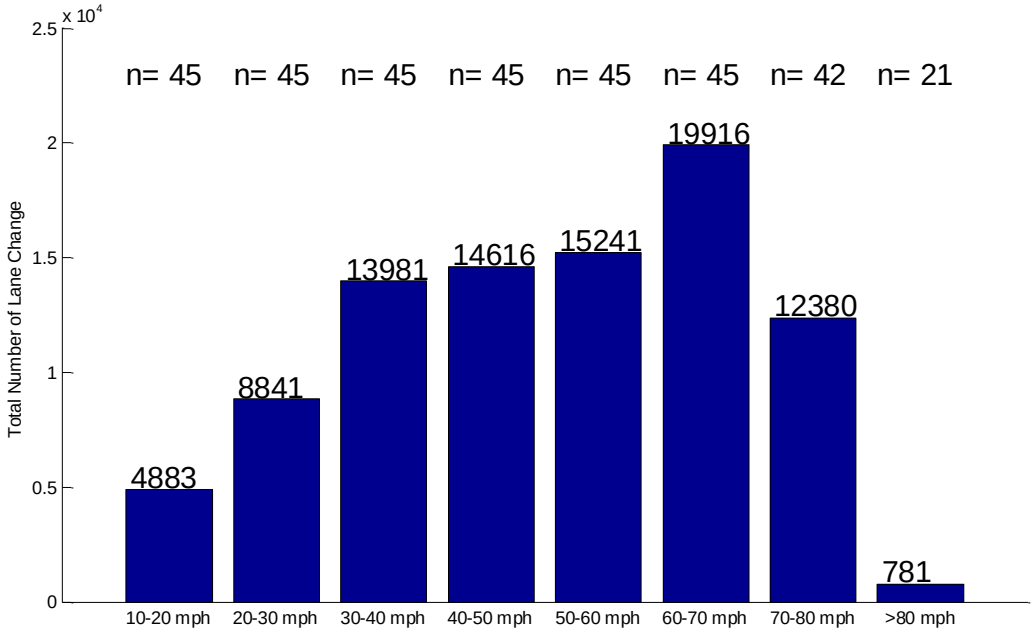

Figure 6: Distribution of Lane Change with Closing Lead Vehicle and TTC (90,639 total lane changes)

Figure 7 shows the distribution of minimum TTC by driver in each speed bin. Each driver has one point in each 10 mph speed bin. The cumulative distribution of minimum TTC for each driver are plotted below for each speed bin. In Figure 7, TTC was computed as the TTC at the start of lane change, and minimum TTC for each driver refers to the lowest (closest to zero) TTC for trips within each speed bin.

In general, the minimum TTC increases with travel speed. The variability in minimum TTC between drivers also increases with travel speed. The plot shows the number of drivers who made lane changes with a closing lead vehicle in each speed bin (ex. N = 45 in the 60-70 mph bin). Due to the relatively low percent of lane change at high speeds, not all drivers had lane changes with a closing lead vehicle in the >80 mph speed range.

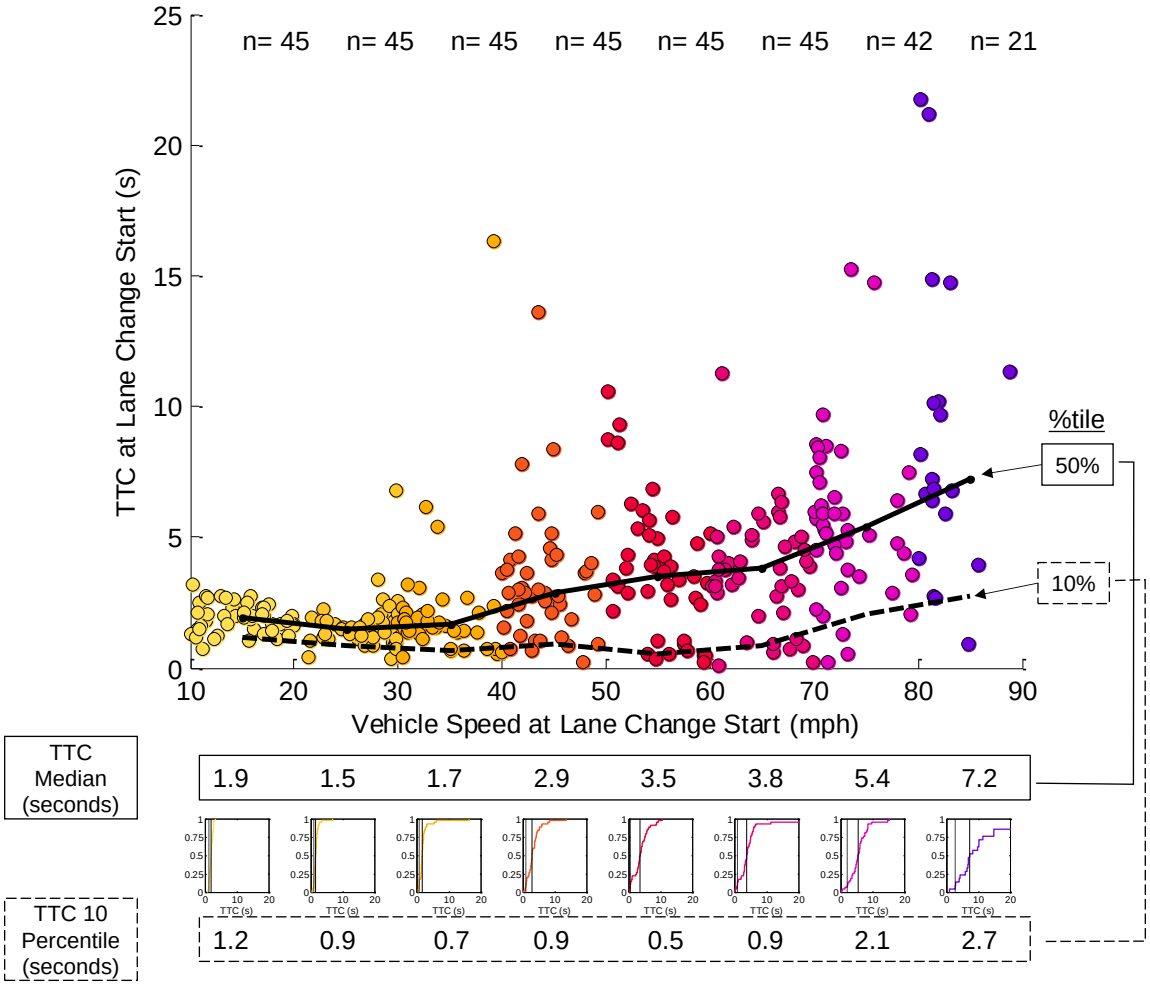

Figure 7: Minimum TTC for Each Driver by Travel Speed Bin (n =90,639 lane changes)

CONCLUSIONS

The objective of this study was to characterize driver behavior during lane change events. This study presented a methodology to detect lane change events and a methodology to identify lead vehicle in lane change events in the 100-Car NDS. The performance of these two algorithms was validated against 126 trips, in which the researchers manually examined the video footage to determine the time frame of lane change and lead vehicle presence. Finally these algorithms were applied to the 100-Car NDS dataset to obtain the distribution of TTC at lane change with a lead vehicle in front of the subject car.

In 126 randomly selected validation trips, the researcher manually reviewed the trip video and identified 1,425 lane change events. 420 events were found to be overtaking event, in which the video shows a driver passing a lead vehicle. A total of 439 merging lane change was found in the sample, in which a closing lead vehicle was present in about 42% of the merging events. For lane change events other than overtaking and merging (566 events), a closing lead vehicle was present in 37% of these events. Overall, the lane change detection algorithm performed very well. For lane changes with sufficient lane tracking information, the validation process showed that the automated lane change algorithm had a sensitivity of 0.87 and a specificity of 0.983. The lead vehicle identification algorithm also

performed reasonably well, and correctly identified the car following situation in 84% of the validation sample of lane change events. Our lane change detection algorithm is highly dependent on the presence of lane markings, therefore our analysis was restricted to marked roadways. It is currently uncertain how our results will be generalized to roads with poor lane markings.

A total of 326,238 lane change events were identified by the algorithm in 46,250 trips, totaling over 400,000 miles of driving. In addition, the breakdown of lane change frequency by speed bin also shows that drivers are more likely to change lanes in travel speeds ranging from 30-60 mph. The lower lane change frequency in the lower speed range bins can potentially be due to several reasons. First, it is likely that lane changes events in lower speed ranges (i.e. lower than 30 mph) were on roads with lower speed limit and were 2-lane roads with no adjacent lanes to change into. We also hypothesize that when drivers initiate lane change maneuvers in lower speed ranges, they were more likely to speed up and thus are grouped into higher speed bin ranges. An example scenario would be during congested traffic conditions in the Washington D.C. area, the instrumented vehicle are in a lane with slow moving traffic but speeds up to change into the adjacent lane in order to move ahead of traffic.

The distribution of TTC showed that minimum TTC, as well as variability of TTC between drivers, generally increased with travel speed. The increase in TTC can be attributed to drivers generally becoming more cautious and increase following distance as travel speed increases. As driver begins to increase following distance, lead vehicles are more likely to be outside of the range of the front radar. Although our study was highly dependent on the radar to detect lead vehicle, therefore biased towards what the radar can "see", the results are still relevant in improving the effectiveness of a FCW system in a crash imminent situation.

The characterization of lane change events presented in the current study will provide active safety system designers further understanding of driver action in overtaking maneuvers and can improve designs of FCW systems. Specifically, the results of this study show that the frequency and TTC in lane change events vary by vehicle speed. The large variability in minimum TTC between drivers shows the need for FCW systems to implement several different warning thresholds depending on different driving style.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Toyota Collaborative Safety Research Center (CSRC) and Toyota Motor Corporation for funding this study. Our special thanks to Rini Sherony, Hiroyuki Takahashi, and Satoshi Taniguchi of Toyota for sharing their technical insights and expertise throughout the project. We also gratefully acknowledge Stephen Hunter and Stephani Martinelli, our Virginia Tech undergraduate research assistants, for their greatly appreciated contributions to the search algorithm validation.

REFERENCES

[1] National Highway Traffic Safety Administration, "Traffic Safety Facts 2012: A Compilation of Motor Vehicle Crash Data From the Fatalit Analysis Reporting System and the General Estimation System." U.S. Department of Transportation, Washington, DC, Report no. DOT HS 812 032, 2014.

[2] K. D. Kusano and H. C. Gabler, "Safety Benefits of Forward Collision Warning, Brake Assist, and Autonomous Braking Systems in Rear-End Collisions," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1546–1555, Dec. 2012.

[3] J. McClafferty and J. Hankey, "100-Car Reanalysis : Summary of Primary and Secondary Driver Characteristics." The National Surface Transportation Safety Center for Excellence, Blacksburg, VA, p. Report Number 10–UT–007, 2010.

[4] T. Dingus, S. Klauer, V. Neale, A. Petersen, S. E. Lee, J. Sudweeks, M. Perez, J. Hankey, D. Ramsey, S. Gupta, C. Bucher, Z. . Doerzaph, J. Jermeland, and R. . Knipling, "The 100-Car Naturalistic Driving Study, Phase II – Results of the 100-Car Field Experiment," Report no. DOT HS 810 593, 2006.

[5] K. D. Kusano, J. Montgomery, and H. C. Gabler, "Methodology for identifying car following events from naturalistic data," *IEEE Intell. Veh. Symp.*, pp. 281–285, 2014.

Driver Vehicle Interface Design Assistance for Vehicle-to-Vehicle Technology Applications

Christian Jerome, Ph.D.
Chris Monk, Ph.D.
National Highway Traffic Safety Administration
U.S. Department of Transportation
United States

John Campbell, Ph.D.
Battelle
United States

Paper Number 15-0452

ABSTRACT

This paper provides an overview of the program of research involved in the development of a set of driver-vehicle interface (DVI) design assistance intended for vehicle-to-vehicle (V2V) technology applications. The research reviewed and performed under the NHTSA sponsored program of research (Human Factors for Connected Vehicles (HFCV), Safety Pilot DVI Evaluations, Driver Issues, and Integration Strategy) will culminate in a DVI Design Assistance Document for V2V technology applications. The results of the research will inform this document, which will be a set of Design Assistance topics and will describe appropriate human-centric design attributes for the DVI.

INTRODUCTION

Automotive DVI research in general has typically focused on the design of safety system DVIs. Therefore, the available research used to inform this DVI Design Assistance Document is primarily drawn from safety research. However, the basic design assistance that this document provides may help inform the design of non-safety related DVIs (i.e., infotainment and driver convenience systems) for V2V. Additionally, this document provides information from recent and ongoing research in the emerging field of V2V technology.

Objectives

The overall objective of the HFCV research program is for V2V technologies and applications to have DVIs that effectively communicate safety information while managing driver workload and minimizing distraction. This objective will be realized with the product of this research program, the DVI Design Assistance Document, and will aid in the design of relevant system interfaces to be more effective without increasing distraction or creating high workload.

Goals

The goal of the DVI Design Assistance Document is to aid the design of interfaces that are effective without increasing distraction**.** Warning system interfaces are effective if they (1) attract the driver's attention, (2) communicate the intended information, (3) communicate the information in a way that is immediately understandable so the driver can process the information, and (4) directs attention to a threat in time for the driver to take corrective action. The DVI Design Assistance Document were informed by previous research and current ongoing research performed in the course of the HFCV research program. The research program and design assistance provide information for a better understanding of driver behavior, improved driver assistance

technologies, interface design assistance for vehicle-to-vehicle communications technologies aimed at passenger vehicles, heavy trucks, and transit vehicles. The NHTSA's primary concern is safety applications, but the research and design assistance also cover some mobility and sustainability applications for V2V.

In order to realize these goals and objectives, the research program approach included a number of phases and simultaneous tracks that informed the DVI Design Assistance Document. The program began with a definition of the problem, which split into two main tracks of research: one focusing on V2V information and how it should be integrated and managed to better suit the driver's needs, and the other focused on characterisitcs of the DVI and how different safety messages should be delivered to the driver. The HFCV program also engaged stakeholder throughout the DVI Design Assistance development in order to improve the format and content of the DVI Design Assistance Document.

DVI DESIGN ASSISTANCE DOCUMENT

The topics within the V2V DVI Design Assistance Document are informed by the best available scientific research, and the DVI Design Assistance Document is a collection of focus areas uncovered by research that are intended to inform and assist DVI designers in making decisions regarding potential issues with various aspects of the DVI. Each focus area includes an introduction that first defines the topic and how it might be problematic, then the focus area summarizes the functional design goal with supporting design assistance. The focus area then includes a discussion of the practical design issues and the supporting available scientific research.Each focus area of the V2V DVI Design Assistance Document is organized into an easy-to-read, two-page format.

Overall, this document provides goals and assistance for the design and development of V2V DVIs, for both light- and heavy-vehicles, based on knowledge of driver capabilities and limitations. It consists of a series of chapters containing DVI design assistance. Each chapter provides a set of subtopics relevant to a specific design characteristic or element. Chapter 2 provides an overview of the format and content of these design-specific chapter topics (Chapters 3 through 11). Following the design chapters are a set of reference chapters with supplemental information (Chapters 12 through 18), including tutorials, a glossary, an index, lists of abbreviations and equations used in the document, a list of additional standards and other documents related to DVI design, and a complete reference list of articles and reports used to develop the design assistance and tutorials. Figure 1 below provides a high-level overview of the chapters and topics included in the DVI Design Assistance Document.

