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.

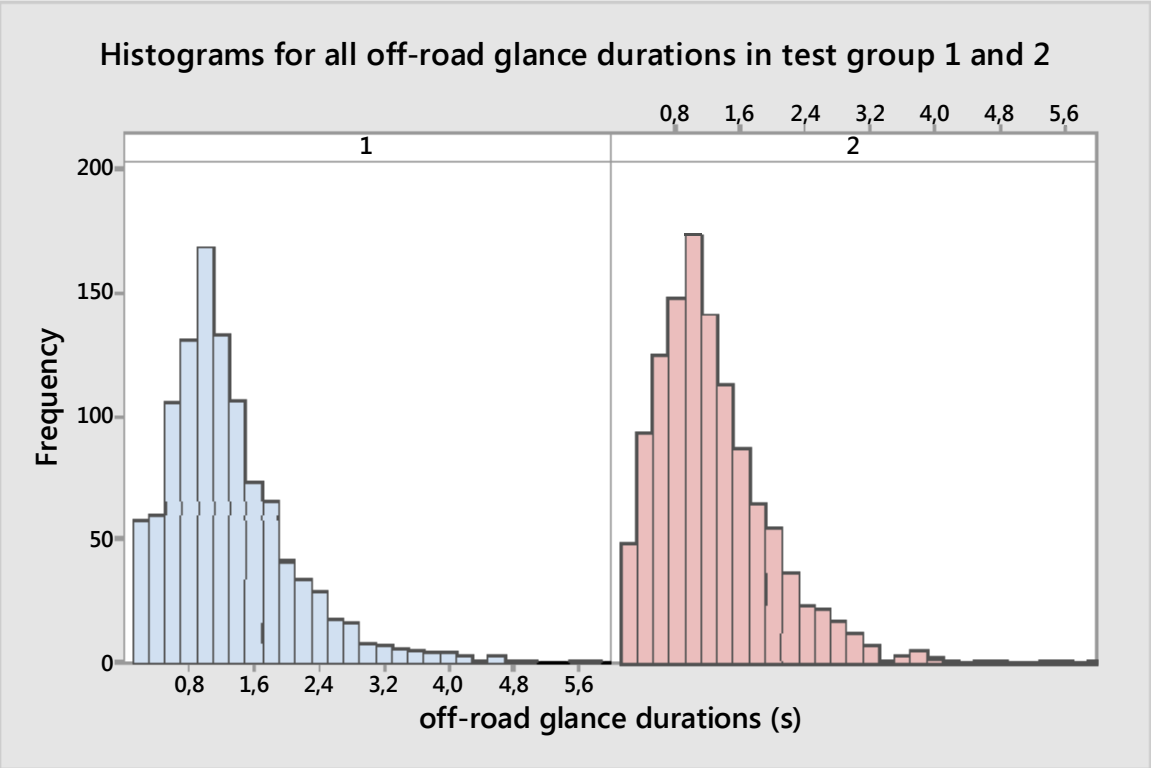

Figure 2: The histograms for the two groups show that they have similar off-road glance duration distributions

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.

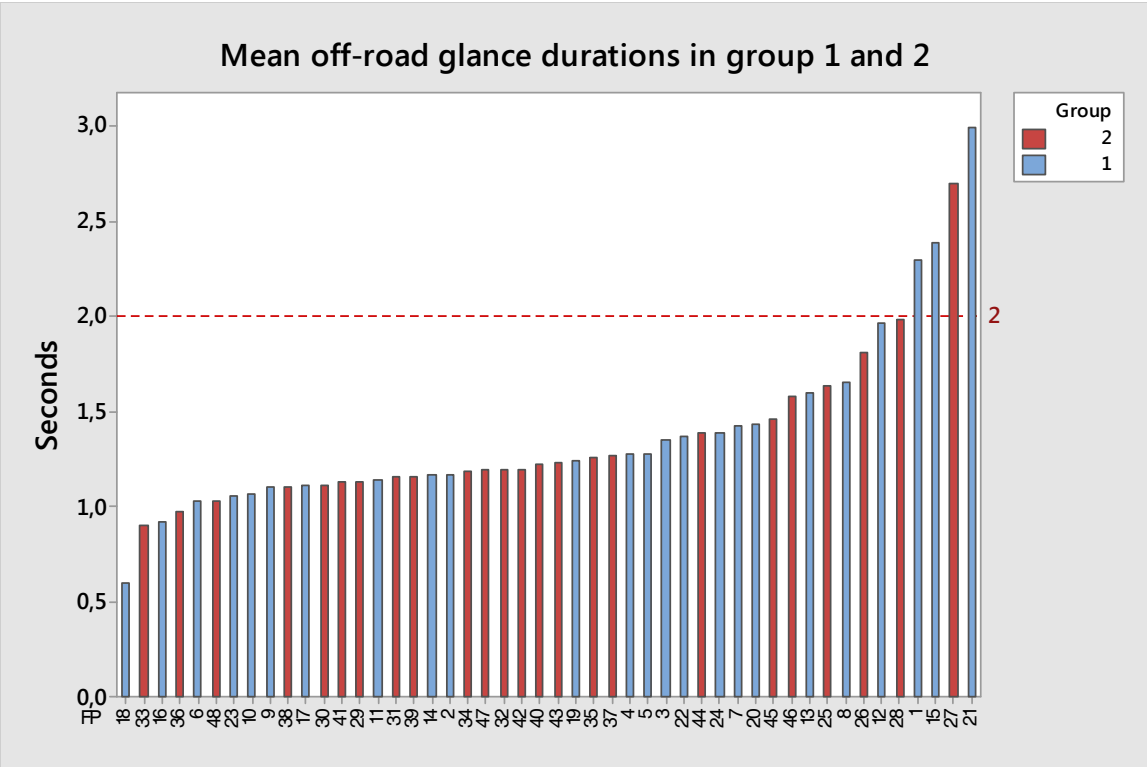

Figure 3: Mean off-road glance durations across all tasks for each test person, ranked according to duration. The reference line indicates the 2.0 s threshold in EGDS.

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.