Chapter 1. Introduction
Chapter 2. How to Use this Document

Chapter 3. General DVI Considerations
- Distraction
- General Workload Considerations
- Workload from Secondary Tasks
- Providing Drivers with Information on System Function and System Messages
- Developing Driver Training Materials

Chapter 4. Principles for Safety Messages
- False and Nuisance Warnings
- Multimodal Warning Messages
- Warning Stages
- Providing Warnings that Accommodate Driver Brake Reaction Time
- Using Coverage Zones to Provide Lane Change Information

Chapter 5. Message Characteristics
- Designing Messages for Driver Comprehension
- Message Complexity
- Selection of Sensory Modality

Chapter 6. Visual Interfaces
- Visual Display Type for Safety-related Messages
- Locating a Visual Display
- Using Color
- Selecting Character Height for Icons and Text
- Characteristics of Legible Text
- Temporal Characteristics of Visual Displays
- Display Glare
- Head-up Displays

Chapter 7. Auditory Interfaces
- Auditory Display Type
- Perceived Urgency of Auditory Warnings
- Perceived Annoyance of Auditory Warnings
- Loudness of Auditory Warning Signals
- Distinctiveness of Warning Messages
- Using Localization Cues to Indicate Direction
- Presenting Warnings Using Speech Messages

Chapter 8. Haptic Interfaces
- Selecting a Haptic Display
- General Characteristics for Haptic Displays
- Improving Distinctiveness of Haptic Displays
- Accommodating for Vibrotactile Sensitivity Across the Body
- Generating a Detectable Signal in a Vibrotactile Seat
- Presenting Spatial Information Using a Vibrotactile Seat

Chapter 9. Driver Inputs
- General Principles for Driver-DVI Interactions
- Control Placement
- Selecting Physical Control Type
- Control-movement Compatibility
- Control Coding
- Labels for Controls
- Voice Recognition Inputs

Chapter 10. System Integration
- Prioritizing Messages Presented to Drivers
- Using "Master" Warnings in Integrated Warning Systems
- Overview of the HFCV Integration Architecture

Chapter 11. Application of HF Design Principles to Heavy Vehicle DVIs
- Design Considerations for Warning Signals in Heavy Vehicles
- Selection of Sensory Modality for Heavy Vehicle Warnings
- Design Principles for Visual Displays in Heavy Vehicles
- Visual Display Location in Heavy Vehicles
- Design Principles for Auditory Displays in Heavy Vehicles
- Design Principles for Haptic Displays in Heavy Vehicles
- Driver Controls for Collision Warning Systems in Heavy Vehicles
- General DVI Considerations for Heavy Vehicles

Chapter 12. Tutorials
- Tutorial 1: Procedures for Assessing Driver Performance: Visual Demand Measurements
- Tutorial 2: Priority Order Index Look-up Table for Message Prioritization
- Tutorial 3: Preliminary HFCV Integration Architecture
- Tutorial 4: Heavy Vehicles Characteristics and Driving Environment Relevant to DVI Design

Chapter 13. Glossary
Chapter 14. Index
Chapter 15. Abbreviations
Chapter 16. Equations
Chapter 17. Relevant Documents from USDOT, SAE International, ISO
Chapter 18. References

Figure 1. Overview of the Contents of the DVI Design Assistance Document.

In the V2V DVI Design Assistance Document, a consistent two-page format is used to present the individual human factors topics provided in Chapters 3 through 10. On each page, the chapter title is indicated by centered, bold type within the header. As described in more detail below, the left-hand page presents the title of the topic; an introduction and overview of the topic; a high-level design goal; supporting design assistance; and a graphic, table, or figure that augments the text information. The right-hand page provides the more detailed supporting rationale for the topic, as well as special design considerations, cross-references to related topics, and a list of references. A sample topic, with key features highlighted, is shown below in Figure 2:

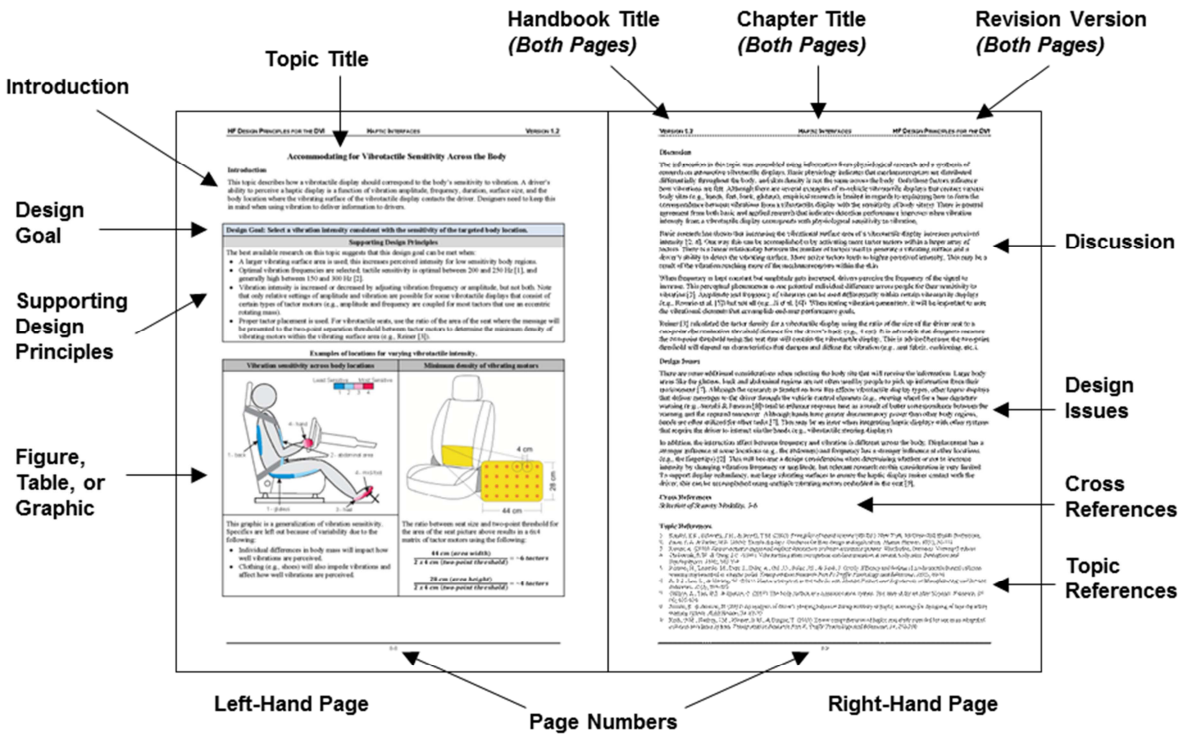

Figure 2. Topic format used in the design assistance document.

Overview of Example Topics

Below, we have summarized the contents from a small sample of the design topics addressed in the DVI Design Assistance Document:

- Locating a Visual Display,
- Perceived Urgency of Auditory Warnings,
- Accommodating for Vibrotactile Sensitivity Across the Body, and
- Selection of Sensory Modality for Heavy Vehicle Warnings

Only 4 of the 52 design topics were included in the interest of the limits of this publication, and these particular 4 topics were chosen due to the expected high interest in them (i.e., they are key topics of interest in current trends for DVI design).

Locating a Visual Display

This topic discusses the issues related to the location of a display where it will be easily seen and comprehended by drivers. As new warning and information systems on the market can increase dashboard complexity, selecting an appropriate location for a visual display is important to minimize information-seeking behavior that distracts from the driving task. Recommended design assistance to facilitate rapid extraction of information while minimizing eyes-off-road glances and negative impacts on driving performance are:

- Critical displays for continuous vehicle control or critical warnings related to vehicle forward path are located within ± 15 degrees of the central line of sight but as close to the central line of sight as practicable. ISO (1984) recommends that messages that require immediate detection be located within 5 degrees of the forward view when possible. (A).
- Displays are placed in locations that are generally compatible with established expectations or with location cues from other warnings, such as auditory or haptic (B).
- The display location is compatible with the desired response, such as a display in the mirror for alerts for looking to the blind spot; a Head-Up Display (HUD) that is used to direct attention to the forward view for critical warnings (C).
- The design and location minimizes glare from external sources or other displays in the vehicle (e.g., in the instrument panel or under a protective cover; D).

Figure 3 below illustrates potential visual display locations.

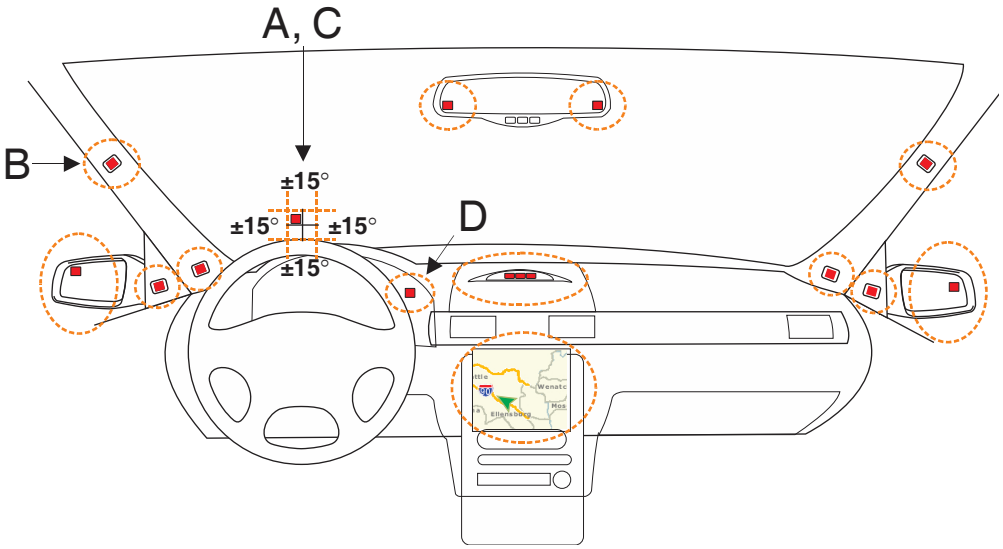

Figure 3. Potential Locations for Visual Displays

Additional discussion in this design topic includes research results and recommendations for keeping warnings near to the driver's line of sight, use of color, and correspondence between the driver's gaze and directional hazard location. Benefits of highly compatible display-response relationships include development of strong mental models by the operator as to how the system works, reduced driver response times, and fewer response errors. Also noted, is the lack of current data or standards regarding messages in heavy vehicle (HV) applications.

Perceived Urgency of Auditory Warnings

This topic discusses and recommends a number of design assistance topics to create auditory warnings that clearly communicate a level of urgency consistent with the urgency of the hazard. Auditory attributes that can be engineered to higher or lower levels of urgency include:

To increase the perceived urgency:

- Use faster auditory signals (e.g., 6 pulse/sec).
- Use regular rhythms (all pulses equally spaced).
- Use a greater number of pulse burst units (e.g., 4 units).
- Use auditory signals that speed up.
- Use high fundamental frequencies (e.g., ≥1000 Hz).
- Use random or irregular overtones.
- Use a large pitch range (e.g., 9 semitones).
- Use a random pitch contour.
- Use an atonal musical structure (random sequence of pulses).
- Use more urgent words (e.g., "Danger").

To decrease the perceived urgency:

- Use slower auditory signals (e.g., 1.5 pulse/sec).
- Use irregular rhythms (pulses not equally spaced).
- Use a fewer number of pulse burst units (e.g., 1 unit).
- Use auditory signals that slow down.
- Use low fundamental frequencies (e.g., 200 Hz).
- Use a regular harmonic series.
- Use a small pitch range (3 semitones).
- Use a down or up pitch contour.
- Use a resolved musical structure (from natural scales).
- Use less urgent words (e.g., "Caution").

Figure 4 below provides a visual comparison of higher urgency auditory signal characteristics (tops) with lower urgency signal characteristics (bottom).

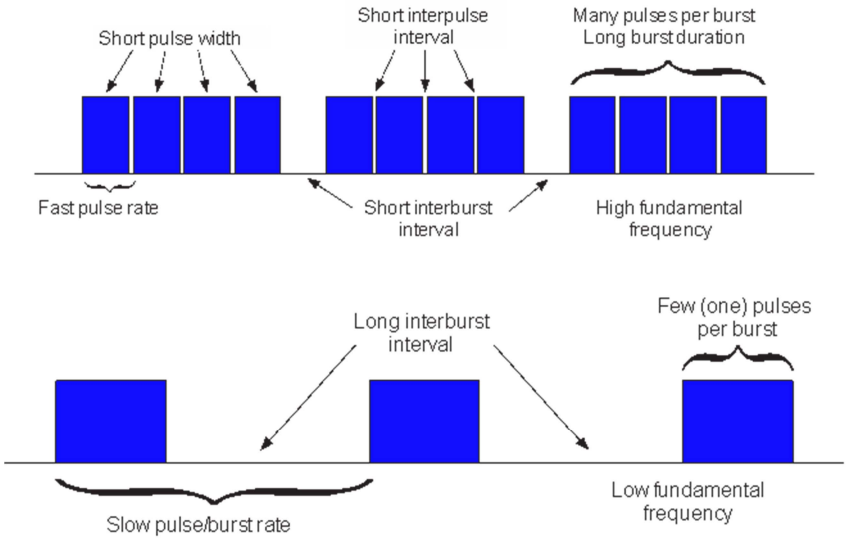

Figure 4. Comparison of Auditory Signal Characteristics Associated with Higher vs. Lower Levels of Perceived Urgency Characteristics

The discussion for this topic includes research results and recommendations indicating that accurate urgency mapping leads to more effective driver response. While more urgent signals are associated with faster reaction

times, confusion, startle responses, or distraction can result if the perceived urgency is higher than the situation demands.

Some of the design trade-offs to be considered include compatibility with other signals, driver annoyance, and distinctiveness of the signal. Message semantics for both signals and speech can affect driver response. Study findings suggest that perceived urgency can be increased using high-urgency semantics (whether with speech or with familiar non-speech signals) into auditory messages that have lower-urgency acoustic characteristics (Guilluame et al. 2002; Baldwin & May, 2011).

Accommodating for Vibrotactile Sensitivity Across the Body

This topic provides design assistance for how to design vibrotactile displays so that they correspond to the body's sensitivity to vibration. Vibrotactile display attributes include vibration amplitude, frequency, duration, surface size, and the body location of the vibrating surface. Displays are most effective when they are readily perceived by the driver thus matching vibrotactile intensity with the differing sensitivity of various body locations will increase detection.

Figure 5 below provides examples of body locations for varying vibrotactile intensity.

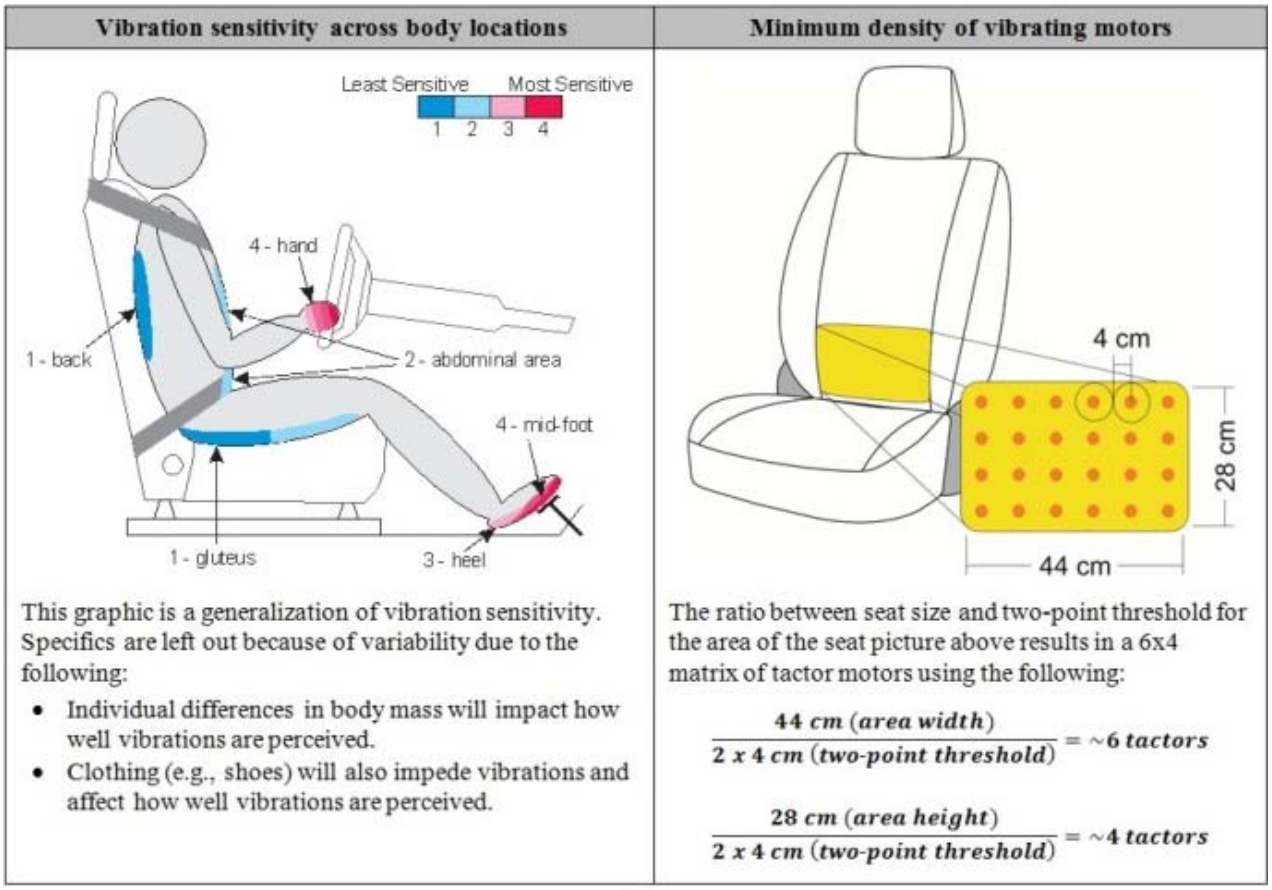

Figure 5. General Design Assistance for Vibrotactile Displays

Information on this topic is derived from research on both physiological responses to vibration and vibrotactile message delivery to drivers. As mechanoreceptors are distributed differentially throughout the body, and skin density is not the same across the body, vibrations are felt differently at different body sites. Both basic research and applied research indicate detection of the vibrotactile signal is improved when the vibration intensity corresponds with the physiological sensitivity to vibration.

Generally, drivers perceive increased intensity by either: 1) increasing surface area or number of tactors activated within the surface area, or 2) increasing the amplitude of a constant frequency. A two-point discrimination threshold can be used. Prototype testing using the exact seat to be used with the device is recommended as the threshold will be affected by the fabric, cushioning and other characteristics of the seat.

Additional design issues for consideration include driver familiarity, i.e., most drivers are not used to receiving information via vibrotactile sensation to large body areas such as the gluteus. Using the steering wheel to deliver a haptic cue for lane change warning, for example, would obtain a quicker response due to the correspondence between the warning and the corrective action. Consider also that signals to the hands, while effective sensation-wise, may be less effective when other driver activities using the hands interfere with reception of the signal. Display redundancy can be supported by using larger vibrating surfaces such as the seat.

Selection of Sensory Modality for Heavy Vehicle Warnings

This topic discusses how to select an appropriate presentation, such as visual, auditory, or haptic, for use in a heavy vehicle (HV) environment. The nature of heavy vehicle driving creates additional issues to be considered in light of higher visual scanning demands, the long-term exposure to alerts, and passenger needs. The selected DVI modality should be consistent with HV driver tasks, needs, and expectations. Table 1 below provides a summary of the recommendations for CWS functions and control types. In addition:

- Avoid using exclusively visual warnings for imminent collision warnings.
- Use auditory or haptic signals as the primary mode of conveying collision warning information. Use an auditory or haptic signal in conjunction with a visual display to increase warning conspicuity. Auditory signals have been shown to provide effective cautionary and imminent warnings, particularly as part of multimodal warnings.
- The use of haptic signals may be preferred over auditory signals in transit bus applications because they are less obtrusive and less likely to be noticed by passengers. However, haptic signals have additional considerations to ensure they are perceived. Design Assistance for Haptic Displays in Heavy Vehicles in the DVI Design Assistance Document provides issues, caveats, and recommendations associated with using haptic signals in heavy vehicles.

Table 1.
Recommendations and Design Assistance for Collision Warning Systems (CWS) in the Heavy Vehicle Cab

CWS Function	Recommendation	Use Discrete Control	Use Continuous Control
On/Off Enables and disables the CWS. *Note:* Drivers should be notified of system status for automatic system startup (i.e., ignition activated systems).	Not Recommended	Yes	Not Applicable
Sensitivity (Warning Timing, Warning Threshold, Range, TTC) Controls the physical or temporal proximity threshold for which warnings are activated.	Neutral Recommendation (i.e., can use if desired). Limited Range of Settings[1]	Yes Between 2 and 6 Sensitivity Settings	Yes Precise Adjustment
Master Intensity Master control for intensity of all displays within a modality (i.e., visual, auditory, or haptic). May include non-warning displays (e.g., IP brightness).	Recommended Limited Range of Settings[2]	Yes Multi-position	Yes Limited Range
Auditory Intensity Controls the intensity of the auditory warning signals.	Recommended Limited Range of Settings[2]	Yes Multi-position	Yes Limited Range
Visual Luminance Controls the intensity of the visual warning signals.	Recommended Limited Range of Settings[2]	Yes Multi-position	Yes Limited Range

Most current passenger vehicle CWSs allow disablement of the system by drivers; some current HV CWSs do as well. Disablement of the system, however, defeats the purpose of the fleet owner/operator. Additionally, some research notes that most drivers would not disable the system and prefer an integrated CWS system.

Allowing adjustments to CWS systems can accommodate both owner and driver needs. For example, reducing system sensitivity in a cluttered driving environment would correspondingly reduce the frequency of nuisance alarms, although this may mean a delayed alert which would not provide enough time for response. Another example is the decibel level of alarms, which can cause driver annoyance if set too loud. While research indicates that, for a forward imminent-only crash warning, 87dBA was rated appropriately high by 81% of drivers, 15% did not consider it loud enough (Tidwell et al., in press).

These design assistance topics recommend that HV system sensitivity could be reduced temporarily by drivers to mitigate nuisance/false alarms, through duration and frequency of reduction limits, and based on a standard sensitivity setting using empirical analysis of driver response times and vehicle stopping distances. Minimum audible and luminous settings could employ adaptive capability to accommodate varying levels of ambient noise and nighttime or low luminance conditions.

Involvement of the DVI Design Community

Stakeholder Groups The process for developing the DVI Design Assistance Document included participation of the DVI design community through two stakeholder groups: one for light vehicles and one for heavy vehicles. The stakeholder groups included individuals from OEMs, suppliers, and other industry companies and organizations.

Information Gathering and Sharing When each new organization joins a stakeholder group, the first activity they participate in is a User Requirements Analysis. This activity provides: 1) an opportunity to get a better understanding of the organization's needs as end-users, 2) information on how the DVI Design Assistance Document might fit into the design process at the organization, and 3) a first opportunity for an organization to provide feedback on the general content, format, and organization of the document. Other means of communicating and working with stakeholders was through webinars, workshops, and group or individual interviews.

Feedback and input from the community was used to:

- Support development of the content, format, and organization of the DVI Design Assistance,
- Provide insight into the DVI design process at their organization,
- Evaluate preliminary concepts and plans for the materials,
- Suggest additional research studies and source materials, and
- Review and provide comments on draft materials.

Key questions presented to the stakeholders covered all aspects of the final DVI Design Assistance, were organized around four main topics. Below each of the following topics include some of the main issues discussed:

- What content would you find useful?
    - Key needs for human factors information across a range of in-vehicle issues and applications.
    - Requests for design assistance on specific topics.
    - Data sources that we should use to develop the design assistance.
    - When, where, how, and how often the design assistance would be used.
    - Degree to which initial ideas for content meet the information needs of end-users.
- What format would work best?
    - Overall presentation of the V2V DVI Design Assistance Document – how to best achieve clarity, relevance & ease-of-use.
    - Layout of individual sections or topics.
    - Amount of text vs. non-text elements.
    - Flow and structure of text portions.
    - Use of graphics, figures, and tables.
    - Length and consistency of individual sections or elements.
    - Presentation of constraints and trade-offs.
    - Presentation of supporting materials.
- How would you like to see the information organized?
    - End-user ideas and preferences for overall organization of the materials.
    - Degree to which final organization of the V2V DVI Design Assistance Document:
        - Matches the design process.
        - Matches information needs.
        - Facilitates understanding of the design assistance.
        - Facilitates overall ease-of-use.
- When and how would this information be used within your organization?
    - During initial specification of the driver-vehicle interface
    - During testing and evaluation
    - During a final quality check

CONCLUSIONS

The V2V DVI Design Assistance Document provides a set of 52 human factors design assistance for driver-vehicle interfaces (DVIs) of systems for V2V communication. The design assistance are based on the findings of current high-quality research (including both the best-available scientific literature, and current research being conducted by agencies of the United States Department of Transportation), as well as basic human factors concepts. These design assistance are provided as a tool and a complementary resource to other documents and resources, as well as an expansion to industry research and existing guidance from the National Highway Traffic Safety Administration. The information within this document may be useful to researchers, designers, and original equipment manufacturers and Tier-1 suppliers seeking to ensure that V2V DVIs are designed to better reflect driver limitations and capabilities.

Discussion and Limitations

While DVI design assistance topics such as these can be a valuable tool and resource for designers, they are not without limitations. Many factors must be considered, and tradeoffs examined, prior to finalizing a DVI design. Some of these factors include regulation and industry or international standards. These design assistance topics are intended to support developers as they design DVIs for V2V communication systems. Note that this document is not meant to serve as a standard or design guideline. Resources such as Federal Motor Vehicle Safety Standards (FMVSS), SAE and ISO standards, and the Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices (NHTSA, 2013) exist that provide design guidance for DVIs. Instead, this document it is meant to serve as a complementary resource for original equipment manufacturers (OEMs), Tier-1 suppliers, and the automotive research community in designing V2V DVIs that enable rapid, consistent, and reliable communication between the vehicle and driver.

REFERENCES

Baldwin, C. L. & May, J. F. (2011). Loudness interacts with semantics in auditory warnings to impact rear-end collisions. *Transportation Research Part F: Traffic Psychology and Behaviour, 14*(1), 36-42.

Guilluame, A., Drake, C., Rivenez, M., Pellieux, L., & Chastres, V. (2002). Perception of urgency and alarm design. *Proceedings of the 8th International Conference on Auditory Display*.

International Organization for Standardization (ISO) (1984). *Development and principles for application of public information symbols.* (ISO/TR 7239). Geneva: Author.

Tidwell, S.A., Blanco, M., Trimble, T., Atwood, J., & Morgan, J.F. (in press). *Evaluation of heavy vehicle crash warning interfaces* (Report No. DOT HS XXX XXX). Washington, DC: National Highway Traffic Safety Administration.

AUDITORY ALERTS IN VEHICLES: EFFECTS OF ALERT CHARACTERISTICS AND AMBIENT NOISE CONDITIONS ON PERCEIVED MEANING AND DETECTABILITY

Jeremiah Singer
Neil Lerner
Westat
USA

Carryl Baldwin
George Mason University
USA

Eric Traube
National Highway Traffic Safety Administration
USA

Paper Number 15-0455

ABSTRACT

Two complementary studies were conducted. In Study 1 (categorical perception of alerts), a series of experiments examined the key parameters that result in listeners perceiving a given sound as an urgent warning versus other less urgent message categories. Initial perceptual sorting experiments identified the most significant parameters and subsequent driving simulator experiments confirmed and refined the findings. Study 1 found that four auditory characteristics accounted for most of the variance in classification of auditory alerts as urgent warnings. Sounds were classified as alarms over 90 percent of the time when they had a peak-to-total-time-ratio (ratio of the time that the sound is at full intensity to the entire pulse duration including onset and offset) of greater than or equal to .7, an interburst interval of less than or equal to 125 ms, at least 3 harmonics, and a base frequency of greater than or equal to 1000 Hz. These results were observed initially in laboratory studies, and replicated during simulated driving. In Study 2 (warning perception in ambient noise environments), an experiment was conducted to investigate the effects of different in-vehicle ambient noise conditions on auditory signal detection and perception. Participants driving on a freeway experienced three ambient noise conditions (windows closed and no music, music playing, front windows open). Fifteen auditory alerts, presented at 65 or 75 decibels (A-weighted), occasionally occurred. Participants pressed a button as soon as they detected the sound, then provided ratings of the noticeability, urgency, and intended meaning of the sound. Study 2 found significant main effects for ambient noise condition and for alert sound for perceived noticeability, urgency, and response time to alert. Detection was impaired by the presence of music, and even more so with the front windows open. Even when auditory signals were heard, noise conditions modified their perceived urgency and meaning. There were also interactions between ambient noise condition and sound, indicating differences in how well sounds of similar loudness tolerated interference from noise. Results also demonstrate that the perceived urgency and meaning of auditory messages can change under noisier ambient conditions, and some features of more noise-resistant signals were suggested by the data. The findings of this research may help interface designers to create auditory signals that indicate the appropriate type and urgency of message.

INTRODUCTION

Advanced technologies in vehicles and on mobile devices when used to complete the driving task such as smartphones are expanding the amount of informtion available to drivers. Information is increasingly available about safety-critical events, vehicle status, traffic and navigation, and so forth. In this increasingly information-dense environment, it is important that drivers can quickly and easily recognize and distingish time-critical safety information from other, less urgent information.

This paper describes two complementary studies conducted within the National Highway Traffic Safety Administration's (NHTSA) Crash Warning Interface Metrics (CWIM) research program. The CWIM program included a series of research studies intended to address research needs related to the driver-vehicle interface for advanced crash warning systems. Study 1 (Categorical Perception of Alerts) included a series of research tasks designed to establish criteria regarding the key characteristics of a warning sound that would enable the sound to be quickly and reliably perceived as representing a highly urgent collision warning. Study 2 (Warning Perception in Ambient Noise Environments) investigated the effects of different in-vehicle ambient noise conditions on auditory signal detection and perception. Together, these two studies provide insights into the criteria required for an acoustic

signal to be unambiguously interpreted as an urgent warning with the consideration of the effects of varied in-vehicle ambient noise conditions on the detection and interpretation of signals.

STUDY 1: CATEGORICAL PERCEPTION OF ALERTS

This study, which was led by George Mason University (GMU), included a series of research tasks designed to establish criteria regarding the key characteristics of a warning sound that enable the sound to be quickly and reliably categorized as an urgent warning rather than a vehicle status message or a social notification. Three methodologies were implemented across six research tasks to specifically examine the range of key parameters that result in 85 percent of listeners perceiving a given sound as an urgent collision warning. The 85-percent criterion was selected because it is a common de facto criterion to establish a boundary of typical behavior or perception.

Before beginning the experiments, the research team conducted a survey of forward collision warning (FCW) sounds currently in use in automobiles. The objective of this task was to collect an inventory of warning sounds that could be used to develop an initial list of characteristics common among acoustic warnings. A focus was placed on obtaining representative OEM FCW sounds (i.e., those with the greatest vehicle fleet penetration). Once obtained, the sounds were analyzed in terms of their acoustic properties (e.g., spectral frequency components, pulse duration, onset and offset, harmonics, tempo, etc.). From this inventory, as well as existing published guidelines (e.g., Campbell, Richard, Brown, & McCallum, 2007), a list of key parameters was developed and examined in subsequent experiments. Select non-warning sounds (e.g., vehicle status notifications, ringtones) were also surveyed and recreated for use in this research.

The first four research tasks in this study were a sort task and a series of three method-of-adjustment laboratory tasks. These studies were designed to define and refine the key acoustic characteristics that cause individuals to categorize a message as an urgent crash warning as opposed to other less urgent categories, using the characteristics identified in the inventory of existing in-vehicle warning sounds as a starting point. Specifically, the research team sought to define the key parameters that would be associated with at least 85 percent of listeners perceiving the sounds as a highly urgent, time-critical collision warning. The final two tasks validated the findings of the first four tasks using a sort task and a driving simulation task, respectively. Methods and results for each of the six research tasks are briefly discussed below.

Perceptual Space Sort Task

Introduction The goal of this series was to define the perceptual space associated with a highly urgent FCW sound. Specifically, the research team sought to define the key parameters that would be associated with at least 85 percent of listeners perceiving the sounds as a highly urgent, time-critical collision warning.

Method Twenty-one GMU students participated in this research task. Stimuli initially included in the Perceptual Space Sort Task were 29 varying driver-vehicle interface (DVI)-related sounds. Sounds were created in Adobe Audition CS5.5 and Audacity and were based closely on sounds used in real, in-vehicle systems. Stimuli included sounds from many categories of in-vehicle system including those similar to FCW, LDW, backup warning, fatigue alerts, door open, low fuel, seatbelt, and park assist sounds. Stimuli were equated for loudness using Adobe Audition and were presented at 84 decibels, A-weighted (dBA).

Sounds were embedded in a large Microsoft PowerPoint slide such that they could be played by clicking on the sound and a pressing a small play triangle. After a participant played a sound, he or she could move it using the cursor into one of four categories. The available categories were “Warnings,” “Alerts,” “Status Notifications,” and “Social Notifications.” Each category also held a small section labeled “Prototype.” Under each category were two text boxes. The first text box held the text “Click here to tell us why…” and was intended to allow participants to give justification for how or why they grouped sounds together. The second box held the text “Best Urgency Level:” and was intended to allow participants to give a numerical rating of the desired urgency level for each category between 1 and 100.

Participants were instructed to play each sound and then categorize it into one of the four available categories. “Warning” sounds were suggested to “include sounds that you believe to be time-critical, collision warning sounds.” “Alert” sounds were suggested to “include non-critical alerting sounds, for example, a sound that might indicate a

lane departure or a parking assist sound." "Status Notifications" were suggested to "include sounds that indicate something about the status of your car, for example, low windshield wiper fluid or low tire pressure." "Social Notifications" were suggested to "include sounds used by a car's social media system to indicate a social media (like Facebook or email) notification." Participants were instructed that they could have as many sounds as they desired in each category but must have at least one sound in every category. Participants were further instructed that after categorizing each sound, they should play every sound again and choose one sound that was the most prototypical of each category and move it into the "Prototype" section. Participants were finally instructed that they should then attempt to explain their grouping choices and choose an urgency level for each category between 1 and 100. After the practice task, participants were given the experimental task which was identical to the practice but with more sounds.

Results Results were first analyzed only in terms of the percentage of times each sound was placed in a given category. While participants distinguished between warnings and alerts as opposed to status or social notifications, there was little agreement on what constituted a warning or an alert. This was further made obvious based on explanations for categorization where it was possible for two participants to categorize based on the same criteria but to choose opposite categories for which those criteria apply. Therefore, all "warning" and "alert" categorizations were reclassified into a new category called "alarms" after data collection.

The four most important (i.e., explaining the most variance) properties that related most to alarm-like warnings were determined using a backward stepwise logistic regression to predict membership in the category "alarm." These properties were interburst interval (IBI), base spectral frequency, number of harmonics (which contributed to the harshness of the tone) and peak to total time ratio. These properties explained 58 percent of the variance in alarm classification ($R = .800$, adj. $R^2 = .581$). Peak to total time ratio defines a property to encompass the perceptions created by longer or shorter onset or offset values. IBI is the gap between multiple bursts of sound, which contributes to the perceived tempo of a sound.

Using the results of the backward stepwise regression, criteria were defined in order of decreasing importance for each property of interest and cutoffs were determined. Cutoffs allowed the researchers to organize the data in terms of whether or not it met each criterion sequentially, and eliminate any sound that did not pass at each level. Cutoffs in this case were based partially on results from the subsequent Method of Adjustment experiments and partially based on the research team's experiences from previous studies and examples from the literature. Sounds that met all four of the criteria below were classified as alarms by at least 90 percent of the listeners. The criteria identified were:

- Peak to total time ratio $\geq$ .7
- Interburst interval (IBI) $\leq$ 125 ms
- Number of harmonics $\geq$ 3
- Base frequency $\geq$ 1000 Hertz (Hz)

Method of Adjustment – Single Parameter

Introduction The primary purpose of this experiment was to see at what point participants perceived an auditory tone to either be a highly urgent time-critical collision warning sound or to no longer be a highly urgent time-critical collision warning sound. A psychophysical method of adjustment procedure was implemented using ascending and descending thresholds for the key parameters. Participants began with a high or low parameter value and then adjusted it up or down, respectively, until the parameter fell within what they perceived to be a level consistent with a critical warning. For this experiment, frequency, IBI, pulse duration, and pulses per burst were examined one at a time. Averages of the crossover points using an ascending and descending series of parameter adjustments could then be used to further define the perceptual space. Furthermore, these results could be compared to the parameters obtained from the previously described Sort Task and any similar parameter values for the cutoffs would validate the Sort Task results.

Method Twenty GMU students participated in this experiment. The base sound used for this experiment was a single 300 Hz tone lasting 200 ms with a 20 ms onset and offset. The tone was repeated eight times with 118 ms between each burst, lasting for a total time of 2426 ms. The sound was played through Sennheiser headphones via a Matlab program. The Matlab program allowed participants to adjust the sound by frequency, pulse duration, tempo

(IBI), or pulses per burst. Sounds were presented at 84 dBA. All parameters included a minimum and maximum possible value with a set increment by which the parameter could be adjusted. For this experiment participants only adjusted one parameter at a time. Each screen indicated which parameter would be adjusted and whether the parameter should be adjusted to sound like a highly urgent, time-critical collision warning or no longer like a highly urgent, time-critical collision warning. Participants adjusted each parameter six times. In a randomized order participants adjusted each parameter three times in ascending order to make it sound like a highly urgent, time-critical collision warning sound and three times in descending order to make it no longer sound like a highly urgent, time-critical collision warning sound. Participants received real-time feedback, such that each time an adjustment was made, a new sound was played based upon the adjustment. Participants could continue to make adjustments until they were satisfied that the sound sounded like or no longer sounded like a highly urgent, time-critical collision warning.

Results Table 1 shows the mean and 85th percentile values at which at least 85 percent of participants found the sound to be a highly urgent, time-critical warning.

Table 1.
Mean values for parameters and value in which at least 85 percent of all participants found the sound to be a highly urgent, time-critical warning

Parameter	Mean Value	85% Value
Frequency	931.71 Hz	1200 Hz
Tempo (IBI)	330 ms	240 ms
Pulse Duration	460 ms	360 ms
Pulse Per Burst	2.73	4

Method of Adjustment – Multiple Parameters

Introduction This experiment was conducted to explore the possibility that some acoustic parameters may play a stronger role than others in shaping the perceptual space. This experiment allowed individuals to adjust all four parameters simultaneously, allowing them to change the parameter they felt was more important and thus dictating their perception. Furthermore, this experiment makes it possible for the participant to always make a sound seem like or unlike a highly urgent, time-critical collision warning sound, where in the single parameter adjustment it was possible that someone may not have been able to effectively create a highly urgent sound by changing levels of only one parameter. For example, it is possible that participants could not adjust the sound enough to perceive it as being either a highly urgent, time-critical collision warning sound or no longer being a highly urgent, time-critical collision warning sound..

Method Twenty GMU students participated in this experiment. For this experiment the same base sound from the Single Adjustment Task was used in addition to two DVI-related FCW sounds. Sounds were played through Sennheiser headphones via a Matlab program. The Matlab program allowed participants to adjust the sound on frequency, pulse duration, tempo (IBI), and pulses per burst simultaneously (by the same values as presented for the Single Adjustment Experiment). Sounds were presented at approximately 78 dBA. This sound pressure level was 6 dBA lower than used in the previous Method of Adjustment study. This change was made to reduce signal loudness to a level closer to what might be found in vehicles.

The procedure was exactly the same as the Single Adjustment Task with the exception that participants in this experiment were asked to adjust each of the four parameters for a single sound. Participants were allowed to adjust all four parameters at the same time, allowing them to freely choose which parameter to adjust first and allowed them to continually switch between parameters until they were satisfied with the resulting sound. As in the previous adjustment study, participants were instructed that their task was to make the sound presented seem either like a highly urgent, time-critical collision warning (in the ascending trials) or to no longer sound like a highly urgent, time-critical collision warning (in the descending trials).

Results Table 2 shows the average value for each parameter as well the value at which at least 85 percent of participants found the sound to be a highly urgent, time-critical collision warning. The table shows that the values were similar to those found in the previous method of adjustment experiment.

Table 2.
Average values for parameters and value in which at least 85 percent of all participants found the sound to be a highly urgent, time-critical collision warning

Parameter	Average Value	85% Value
Frequency	1094.17 Hz	1300 Hz
Tempo (IBI)	500 ms	360 ms
Pulse Duration	510 ms	440 ms
Pulse Per Burst	3.03	4

Method of Adjustment – Prototype Development

Introduction The primary purpose of this experiment was to gain information regarding what values of each of the key parameters would result in individual perceptions of that sound being the "ideal or prototypical" collision warning sound.

Method Twenty-five GMU students participated in this experiment. For this experiment the same base sound from the "Method of Adjustment – Single Parameter" task was used. The sound was played through Sennheiser headphones via a Matlab program. The Matlab program allowed participants to adjust the sound on frequency, pulse duration, tempo (IBI), and pulses per burst simultaneously. Sounds were presented at approximately 78 dBA.

The procedure was the same as the Multiple Adjustment Task with the exception that participants were now asked to adjust the sound until it matched their prototype of the ideal highly urgent, time-critical collision warning sound.

Results Table 3 shows the average value for each parameter as well as the value at which at least 85 percent of participants found the sound to be a highly urgent, time-critical collision warning. Compared to the results of the previous method of adjustment task, these results show that an "ideal" collision warning sound has a higher frequency, faster tempo, and faster pace than a "threshold" collision warning.

Table 3.
Average values for parameters and value in which at least 85 percent of all participants found the sound to be a highly urgent, time-critical collision warning

Parameter	Average Value	85% Value
Frequency	1576 Hz	1900 Hz
Tempo (IBI)	254 ms	40 ms
Pulse Duration	398 ms	200 ms
Pulse Per Burst	4.16	5

Validation Sort Task

Introduction The primary purpose of the Validation Sort Task was to validate the results of the initial sort task and the method of adjustment tasks. To do so, researchers chose to investigate parameters that were found to be important in the previous tasks. Since only one of the previously-evaluated sounds met all four of the criteria established in the first Sort Task, the researchers directly manipulated key acoustic aspects of several sounds to create new sounds designed to examine the robustness of criteria cutoff values. Specifically, the number of harmonics present in several sounds was manipulated by adding harmonic components to sounds that previously had only one harmonic. The presence of multiple harmonics was found in the previous tasks to influence the "harshness"

of the sound, and indirectly, perceived urgency or importance. Further, it was found in the first sort task that some sounds were inadvertently disproportionately long. Therefore in this validation task, sounds were recreated to be of more similar lengths; approximately 1500 ms.

Method Fifteen GMU students participated in this experiment. Stimuli for the Validation Sort Task were similar to the Perceptual Space Sort Task but included additional stimuli using added harmonics such that 52 total stimuli were presented to participants. All apparatus and presentation methods were identical to the Perceptual Space Sort Task discussed above.

Results Results were analyzed identically to the Perceptual Space Sort Task, using the same criteria and cutoffs. A backwards stepwise regression was able to account for 61 percent of the variance in Alarm classification ($R=.802$, adjusted $R^2=.612$). Sounds that met all criteria (or even criteria after the first cutoff) were much more likely to be considered "alarm" sounds than status or social sounds.

Rapid Categorization under Divided Attention

Introduction The Rapid Categorization under Divided Attention task was designed to extend and provide additional validation of the results from the earlier tasks in this series. It has previously been shown that cognitive load can degrade participants' responses to warnings while driving (Lewis, Penaranda, Roberts, & Baldwin, 2013; Santangelo & Spence, 2007). Therefore, this task was designed to examine whether or not people's rapid categorization of sounds into critical and noncritical categories differed when they were experiencing the multiple demands of simulated driving while being engaged in a distracting secondary task. In this experiment, participants first completed a sort task similar to the two reported earlier in this study. Then they were asked to categorize sounds by providing a behavioral response (either a brake or button press) while they were concurrently engaged in both a simulated driving task and an n-back number recall task. The n-back task is described in the Method section.

Method Twenty-two GMU students participated in this experiment. For the sort task, 26 sounds were used, with specific emphasis placed on those sounds that were shown to be consistently categorized across the sort tasks. The sort task was similar to the two previous sort task experiments with some small changes. In this case the separate categories of "warning" and "alert" were merged into one single "alarm" category, as previous experiments indicated that there is little consistent differentiation between "warnings" and "alerts" by sound parameters. The prototype section of the experiment was removed; otherwise, all aspects of the sort task were identical to the previous sort tasks.

Upon completion of the sort task, participants were seated in the driving simulator. A Realtime Technologies desktop driving simulator was used, with a Logitech Driving Force GT steering wheel. Ambient sound was presented via two monaural computer speakers at 60 dBA. Sounds were presented via a pair of Bose computer speakers at 75 dBA. The same sounds were used as in the sort task. All sounds were repeated twice except for some of the social sounds, which were repeated four times to increase the temporal gap between more urgent sounds. This set of sounds included several sounds based on criteria that were earlier established and were expected to consistently be classified as alarms. Additionally, several sounds were used that shared several but not all attributes of an alarm. The research team theorized that these sounds would be more difficult to classify and potentially result in not being classified as an alarm and/or requiring more time to make a classification response.

Participants practiced driving while maintaining a speed of 45 mph and holding a steady lane position. Once participants were comfortable with the simulator they were instructed that they would be doing a visual 1-back task. For this task, numbers between 1 and 10 were presented on a small screen to the right of the steering wheel. Participants responded via the steering wheel positively or negatively depending on whether or not each number matched the number presented directly before it. After a variable amount of time on the 1-back task, a sound played and participants indicated via the brake, left button, or plus button whether they considered the sound to be an alarm, a status notification, or a social notification, respectively. For the initial practice task, instead of actual signals, the experimenter spoke the words "alarm", "status" or "social." Participants then repeated the category name that they heard. This allwed participants to become familiar with the response actions without exposing them to actual signals or requiring them to make a decision about the categorization of a signal. Midway through the practice, participants began driving and practiced all three tasks simultaneously. After completing the practice task, participants completed the experimental task. The experimental task was similar to the practice task with the main difference

being that sounds were now abstract tones (drawn from sounds used in the sort task experiments). Participants were instructed to make their classifications as quickly as possible while maintaining control of the vehicle.

Results Results from the sort task, using the same cutoffs and criteria as in the Perceptual Space Sort Task and Validation Sort Task, indicate that the sounds from Rapid Categorization Under Divided Attention are categorized similarly to previous experiments and that the criteria and cutoff results are highly predictive of alarm categorization (R=.865, adj. R^2=.701). If the abstract sound met the first two of the previously-established criteria (peak to total time ratio > .7; IBI < 125 ms) it was classified as an alarm at least 70 percent of the time. Sounds that met all four of the previously-established criteria (ratio and IBI, plus number of harmonics > 3; base frequency > 1000 Hz) were classified as alarms nearly 90 percent of the time. Note that these results are highly similar to the results of the previous two sort tasks. This provides additional support for the reliability and validity of the previous results. The next aim of the current study was to determine if the previously-established alarm criteria were also predictive of alarm perception under divided attention rapid categorization conditions.

Figure 1 shows the results of the rapid categorization task while participants were engaged in the driving and 1-back distraction task. Results are remarkably similar to those obtained from the sort task without concurrent load. Alarm categorization in the sort task obtained while driving was highly correlated with alarm categorization during the rapid categorization task (Pearson correlation=0.962, p<0.001). Additionally, status and social categorization during the sort task were highly correlated with status and social categorization during the rapid categorization task (Pearson correlation for status categorization =0.934, p<0.001 and for social categorization =0.982, p<0.001).

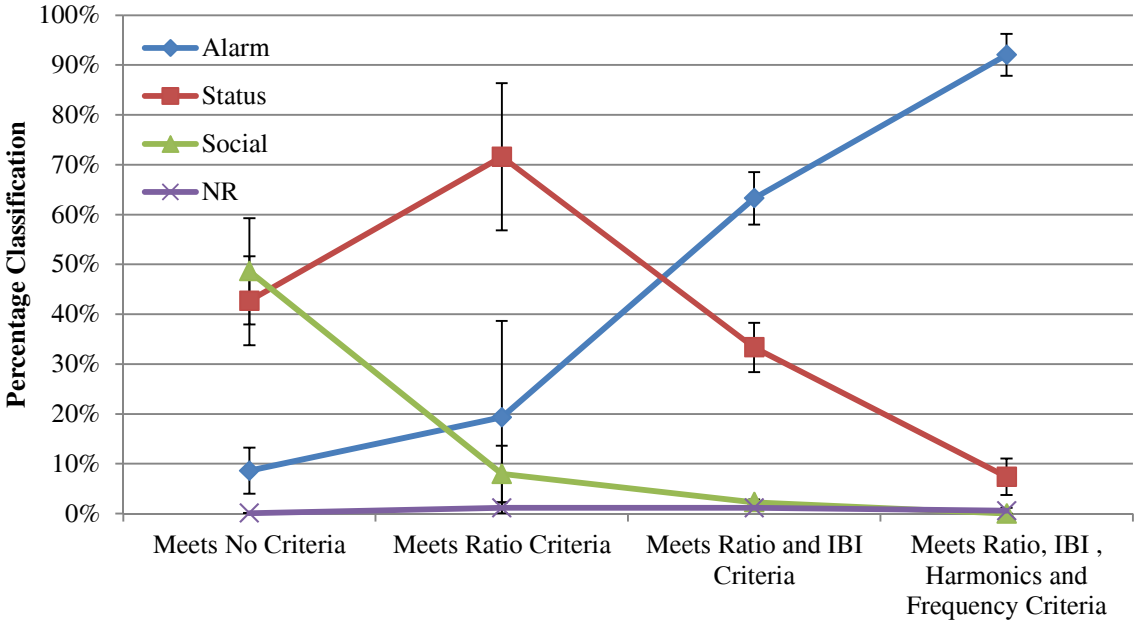

Figure 1. Percent categorization while driving by criteria met during driving and distraction task performance.

Results indicate that the number of criteria met was indicative of response times, where the fastest response times were for sounds that met all four of the previously-established alarm criteria, and slowest response times were for sounds that met only the ratio criterion. These results show that when a sound matches with the criteria, people are able to decide more quickly whether or not the sound is a critical alarm relative to a "status" or "social" sound.

Additionally, the more likely a sound was to be classified as an alarm, the faster the response time for that alert, with alarm classification being significantly predictive of lower response times, B = -0.398, p=0.004. Further, alarm classification even in the sort task was significantly predictive of faster response times during the driving task, B = -0.296, p=0.031. These results indicate that the more homogenous the classification (the more participants classify a sound as an alarm), the faster participants are able to respond to that sound. There was a potential confound, however, because participants categorized sounds by pressing a different button on a steering wheel for

each category. Therefore, differences in button response times could have potentially influenced results, though there is no direct evidence that this was the case.

Study 1 Discussion

The primary aim of this series of perceptual space investigations was to determine the key acoustic characteristics and the parameter values associated with sounds being unambiguously perceived as being or not being highly urgent collision warnings. Toward this aim, the research team employed three methodologies across a series of research tasks. The first task (Perceptual Space Sort) employed regression techniques to identify key characteristics that accounted for a substantial amount of the variance in category classifications (see Perceptual Space Sort Task Results section). Cutoff scores for each of these key parameters were then established by examining the values of each characteristic that resulted in an 85 percent probability of classification as an alarm. The cutoffs established were a peak to total time ratio of greater than or equal to .7, an IBI of less than or equal to 125 ms, at least three harmonics, and a base frequency of greater than or equal to 1000 Hz. When all four of these criteria were met, the sound was classified as an alarm over 90 percent of the time. Meeting only the first of these criteria (peak to total time ratio greater than or equal to .7) resulted in an alarm classification rate of 70 percent. However, because only one sound in the initial set actually met all four criteria and since the criteria were established by the same data set that was being used to examine classification performance, further validation was warranted.

Using alternative methodologies (method of adjustment with single and multiple parameters) yielded results that confirmed several of the parameters (base frequency, presence of at least three harmonics). Though not specifically manipulated, the peak to total time ratio parameter was also met by the resulting sound created in each of the method of adjustment procedures. Though the method of adjustment yielded a longer IBI than the criteria had established, it could reasonably be determined to be an artifact of the fact that when adjusting IBI, participants were actually listening to bursts that contained three pulses that each had onset and offset times of 20 ms. Using criteria established in IEC 60601-1-8 (a medical alarm standard) a sound with onset and offsets create the perception of tempo when they alternate in and out of 90 percent of their peak amplitude. Therefore, the sounds being adjusted would have inadvertently yielded a tempo within the criteria established in the first sort task.

A final validation in the series examined the predictive capabilities of these criteria while participants were engaged in a simulated driving task and were concurrently performing a distraction task consisting of a visual version of an n-back task. The criteria had strong predictive capabilities under these conditions. Sounds that met all four of the criteria were classified as alarms over 90 percent of the time. Furthermore, when exposed to sounds meeting all four criteria participants were able to make their classifications much faster than when they met only one to three of the criteria. This suggests that sounds meeting these criteria not only have a high probability of being understood to be a warning but also that participants come to this understanding much faster than sounds that meet only a few of the criteria.

Together, the results of this research suggest that it is possible to define key characteristics and associated parameter values that could yield effective alarms for use in collision warning systems. A limitation of this research is that the laboratory methods had participants categorize sounds rather than respond to them in a more ecologically valid context (e.g., orientation and response to a perceived threat while driving). Follow-up research funded by NHTSA, and conducted by Westat and GMU, is ongoing to investigate signal categorization and response to unexpected signals during simulated and on-road driving. This follow-up research also includes haptic signals, a wider range of acoustic signals, and a broader age range of participants.

STUDY 2: WARNING PERCEPTION IN AMBIENT NOISE ENVIRONMENTS

Introduction

This study, which was conducted by Westat, was an experiment to measure aspects of driver perception of warnings and alerts under a range of ambient noise driving conditions on actual roads. Considerable research has addressed perceptions of in-vehicle warnings and messages, but the vast majority of this work has been conducted under relatively benign in-vehicle ambient noise conditions. Warnings, however, must remain effective under the likely range of noise conditions that may be anticipated in vehicles. Very little information exists on perception of meaning and urgency in noise.The characteristics and sound level of in-cab ambient noise may vary due to the vehicle's physical characteristics, the road surface, surrounding traffic, travel speed, and interior noise sources.

Method

Design The experiment was a three-factor design, with one between-groups factor (vehicle type) and two within-groups factors (interior noise condition, acoustic signal). Three different vehicles were used in the experiment in order to provide a representative range of vehicle types: (1) a small car, (2) a larger sedan, and (3) an SUV. Each participant drove only one of these vehicles. During the drive, data were collected under three different interior noise conditions: (1) windows up, music off; (2) windows open, music off; and (3) windows up, music on. The order in which each noise condition block was presented to participants was counterbalanced within each vehicle condition.

A set of 15 different acoustic signals was presented under each ambient noise condition. These included three unique voice messages and eight unique non-voice sounds. All eleven of the unique sounds and voices were presented at a sound pressure level (SPL) of approximately 65 dBA as measured near the driver's right ear. One of the voice messages and three of the non-voice sounds were also presented at 75 dBA, with the resultant total of 15 signals. The lower 65 dBA level is representative of a number of acoustic alerts as measured in production vehicles (Lin & Green, 2013). The higher 75 dBA level is more consistent with human factors guidance (e.g., Campbell, Richman, Carney, & Lee, 2004), assuming a moderate level of ambient vehicle cab noise.

Four different dependent measures were recorded to evaluate driver response. These included: (1) a measure of reaction time for the participant to detect the occurrence of a signal; (2) a rating of signal noticeability; (3) a rating of signal urgency; and (4) perceived meaning of the signal (chosen from a set of four alternatives).

Participants Participants included 34 drivers aged 22 to 49, with 13 males and 21 females. No participants reported having hearing decrements or using hearing assistive devices. All drove regularly, held valid U.S. driver's licenses, and passed a screener of their motor vehicle records.

Vehicles Three different classes of passenger vehicles were used in order to provide a range of vehicle types. These types were small car (2013 Hyundai Accent GLS), sedan (2013 Toyota Camry LE), and SUV (2013 GMC Terrain SLT). The specific vehicles used were selected from among the best-selling models in their class and with good rental availability. Twelve participants drove the small car, 11 drove the sedan, and 11 drove the SUV.

Roadway Data collection took place on a limited access toll highway (Maryland Route 200) with a 60 mph speed limit. This is a relatively new highway with smooth and uniform asphalt over most of its length. It is also generally free-flowing, without congestion. These attributes permitted good control over ambient road noise and speed conditions. The roadway has three travel lanes in each direction. Participants were instructed to travel in the right lane except when needing to pass slower vehicles.

Noise conditions All drives were conducted during clear weather on dry roads, with a target speed of 60 mph. The fan on the climate control system was set to a low, inaudible setting. During the baseline condition, all windows were closed and music was off. During the "windows open" condition, the front windows on both sides of the vehicle were fully opened. During the "music on" condition, the smooth jazz song "Café Amore" by Spyro Gyra played in a continuous loop. The volume of the music was adjusted by the participant to the volume they would typically use for their own music while driving alone in their own car. However, the experimenter required participants to set the volume at a level equating to at least 60 dBA, as measured in an otherwise silent vehicle

interior. The minimum SPL was established to ensure that the music could potentially affect participants' detection and ratings of messages.

Auditory signals and stimulus presentation Fifteen auditory signals were compared. There were 11 unique alerts presented at approximately 65 dBA. Four of these sounds were also presented at approximately 75 dBA. All sounds were initially volume-adjusted to these levels, but were then adjusted for perceptual equivalence of loudness, as determined by a jury of six individually tested raters. Thus, the 65 dBA sounds were of subjectively equivalent loudness to a 65 dBA continuous pink noise signal (i.e., a signal of random noise containing equal amounts of energy per octave). The 75 dBA sounds were digitally amplified by 10 dB from the 65 dBA level.

The alerts used in this experiment were adapted from examples of current in-vehicle warnings and alerts of various types, other sounds found in various sources, and synthetic speech messages created using an online text-to-speech generator.[1] It is important to note that the signals that were sourced from current in-vehicle systems were presented using a different speaker in a different vehicle interior, and are not necessarily presented at the same SPL as the original alerts. Therefore, the results of this experiment do not necessarily reflect upon the messages as used in their native vehicle environments. The alerts used in this experiment are briefly described below. Alerts 1 through 8 are sounds presented at the 65 dBA level; alerts 9-11 are voice messages at the 65 dBA level; and alerts 12-15 are the subset of alerts presented at the 75 dBA level.

1. FCW 1: One burst of 20 fast beeps with a relatively high frequency profile.
2. FCW 2: Four bursts of four fast beeps with a relatively low frequency profile.
3. Blind spot warning: Three bursts of four fast beeps, each with a smoothed onset and offset and a sustained low intensity sound between beeps.
4. Pedestrian warning: A single sustained beep with a duration of 2 s.
5. Seat belt alert 1: A single chime that decays to silence in the span of about 2 s, with intensity varying in a wavelike pattern.
6. Seat belt alert 2: Two chimes, each of which decays to silence in the span of about 1 s.
7. Park assist 1: One burst of eight beeps.
8. Park assist 2: Two bursts of three beeps.
9. Female voice – not urgent: Female voice says "Attention."
10. Female voice – urgent: Female voice says "Warning, warning."
11. Male voice – urgent: Male voice says "Warning, warning."
12. FCW 1 (high): Same as FCW 1, but presented at 75 dBA.
13. Blind spot (high): Same as (3), but presented at 75 dBA.
14. Park assist 1(high): Same as (7), but presented at 75 dBA.
15. Female voice – urgent (high): Same as (10), but presented at 75 dBA.

During the experimental drives, the auditory signals were presented by an X-Mini II XAM4 capsule speaker mounted on top of the dashboard immediately behind the steering wheel.

Within each noise condition block, the experimental control software generated a random presentation order for the 15 auditory signals. The software provided a random time gap that ranged from 10 to 50 s and averaged 30 s from the completion of the previous sound's ratings to the presentation of the next sound. Once the random time had passed, the software indicated to the experimenter that the next signal could be activated. The actual triggering of the trial was done by the experimenter, who first determined that there were no unusual acoustic circumstances (e.g., a large truck passing or a patch of noisier roadway surface). When the participant detected the signal they pressed a microswitch button, worn on their finger or thumb, to provide a response time. The data collection system recorded the response time and then cued the experimenter, who was seated in the right rear seat, to verbally present a series of rating and choice questions. The participants rated:

- Noticeability (defined as "The sound is easily noticeable among other sounds and noises in the vehicle"); 1=not very; 7=extremely

[1] http://www.oddcast.com/home/demos/tts/tts_example.php

- Urgency (defined as "The sound conveys a sense of importance, motivating you to make an immediate response"); 1=not very; 7=extremely
- Perceived meaning. The sound was placed in whichever of four categories the participant felt best matched the meaning conveyed by the signal
    1. Urgent crash warning
    2. Safety information (other than urgent crash warnings)
    3. Information not related to safety
    4. Incoming personal communication (e.g., call, text message, email)

If the participant did not respond to the alert by pressing the microswitch within eight seconds of signal initiation, the participant was deemed to have failed to detect the signal and no rating questions were asked. The participant received no feedback that there was a missed signal.

Results

Driver perception of the auditory signals was influenced by both the nature of the signal and the ambient noise background. Figure 2 shows the percentage of signals detected by the driver for each signal under each ambient noise condition. Few participants missed any signals in the baseline condition (windows up, music off). In the music condition, a few signals were detected in the 70 percent or fewer range, but most were detected at least 80 percent of the time. The open windows condition interfered with detection more dramatically. All four of the signals presented at the 75 dBA level continued to be detected well even with the windows open, but for the 65 dBA signals there was a wide disparity in performance, with a few detected by more than 90 percent of participants and three others detected by only around 10 percent of participants.

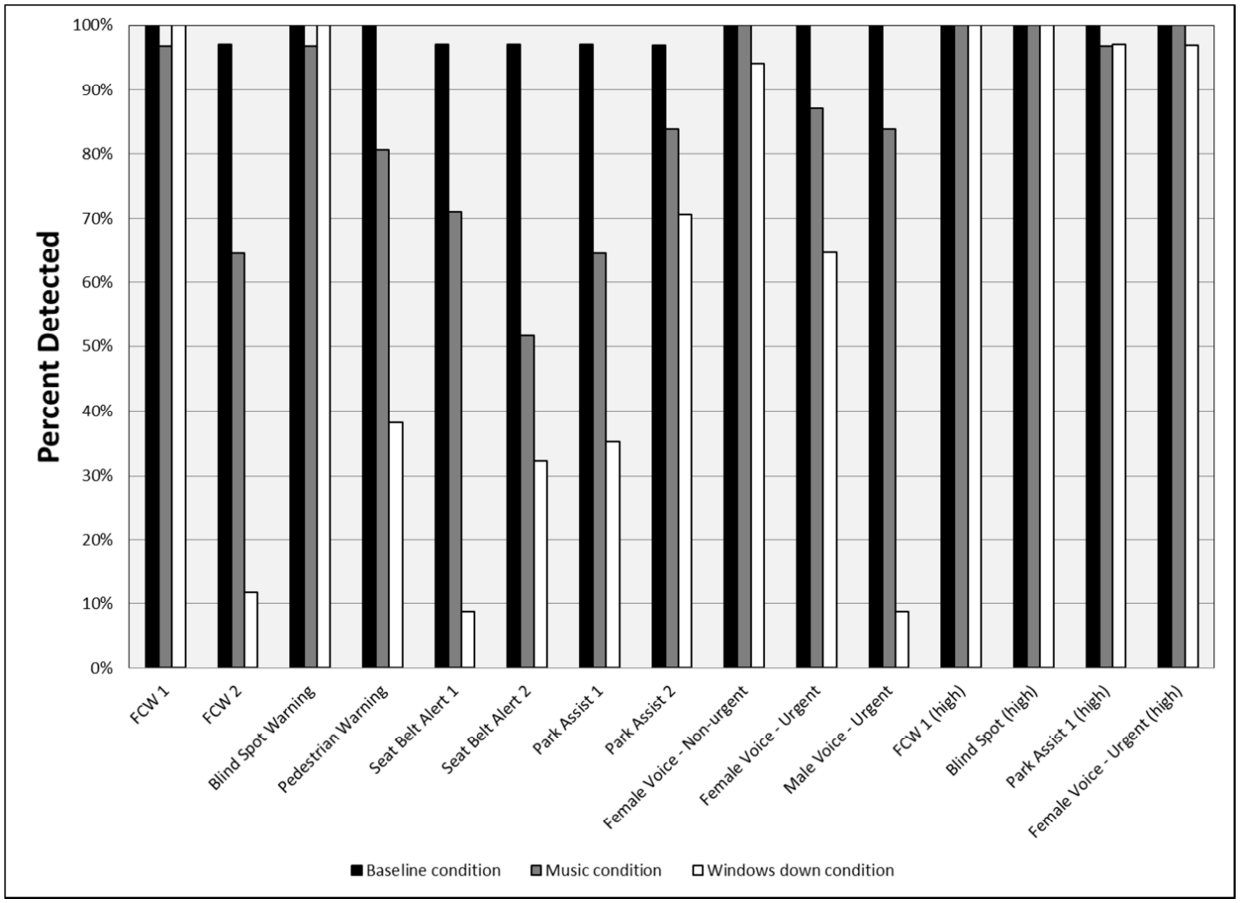

Figure 2. Percentage of participants who detected alerts under each ambient noise condition.

Three-factor (signal, ambient noise background, vehicle type) ANOVAs were conducted for the measures of rated noticeability, rated urgency, and response time. The three parallel analyses yielded identical conclusions. Signal, noise condition, and their interaction were all statistically significant ($p<0.0001$). There was no main effect of vehicle type and no interaction of vehicle type with ambient noise background. There was a significant interaction of signal with vehicle type, though there was no clear pattern of effects. It may be expected that due to the complex and varied geometry of the vehicle cabin space and the nature of the reflective and absorbing materials, the differences in the acoustic space might differentially affect some signals. There was no three-way interaction. Figure 3 shows the group mean rating of urgency for each signal under each ambient noise condition. The figure shows that there were dramatic differences in perceived urgency, even among sounds that were equated for similar perceived loudness under relatively quiet conditions.

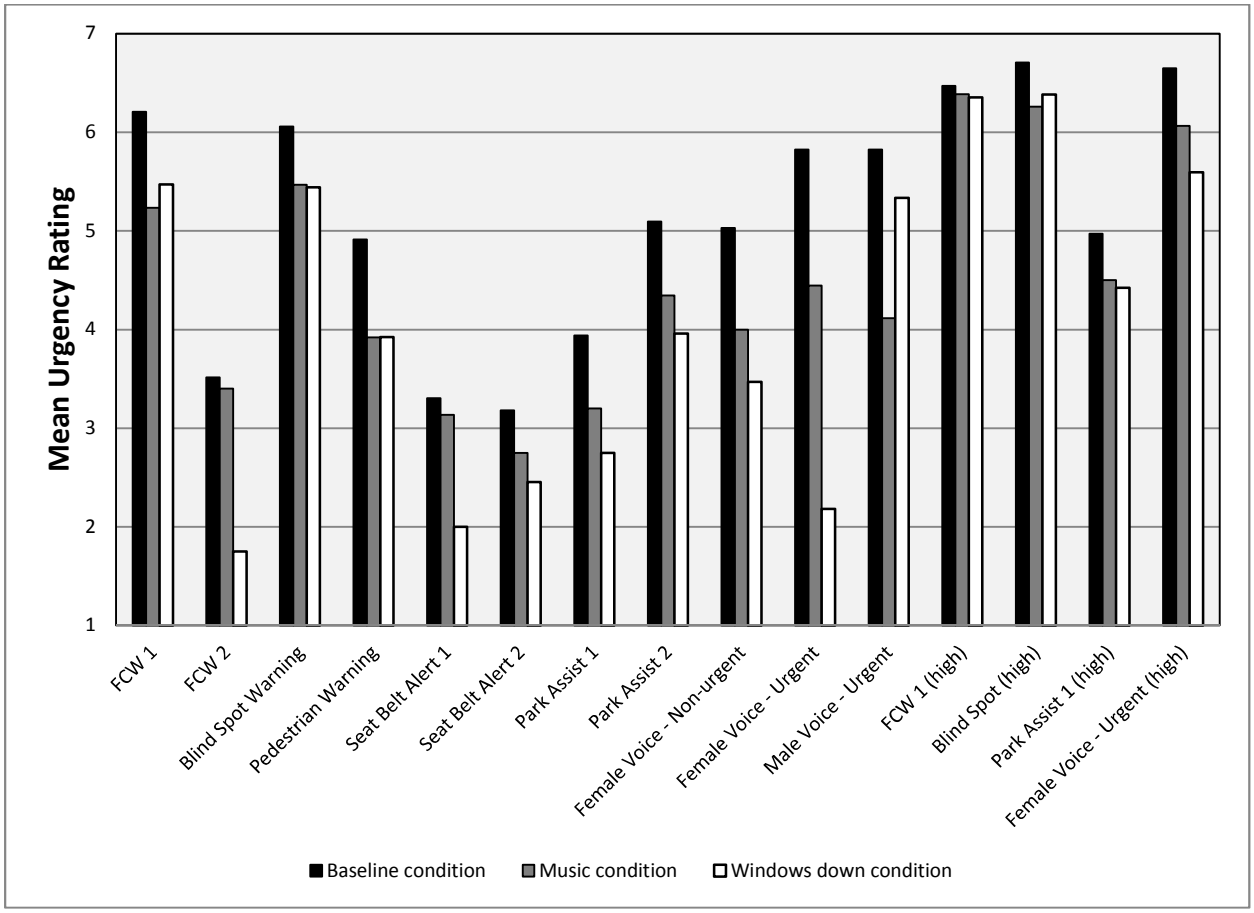

Figure 3. Mean urgency rating for each combination of signal and ambient noise condition.

The ambient noise condition influenced the category of meaning that a listener assigned to a particular alert. Participants had the option of classifying a given alert as "urgent crash warning," "safety information," "information not related to safety," and "incoming personal communication." As expected, the various alerts differed in terms of how they were interpreted, with some predominantly viewed as urgent crash warnings and others predominantly view as unrelated to safety at all. Multinomial logistic regression was used to analyze the perceived meaning classifications. Multinomial logistic regression is used to predict the probability of category membership on a dependent variable based on multiple independent variables. This approach is an extension of binary logistic regression that allows for $k>2$ categories of a dependent variable. Maximum likelihood estimation is used to evaluate the probability of category membership. The current model analysis was performed in SAS and used a cumulative logit model with Fisher's scoring as an optimization technique. Differences of least square means are reported with Sidak adjusted p-values. The analysis found signficant effects of ambient noise ($p=0.0036$), acoustic signal ($p<0.0001$), and an interaction between these two factors ($p<0.0001$). The key outcome to note here is that

ambient noise condition had a significant effect on participants' categorization of signals, and the significant interaction term reveals that different signals were affected in different ways by the various ambient noise conditions, revealing a complex relationship between ambient noise and acoustic signals with regard to perceived category of meaning.

Study 2 Discussion

As an initial study on this topic, the experiment demonstrated very sizable effects of ambient noise conditions that might reasonably be expected to occur on occasion. Background noise from music, and especially from open windows, interfered with the perception of auditory signals presented at 65 dBA. Interference was not very pronounced for the set of 75 dBA signals, although only four signals were included at this level. The set of sounds and voice messages equated for approximately equal loudness under relatively quiet listening conditions differed substantially in noticeability and urgency even under the baseline condition and even more under the music and open windows conditions. Under noise conditions, 65 dBA signals were less likely to be detected, and when they were detected they typically lost much of their perceived urgency, which may compromise their effectiveness as urgent crash warnings. Some sounds suffered low detection rates under noise, particularly the windows open ambient condition.

This experiment was designed to provide an initial examination of the extent to which possible ambient noise conditions might interfere with signal detection and meaning. It was not intended to provide any systematic evaluation of signal features or parameters regarding their resistance to noise effects. However, based on the limited sample of sounds and conditions, it appeared that the predominant frequencies that characterize a signal may relate to perceived urgency under noise. Sounds with base frequencies below 1000 Hz generally performed worst and those with base or significant components above 1500 Hz performed best. However, this observation is based on a limited sample of sounds that also differed in a number of other respects, and so it should be considered tentative. A series of follow-up laboratory investigations conducted within NHTSA's CWIM research program replicated and extended the findings of this on-road study by investigating additional sounds, ambient noise conditions, and sound pressure levels.

CONCLUSIONS

The National Highway Traffic Safety Administration's Crash Warning Interface Metrics program included a series of research studies intended to address research needs related to the driver-vehicle interface for advanced crash warning systems. The two complementary studies described in this paper investigated the signal characteristics that lead to people clearly distinguish urgent alert sounds from other message categories (Study 1), and how the effects of various in-vehicle ambient noise conditions affect the detection and perceived meaning of signals (Study 2).

In Study 1 (Categorical Perception of Alerts), a wide range of parameters was systematically tested, and four parameters were found to be most important in influencing categorization as a warning or safety alert: base frequency, number of harmonics, IBI, and peak to total time ratio. The cutoff values at which a sound becomes unambiguously perceived as an alarm were refined and validated in a series of tasks including a driving simulator task under cognitive load. The findings can potentially help designers create in-vehicle systems in which intended crash warnings are reliably perceived as urgent and critical and in which intended lower urgency message types do not lead to confusion with urgent warnings.

In Study 2 (Warning Perception in Ambient Noise Environments), a set of 15 sounds was presented to drivers under three ambient sound conditions. Results show that many sounds that were easily detected and perceived as urgent in a relatively quiet vehicle interior were much less likely to be detected or perceived as urgent in a louder vehicle interior (e.g., with windows open). Main effects of detection rate for sound stimulus suggest that sounds adjusted for equal loudness in a quiet environment are not necessarily equally loud in a vehicle environment. The presence of interactive effects between different sounds and ambient noise conditions suggests that some sounds are more resistant to the interfering effects of elevated ambient noise than others.

Taken together, the results of these two studies show that the acoustic characteristics that cause a sound to be unambiguously interpreted as an urgent alarm or other category are quantifiable, but that the ambient noise condition in which sounds are heard can have a significant effect on signal detection and interpretation. In-vehicle acoustic

signals, then, should be quickly and unambiguously interpreted with the meaning intended by the designer, and should maintain their detectability and meaning in a variety of ambient noise conditions. Study 2 found that sounds played at 75 dBA were substantially less impaired by elevated ambient noise than sounds played at 65 dBA, suggesting that alerts presented at approximately 65 dBA, as is common in production vehicles, might not always maintain their detectability and meaning under elevated ambient noise. The limited number of stimuli investigated in this study does not provide enough data to suggest an ideal loudness, nor suggest features that might make a given sound more detectable or more reliably categorized at a given loudness.

These studies have some notable limitations. As noted above, signal loudness has a substantial effect on signal detection and interpretation, especially in elevated ambient noise. Study 1 did not systematically manipulate signal loudness, and Study 2 only manipulated loudness at two levels for five signals. Study 2 also only investigated three ambient noise conditions. These studies also involved participants focused on an experimental task without the context of a real on-road information environment. As such, there were no conditions or hazards associated with signals, nor were participants expected to react or respond to signals (other than by providing a categorical response). Follow-up research conducted within NHTSA's CWIM program has continued the ambient noise line of research by replicating the on-road findings in a laboratory setting and extending the research to investigate additional sounds, ambient noise conditions, and sound pressure levels. A separate study conducted within NHTSA's Connected Vehicle program will extend the categorical perception research by investigating additional sounds, as well as haptic signals, in driving simulator and on-road experiments.

REFERENCES

Campbell, J.L., Richard, C.M., Brown, J.L., & McCallum, M. (2007). Crash warning system interfaces: Human factors insights and lessons learned. Report no. DOT HS 810 697. Washington, DC: National Highway Traffic Safety Administration.

Campbell, J.L., Richman, J.B., Carney, C., & Lee, J.D. (2004). In-vehicle display icons and other information elements, Volume I: Guidelines. Report no. FHWA-RD-03-065. McLean, VA: Federal Highway Administration.

Lewis, B.A., Penaranda, B.N., Roberts, D.M., & Baldwin, C.L. (2013). Effectiveness of bimodal versus unimodal alerts for distracted drivers. Paper presented at the International Conference on Driving Assessment, Bolton Landing, NY.

Lin, B., & Green, P. (2013). Measurements of driver-assistance warning-signal characteristics in 2013 cars. Report no. UMTRI-2013-03. Ann Arbor, MI: University of Michigan Transportation Research Institute.

Santangelo, V., & Spence, C. (2007). Multisensory cues capture spatial attention regardless of perceptual load. Journal of Experimental Psychology: Human Perception and Performance, 33(6), 1311-1321.

EVALUATION OF IMPACTS OF IT-INCIDENTS ON AUTOMOTIVE SAFETY WITH REGARD TO SUPPORTING REACTION STRATEGIES FOR THE DRIVER

Sven Kuhlmann
Robert Altschaffel
University of Magdeburg
Germany

Tobias Hoppe
Jana Dittmann
Christoph Neubüser
University of Magdeburg
Germany

Paper Number 15-0156

ABSTRACT

Objectives

The integration of modern IT technologies into vehicles brings up several new challenges in automotive systems engineering. While current technology aspires an exclusive use of electrical and electronic control systems for relevant functions, such as engine control or X-by-wire systems, growing dependency on electronic systems increases the vulnerability of modern cars to both accidental and intentionally forged IT incidents. Especially the constantly increasing complexity of and interdependencies between different automotive IT systems makes it difficult for the developers to foresee all potential fault conditions or to prevent unauthorized actions from taking effect. As if these problems were not enough of a challenge, especially in automotive IT environments IT- incidents can possibly also affect the safety of the car, its passengers and other road users.
Building on a study on IT security warnings [1] and comparing with corresponding ASIL levels, we carried out a driving simulator study to evaluate driver reactions to various error and security relevant scenarios.

Methods

Assuming that malfunctions of electronically supported control systems will endanger the safety of the car, a driving simulator study was designed and executed. These tests cover both security-related and safety-related lsources of failures (i.e.: accidental or provoked malfunctions) and scenarios with different criticality (based on ASIL A, B, C, D – [2]). The reactions of 40 uninformed drivers were observed and analyzed. In particular failures of engine, steering and brakes were executed in different road and traffic scenarios (e.g.: slow vs. high speed, low vs. high traffic density). The reactions of the drivers were recorded and, additionally, the controllability of the situation was observed as perceived by the drivers (using a think-aloud test). Furthermore, the study evaluated the potential of appropriate warning and reaction strategies that could support the reaction of the driver in critical situations as developed in [1].

Results

The results show differences in driver behavior within a specific failure situation and an even greater degree between various failure situations. We found different types of accidents following the loss of steering and braking function – but no accidents caused by the loss of engine function. Interestingly, the results show the highest rate of recognition for the engine turn off scenario, where as in the autonomous acceleration and loss of brake function 15-17% of drivers did not recognize the malfunction. Besides this, we introduce different strategies to warn and support drivers in such situations. Especially when losing the ability to steer and brake, the warnings showed positive impact if the driver is warned ahead and stops the car before the complete loss of those functions. When the warning appears together with the function loss, a significant improvement of crash count and severity could not be observed.

Conclusions

This work shows the impact security-related incidents can have on the safety of concurrent and future vehicles. It shows the potential of decreasing the severity of these incidents by using tailored warnings and shows a first evaluation of the feasibility of such an approach. It was shown that a loss of engine function leads to a safe stop of the car while a loss of steering or braking ability or an autonomous acceleration lead to an accident in 45% up to 71% of all cases. The severity of those accidents is not significantly correlated to the type of malfunction.

INTRODUCTION

Concurrent automotive systems are more than mere mechanical systems. With a broad array of embedded systems, a modern car is akin to a driving computer network. This network consists of various electronic control units (ECU) which communicate using different bus-systems (see [3] for further details). These ECUs implement various functions. Simple ECUs control simple components like the window openers while more complex ECUs handle more complex tasks like engine control. Other embedded systems support the driver by giving him information about the environment or suggesting a route. All these factors establish cars as complex multimedia systems (as discussed in [4] On one hand this opens a broad range of possibilities to interact with and support the driver. On the other hand all these interconnected electronic components enlarge the error-proneness and attack surface of the automotive system. Looking at the security aspect, in automotive scenarios there is always the risk that unauthorized tampering with car IT can escalate into a safety incident – with the risk of the driver losing control of his car. Based on prior work from us where we demonstrated real IT attacks on several automotive subsystems ([5], [6]), such common automotive vulnerabilities have also been illustrated later on full cars by [7] and [8]. The spectrum of observed results leads up influencing the brakes, the steering wheel or disabling the engine by specially crafted CAN bus commands. As these practical analyses have substantiated, unforeseen IT incidents (like intentional, IT-based attacks) also bear severe safety implications. This contribution therefore focuses on the impact of (e.g. security-related) IT incidents on the safety of automotive systems. However, also unintended malfunctions (e.g. caused by system/component defects, software coding errors, or unforeseen and unhandled system interdependencies) could cause similar results. With reference to different examples of such potential incident causes, the aim of this paper is to analyze reactions of drivers in different incident scenarios with different grades of safety criticality.

A Model to Describe Safety and Security Relevant Consequences

As a first step we revert to the Automotive Safety Integrity Level (ASIL), a model to describe consequences of safety and security-related incidents. This model allows us to specify various safety-related incidents and to evaluate their severity with regard to automotive scenarios. ASIL is defined in ISO 26262 ([2]) and consists of three components:

- Severity (S) is the severity of potential harm caused by an incident (similar to the SIL). It ranges from S0 (no injuries) via S1 (light and moderate injuries) and S2 (severe and life-threatening injuries with probable survival) to S3 (life-threatening with uncertain survival or fatal injuries)
- Probability (E) shows how probable it is that such an incident occurs. The levels range from E0 (incredible) via E1 (very low probability), E2 (low probability), E3 (medium probability) to E4 (high probability)
- Controllability (C) evaluates if the driver would be able to control the situation if such an incident occurs. It ranges from C0 (controllable in general) via C1 (simply controllable) and C2 (normally controllable) to C3 (difficult to control or uncontrollable)

These three categories allow a categorization of various incidents in respect of their severity. As a further concept ASIL contains an overall integrity level based on these three components. Similar to SIL, four different levels of differing severity (ASIL A, B, C and D) exist with an additional level for no or very low severity (QM – Quality managementt). Table 1 gives an overview on the impact of the three categories on the overall safety integrity level.

Table1.

Overview on the impact on the overall safety integrity level with highlighted cells representing the tested scenarios

		C1	C2	C3
S1	**E1**	QM	QM	QM
	E2	QM	QM	QM
	E3	QM	QM	A
	E4	QM	A	B
S2	**E1**	QM	QM	QM
	E2	QM	QM	A
	E3	QM	A	B
	E4	A	B	C
S3	**E1**	QM	QM	A
	E2	QM	A	B
	E3	A	B	C
	E4	B	C	D

Using ASIL we are able to evaluate different scenarios in respect of their severity. This allows us to create different test scenarios with varying severity either in respect to their overall ASIL or its components.

Factors in an Automotive Environment

In an automotive environment various factors contribute to othe ASIL. We identify three basic groups of factors which can influence the severity of an incident. These groups deal with the vehicle, the traffic condition and the driver himself. It is important to evaluate all these factors to determine a reliable ASIL for the current situation. As an exhaustive list of factors would exceed the scope of this paper we give a few examples for the various groups of factors.

- Vehicle-dependent factors: This group consists of factors which are inherent to the vehicle and generally don't change rapidly. They consist, for example of:
    - Vehicle type (e.g sports car, bus, transporter, truck)
    - Implemented driver assistance systems (e.g. adaptive cruise control with inherent autonomous acceleration and brake functionality or lane keeping assistant with inherent autonomous steering functionality)
- Traffic-dependent factors: These factors describe the current situation of the traffic or the road. They can change rapidly. Examples consist of:
    - Speed of the vehicle
    - Current lane conditions (e.g. clean or dirty street, potholes)
    - Traffic density (e.g. rural road with no other traffic participants or inner city during work traffic)
    - Weather conditions (e.g. sunny and warm or cold and rainy)
    - Noise (e.g. no outside noises or a lane directly next to a noisy construction site)
- Driver-dependent factors:
    - Response time of the driver for the perception of a certain situation
    - Response time of the driver for the execution of a reaction on a certain situation
    - Quality of the reaction of the driver
    - Interpretation of occurred errors or situations by the driver
    - Physical condition of the driver (e.g. fit or drowsy)
    - Distractions that could influence the attentiveness of the driver (e.g. a nearby airport with a starting plane nearby)

METHODS

We decided to simulate various scenarios of differing severity to record and evaluate the reactions of various drivers. Since the failure of electronic computing control functions endangers the safety of the car, its driver and other traffic participants, we decided to perform these tests using a driving simulator. We created the scenarios according to the formerly discussed ASIL.

The Driving Simulator

We used a driving simulator situated in our lab as the hardware for the tests. It consists of electronic control units (ECUs) and the dashboard of a VW Passat B6 with the ability to trigger instruments such as the tachometer. Furthermore, a forceable and force-feedback steering wheel from Logitech with additional accelerator and brake pedal is integrated into the driving simulator. During the simulation, the driving lane is displayed in front of the vehicle on a semi-transparent screen via rear video projection. We used commercial driving simulation software ([9]) for the simulation of the various test tracks. It provides the adjustment of several factors such as traffic density or the type of vehicle, which will be regarded during the evaluation. Furthermore it offers the possibility to create tracks and specify scenarios.
In order to achieve proper evaluation results 46 test persons were invited to take part in the tests. Thereby each test person has performed four test runs including different scenarios. In the different scenarios the failure of various vehicle components was simulated. As result of the evaluation the reactions of all test persons on these situations were observed, as well as the influence of acoustic and visual warnings on the reaction of the test persons. It should be noted that each test person was only warned in one of the four test runs to prevent the drivers of special expectations during the test runs. Furthermore the appearance of the warning was varied for different test persons, as well as the order of the four chosen scenarios. This was done for the prevention of coherent results and succession effects, which could occur if some combinations of scenario orders with the same appearance of warnings are used to often and could therefore distort the overall evaluation. In total one test run took 1 ½ hour including questionnaires.

Scenarios and derived Hypothesis (ASIL Classification)

We based our scenarios on the different ASIL defined in ISO 26262 ([2]). We created a theoretical scenario for each of the four levels.
ASIL A: In this scenario the engine control unit is manipulated in order to allow law enforcement to safely stop the vehicle using remote access. This scenario implies that the law enforcement agents choose a situation in which the forced stop doesn't endanger other traffic participants. Such a situation implies sparse traffic and low speeds. Hence the severity is rated as S1. The manipulation performed by experts implies a high probability of E4 (e.g. by the means of electromagnetic pulses as researched in [10]. In general the driver will still be able to control the vehicle after the engine is shut down and should be able to stop safely. Therefore, as the derived hypothesis we rate the controllability of this scenario as C2 yielding an ASIL A classification.
ASIL B: In this scenario the car owner tries to manipulate the engine of his car himself, aiming to gain better performance. A failure caused by such manipulation could lead the car to accelerate autonomously while driving, endangering the safety of the driver and other vehicles in the surrounding area. In this scenario we choose an urban area and dense traffic. Hence the rapid acceleration of the car would cause a higher severity of S2. It is not certain that a manipulation of the engine would lead to a malfunction as described. The probability would be rated as E3 though, since most home tuners lack the expertise of professional automotive mechanics. Looking at the high traffic density in this scenario, an affected vehicle would be hard to control. Controllability is rated as C3 in this scenario which leads to an overall ASIL B classification.
ASIL C: This scenario covers another manipulation done by the driver. In this case the driver manipulates his infotainment system to show video (e.g.: TV) while driving. In general this functionality is disabled if the current speed exceeds walking pace. In this scenario, a 3rd party tool to inject forged bus messages with a lower speed signal has been installed by the driver, which would allow him to watch video while driving. As a side effect, this could also set steering support to maximum which could lead to fatal effects especially in high speed scenarios taking place on a speedway (Severity S3) – because these usually require little or no steering support. This causes a relative low controllability for the driver (Controllability C3). As in the previous scenario it is not certain that such a signal manipulation would necessarily propagate to all ECUs. Therefore we rate the probability as E3 again and get an overall ASIL C classification.

ASIL D: In this scenario a potential attacker manipulates the compressor of a heavy transporter. The compressor is also responsible to provide air pressure for the braking process. Such a manipulation would make it impossible for the driver to brake. In this scenario we assume that the transporter reaches the end of a tailback with an option of steering on the hard shoulder and slow down the transporter by engine break function. The momentum of the transporter would ensure fatal consequences which we rated as severity S3 – due to the lack of a higher possible rating. As there would be no control over the brakes any more, the controllability is rated as C3 – again, due to the lack of a higher possible rating. As this manipulation would be performed by an expert, the probability is rated as E4 leading to an overall ASIL D classification.

General Hypothesis

Following the results in the automotive warning research our test drivers should benefit from a warning in general [1]. We expect positive effects especially in ASIL A, B and C scenarios, due to the possibility to brake the car in these scenarios. We used a general warning without any information about the affected system (engine, steering, acceleration pedal or brakes) or reaction strategies (e.g. "Stop car immediately!"). This allows us to investigate the reaction of the drivers as a direct response to the vehicles system failure. Given the timeframe necessary to process incident, deriving, planning and executing an action, we expect a higher amount of false reactions (or no reactions at all) to critical incidents, like the loss of brake or control functionality, due to time constraints. For ASIL A scenario we expect the highest number of correct reactions.

Test Drivers

46 test persons took part in our tests. These persons had an average age of 26.9 years and an average driving experience of 7.3 years. Each test person performed one test for each of the four scenarios. From these 46 test persons 42 where able to finish all our tests. The remaining four persons needed to drop out due to simulator sickness at various stages of the test. For three of them a partial evaluation of their results was done. Before the start of our tests the participants weren't familiar with the test scenarios and the failures.

Measurement

We observed the reactions of all test persons on the given scenarios as well as the influence of acoustic and visual warnings. To prevent the test persons from having expectations during the test runs, each subject was only warned during one of the four scenarios. To prevent coherent results and succession effects, we changed the order of the various tests between different test persons. Furthermore the appearance of the warning was varied for different test persons.

The tests themselves started with a questionnaire and an initial trial run to get familiar with the simulator. Afterwards the respective scenarios started. During these scenarios the test persons committed thoughts and feelings using the think-aloud-method [11]. We combined these protocols with the observation how the test persons reacted inside the simulation and what result (accident or no accident) concluded. For doing so we used a simple coding scheme (Did the driver notice anything is wrong? Did the driver brake? Did the driver try to move the car to the sideline? Did an accident occur?). This allowed us to determine if the test persons correctly recognized the situation and if accidents could be avoided. We also looked closely on the health condition of the test persons in order to prevent negative influence due to simulator sickness [12]. In case of simulator sickness, a short break was scheduled and the test was either repeated aborted.

Warnings

Three different types of warnings are mainly used in current automotive and related research: visual, acoustic and haptic warnings. In general visual warnings are displayed on the head-up display (HUD) of the vehicle to inform the driver about possible critical situations. The size and position of the warning sign on the HUD is of great importance because of its influence on the reaction of the driver. An Evaluation of size and position of visual warning signs and their effect on the driver has been performed in [13]. Acoustic signals are also frequently used in current vehicles to direct the attention of the driver to critical situations. In this case duration and frequency of the acoustic warning signal are the decisive factors for the success of the warning. In [13] an analysis of different acoustic signals with different length and frequencies has also been performed to observe how different test persons react on such signals in critical situations. The third type of warning signals, which is used in vehicles, are so called haptic warnings. In

this case the driver is for instance warned through tactile vibration of the steering wheel during his drive. In current vehicles this is for instance used for lane tracking support in case the driver is about to leave his current lane [14]. Following the previous research we selected visual and acoustic warnings for this experiment. The acoustic warning was a sine tone with the duration of one second and a frequency of 1320 Hz which was repeated 4 times. The visual warning was provided as a red triangle with a black exclamation mark inside. It was presented virtually inside the projection of the simulation scene as a head-up display warning. Additional instructions were not given to avoid influences to the subject's reaction.

RESULTS

Due to our observations the majority of test persons were able to determine the particular situation or failure in each of the four scenarios. We observed the highest rate of correct detection in the ASIL A scenario. This is mostly due this being the least severe scenario which also does not require a sudden reaction by the driver. 94.1% perceived this situation correctly. Regarding the other three (more critical) scenarios, we observed a higher rate of misinterpretations. We see the reason for this higher rate in the fact that these situations require a much faster reaction of the driver which leads to greater attention on the task itself. One example can be seen in the circumstances that several test persons failed to notice the fact that the vehicle accelerated by itself in the ASIL B scenario. They blamed a failure of breaking for the fact that the vehicle didn't slow down and didn't notice that the vehicle even accelerated. In the ASIL C scenario most wrong determinations resulted from problems of some test persons to track the lane while driving with high speed.
At first glance it appears that the provided warning has not improved the correct determination of the provided situation. This is due to the fact that most test persons already stopped the vehicle in safe place immediately after the warning was provided. Hence the test persons were not able to correctly determine the situation since they reacted on the warning before the situation itself manifested. By sorting out these occurrences, we observed the results shown in figure 1.

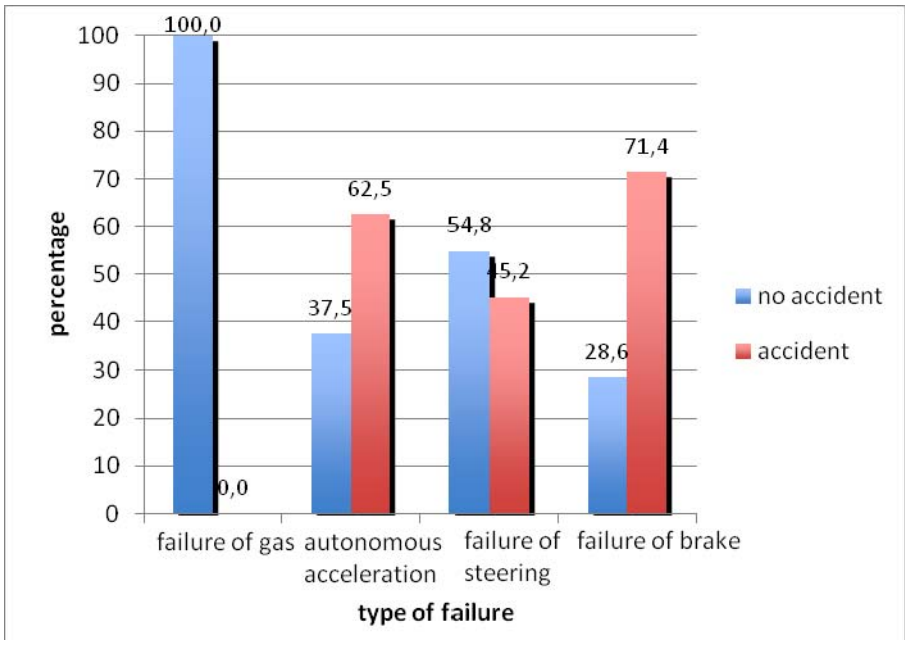

Figure1. Observed reactions of the test persons without a warning

We observed a correct determination rate of 100% in ASIL A (10 out of 10) and ASIL C (7 out of 7). In ASIL B and ASIL D these rates were lower with 82% (9 out of 11) and 87% (7 out of 8) respectively. Therefore it can be assumed that the provided warning had a positive influence on the number of correct determinations although the number of results led to this not being provable by our tests.

The second focal point for our research was how the persons reacted in these situations with regard to prevent harm for themselves and other traffic participants. At first we observed the occurrence of accidents in test runs without a warning before the respective situation. We predicted that the rate of accidents would increase according to the rising severity of the various scenarios. This prediction proved correct in general with 0% accidents in scenario ASIL A while ASIL B, ASIL C and ASIL D led to 62.5%, 42.2% and 71.4% accidents, respectively.

By the inclusion of an error warning before the occurrence of the situation we could reduce the number of accidents in general as shown in figure 2. ASIL B went down from 62.5% to 50% and ASIL D from 71.4% to 55.6%. ASIL C however suffered an increase in the number of accidents to 45.5%.

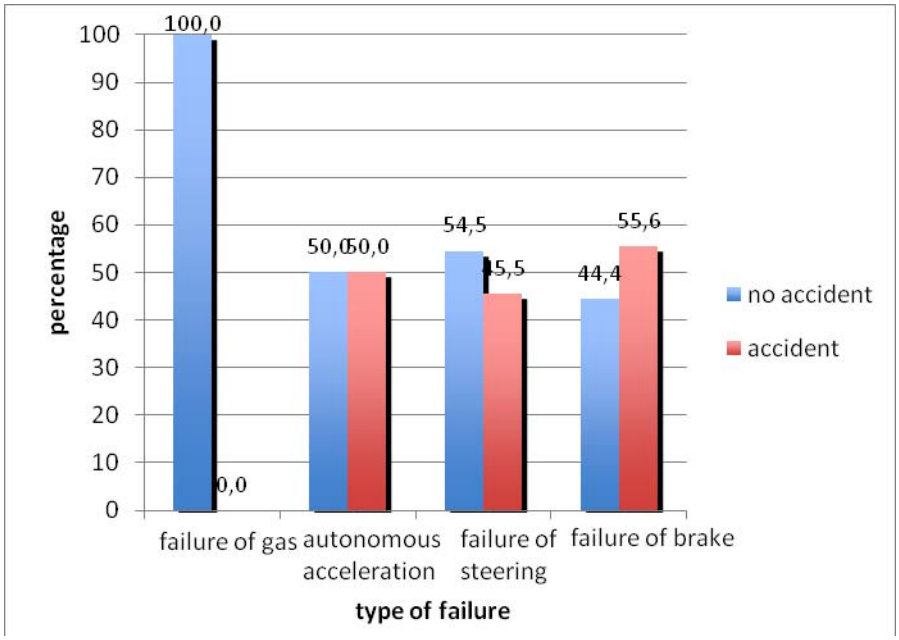

Figure2. Observed reactions of the test persons supported by a warning

This better overall performance is supported by the fact that some test persons almost instantly reacted on the warning by turning on the hazard lights and trying to stop their vehicle in a safe place. The overall attentiveness was also increased since the test persons were more concentrated on determining the reason for the failure. This became most obvious due to the think-aloud-tests. However avoiding an accident was still dependant on the skill of the driver in these sudden occurrences of critical situations as observed in ASIL B scenario. Although traffic density and average speed was rather low for the ASIL B scenario, because of strictly urban traffic, this rather high percentage of accidents had not been expected first. The main reason therefore was that many of the test persons were quite startled of this situation which often led to uncontrolled steering to avoid collisions, because of the decreased braking effect (according to the acceleration of the vehicle). Those who were able to control the car properly, for instance by activating the emergency brake and shifting down the engine, could avoid accidents because they remained concentrated. As this is a situation that rather does not seem likely, the reaction of most of the test persons is not surprising in the end because of its rather sudden appearance in compare to the ASIL A scenario. The results from the think aloud test revealed a general inhibition of the positive warning effect if the expected failure from a warning does not match the failure situation. In the think aloud test and the interview after the simulator study, those drivers reported that it leads to surprise and confusion and therefore to the inadequate reaction. This also applies if the driver is not able to derive a reaction strategy from the given warning, e.g. if the warning is missed, unspecific, appears too early or late or leads interpretation to a wrong system (brake vs. steering). Beside the slight decrease in the statistical numbers it could be observed that the severity of the accidents which still happened decreased. The increase in overall driver attentiveness led to better reaction even if an accident could not been avoided.

Overall the evaluation shows that a preceding warning decreased the amount of accidents or at least their severity. Further research questions were related to the design of the warnings. In summary, the chosen warning design was reflected as appropriate and helpful. Nevertheless, the test drivers suggested additional clear advice for an appropriate reaction strategy (e.g. stop the car immediately).

Inferential Statistics

First of all it was tested if the assumed hypothesis, that a preceding warning affects the number of correct determinations of each situation, could be verified. Therefore a chi square test was performed and the contingency table was created for each of the four observed scenarios regarding the warning condition and the determination condition. As some of the test persons were not able to determine the actual situation according to their immediate reaction on the warning, these test persons were excluded for the chi squared test to gather the correct values. This has been performed for each scenario to verify if the features warning condition and determination condition are stochastically independent.

As each value is lower than the critical value of the chi squared test of 3.841 [15], it could not be verified that the samples differ significantly. Therefore it could not initially be stated that the provision of a preceding warning influences the correct determination of a certain scenario. To prove if the calculated value was just received randomly according to the samples, the p-values for each scenario were calculated as seen in table 2: Failure of gas (p=0.5), autonomous acceleration (p=0.48), failure of steering (p=0.37) and failure of brake (p=0.47).

Table2.

Chi squared values by scenario

Scenario	Chi Squared Value
failure of gas	0.0
autonomous acceleration	0.057
failure of steering	0.48
failure of brake	0.103

Afterwards it was proven if the assumed hypothesis, that a preceding warning has a positive effect on the reaction of the driver, could be verified. As for the verification of the previous hypothesis, the contingency table was created for each of the four scenarios, regarding the warning condition and the accident condition. According to this the chi squared values were calculated for each scenario to verify if the two observed features are stochastically independent. As each value is lower than the critical value of the chi squared test of 3.841 it could not be verified that the samples differ significantly. Because of that it could not initially be stated that the provision of a preceding warning has a positive effect on the reaction of the driver. To prove if the calculated value was just received randomly according to the samples, the p-values for each scenario were calculated: Failure of gas (p=0.5), autonomous acceleration (p=0.375), failure of steering (p=0.49) and failure of brake (p=0.304).

CONCLUSIONS

The evaluation of the performed tests has provided interesting results for security related considerations for present and upcoming x-by wire systems. However, a critical discussion of these results is needed.
First of all the test setup does not precisely reflect the feeling of real driving. This is due to the use of a driving simulator and its properties such as the missing movement of the body and limited field of view provided for the test persons. Therefore they may lack the ability to judge the overall traffic situation. Beside this general limitations the wrong interpretation of the failure in our study may have an important effect on the results. It might happen that a test person attributes the failure (e.g. engine stop) rather to the simulation itself than to the vehicle in the simulation. We implemented engine sound and the real dashboard (incl. warning lights) to avoid a misattribution. Additionally we asked the test subjects during (think-aloud test) and after the tests about their thoughts about the failures. All of them reported that they were confident that the failure was a part of the test and in time critical scenarios they did not reason the cause of failure. We also observed some oddities in the behavior of the computer-controlled traffic.

The test track we used in our driving simulator proved to be appropriate for our evaluation as it included all necessary properties of our test scenarios.
The think-aloud-test we used to gather information about feelings and thoughts of the test persons has been discussed in recent literature. One source [16] points out the unnatural process which could change the demands of the task. This could imply for instance that the test persons believe that they have to fulfill certain expectations which could lead to a distortion of the results. Nevertheless it provided appropriate results for the purpose of the evaluation while different variations could be used to reduce such negative impact.
While the amount of tests persons is appropriate for a first test, the results presented here are not statistically provable but reveal tendencies and qualitative results for further studies.

ACKNOWLEDGMENTS

This work was partly supported by German Research Foundation, project ORCHideas (DFG GZ: 863/4-1).

This work was also partly supported by European Research Foundation, project SAVELEC (Safe control of non cooperative vehicles through electromagnetic means, FP7- SEC-2011, Grant Agreement Number 285202).

REFERENCES

[1] Tuchscheerer, S., Dittmann, J., Hoppe, T., and Krems, J. F. Theoretical Analysis of Security Warnings in Vehicles and Design Challenges for the Evaluation of Security Warnings in Virtual Environments. In G. Saake and V. Koeppen (Ed.). IWDE '10 - Proceedings of the First International Workshop on Digital Engineering (pp. 33-37). New York: ACM Digital Libary, 2010

[2] ISO/FDIS 26262:2010(E). Road vehicles - functional safety, 2010. Available at http://www.iso.org/iso/catalogue_detail?csnumber=43464 Accessed November 14, 2014.

[3] Paar, C. Embedded Security in Automobilanwendungen. Elektronik Automotive, 1:1–6, 2004.

[4] Dittmann, J., Hoppe, T., Vielhauer, C. Multimedia Systems as Immune System to Improve Automotive Security? Safecomp 2013 FastAbstract, Toulouse:
France (2013), Hyper Articles Len ligne (HAL), HALid: hal-00926391, URL: http://hal.archivesouvertes.fr/hal-00926391; SAFECOMP 2013, 24.-27. Sept. 2013, Toulouse, France, 2014.

[5] Tobias Hoppe, Stefan Kiltz, Jana Dittmann: Security threats to automotive CAN networks – practical examples and selected short-term countermeasures. In: Computer Safety, Reliability, and Security, Proceedings of the 27th International Conference SAFECOMP 2008, Newcastle, UK, September 2008; Springer LNCS 5219; S. 235-248; Editors: Michael D. Harrison, Mark-Alexander Sujan; ISBN 978-3-540-87697-7, 2008.

[6] Tobias Hoppe, Stefan Kiltz, Jana Dittmann: Automotive IT-Security as a Challenge: Basic Attacks from the Black Box Perspective on the Example of Privacy Threats. In: Computer Safety, Reliability, and Security, Proceedings of the 28th International Conference SAFECOMP 2009, Hamburg, Germany, September 2009; Springer LNCS 5775; S. 145-158; Editors: Bettina Buth, Gerd Rabe, Till Seyfarth; ISBN 978-3-642-04467-0, 2009.

[7] K. Koscher, A. Czeskis, F. Roesner, S. Patel, T. Kohno, S. Checkoway, D. McCoy, B. Kantor, D. Anderson, H. Shacham, S. Savage: Experimental Security Analysis of a Modern Automobile. In: The IEEE Symposium on Security and Privacy, Oakland, CA, May 16-19, 2010.

[8] Charlie Miller, Chris Valasek: Adventures in Automotive Networks and Control Units. In: DEF CON 21 Hacking Conference, 1.-4. August 2013, Las Vegas, USA, White Paper at http://illmatics.com/car_hacking.pdf (last access: 2015-01-20), 2013.

[9] Astragon Fahr-Simulator 2009. 2009. Available at http://www.astragon.de/produktdetails/article/fahr-simulator-2009.html Accessed November 14, 2014.

[10] Savelec-Project. 2012. Available at: http://savelec-project.eu/ Accessed November 14, 2014.

[11] Janni Nielsen, Torkil Clemmensen, and Carsten Yssing. Getting access to what goes on in peoples heads? - Reflections on the think-aloud technique. In Proceedings of the second Nordic conference on Human-computer interaction, pages 101–110, 2002.

[12] Sonja Hoffmann, Hans-Peter Krüger, and Susanne Buld. Avoidance of simulator sickness by training the adaptation to the driving simulation. VDI- Berichte Nr. 1745. Simulation und Simulatoren - Mobilität virtuell gestalten, 1745:p. 385–404, 2003.

[13] Diana Rösler, Anja Naumann, Sven Tuchscheerer, and Josef F. Krems. Warnen im Kraftfahrzeug: Experimentelle Untersuchung zur Detektion und Bewertung optischer und akustischer Signale. In M. Röttig, G. Woszny, A. Klostermann, and J. Huss, editors, Prospektive Gestaltung von Mensch- Technik-Interaktion, pages 209-214. VDI Verlag, 2007.

[14] Jens E. Gayko. Lane departure and lane keeping. Handbook of Intelligent Vehicles, Volume 1:689-708, 2012.

[15] P.E. Greenwood and M.S. Nikulin. A guide to chi-squared testing. Wiley, 1996.

[16] Sharon McDonald, Helen M. Edwards, and Tingting Zhao. Exploring think-alouds in usability testing: An international survey.IEEE Transactions on Professional Communication, Volume 55:2 19, 2012.