
Article:
Spyridakos, PD, Merat, N orcid.org/0000-0003-4140-9948, Boer, ER et al. (1 more author) (2020) Behavioural validity of driving simulators for prototype HMI evaluation. IET Intelligent Transport Systems, 14 (6). pp. 601-610. ISSN 1751-956X

https://doi.org/10.1049/iet-its.2018.5589

IET Research Journals

Submission for IET Research Journal Papers

IET Journals
The Institution of Engineering and Technology

Behavioural Validity of Driving Simulators for Prototype HMI Evaluation

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Panagiotis D. Spyridakos[1]*, *Natasha Merat*[1], *Erwin R. Boer*[1], *Gustav M. Markkula*[1]

[1] *Institute for Transport Studies, University of Leeds*
* *E-mail: tspds@leeds.ac.uk*

Abstract: In-vehicle interfaces are now part of the vast majority of production vehicles. Such interfaces need to be thoroughly evaluated to ensure they do not pose any risks to the drivers using them. Driving simulators have extensively been used in such a context, yet their reliability in terms of how realistic a driving behaviour they elicit is still in question. An investigation on driving simulator behavioural validity in the context of prototype HMI evaluation is presented in this paper. Using data collected in a dual setting driving study (driving simulator and real world), as well as results from existing related literature, a comparison between driving behaviour in different types of driving simulators and in reality was carried out, for a variety of behavioural metrics. The results are presented in the form of a "validity matrix" that aggregates the level of behavioural validity different simulator settings can achieve for different behavioural metrics.

1 Introduction

Driving is a complex, multi-tasking activity which requires successful acquisition and coordination of various physical, cognitive, sensory and psychomotor skills [1–4]. However, drivers often engage in parallel, non-driving related tasks concurrently to driving. Since drivers are nowadays used to having certain functionalities available while operating their vehicle, from in-vehicle sound systems to hands-free phone access and internet connectivity [5], those tasks are more often than not associated with an in-vehicle interface. Although in-vehicle interfaces can be useful for a range of activities, related to the primary driving task in different degrees, the work presented here revolves around those that provide information and entertainment functionalities to the driver and are, thus, not directly related to the primary driving task. Throughout this paper, such interfaces will be referred to using the generic term Human Machine Interface (HMI) and the secondary tasks associated with them as Human Machine Interface tasks (HMI tasks).

Different types of HMI tasks, based on modality and difficulty, can have different effects on driving performance, with some HMI tasks having an established negative effect on driving performance and increasing driver risk [e.g. 6–8]. Consequently, it is important that HMIs in on-market vehicles do not pose more than the minimum potential risk to the driver. The attentional demands of prototype HMI designs should be thoroughly evaluated before being fitted into a production vehicle. Evaluation in the real world (such as conducting naturalistic studies and real world testing) can be very costly and time consuming. Using driving simulator studies instead, can speed up the evaluation process and make it more efficient. Driving simulators of all shapes and forms, from simple desktop simulators to full scale immersive machines, have been widely used in Human Factors research to investigate driver behaviour under various settings [e.g. 8–10], with thoroughly documented guidelines on simulator testing procedures and behavioural benchmarks also available to researchers [e.g. 11]. The many advantages of using driving simulators to conduct research, can be summarised into providing a safer driving environment, a richer body of data and large economic savings [e.g. 12, 13]. However, results obtained using driving simulator studies should match real-world driving behaviour and, ideally, have minimal deviation from it.

For the first time, this paper presents results from directly comparing driving performance in a hexapod and fixed base simulator to real world conditions, where the participants engaged in HMI tasks while driving. Moreover, this has also been the first attempt to collectively assess the level to which driving simulators can elicit similar to real world behaviour, for all relevant driving simulator types, across the most commonly used metrics in the context of prototype HMI evaluation.

1.1 Behavioural Validity of Driving Simulators

Driving simulators have predominantly been assessed in terms of their *physical* and *behavioural* validity throughout the relevant literature [12, 14, 15]. Physical validity relates to the degree to which a simulator replicates the corresponding real physical system, focusing mainly on simulator characteristics (e.g. what the simulated vehicle looks like, what the simulated outside world looks like, how the simulated vehicle movement matches that of a real vehicle, etc.). Behavioural validity, on the other hand, relates to the degree to which a driver behaves in a similar manner in a driving simulator as they would under real world conditions. Physical validity has been assumed to increase in advanced simulators, e.g. driving simulators employing motion yield higher physical validity than fixed-base ones [15]. However, higher physical validity does not always improve behavioural validity, hence high physical validity is not always necessary in order to acquire useful information on how drivers behave under different conditions [16]. It has been shown, for example, that increasing the validity of visual displays used in a dual-tasking driving experiment, where drivers had to interact with a cell phone, did not have a significant effect on driving performance [17].

When it comes to evaluating performance in different tasks, it has been argued that behavioural validity is more important than physical, as it is the one that describes the correspondence between what is observed in the simulator and what is observed in the real world setting [14, 18].

Behavioural validity can be further classified into two types; *absolute* and *relative* validity [14]. Absolute behavioural validity implies that dependent variables (e.g. driving performance metrics) take on the same numerical values in a driving simulator as in the real world. Relative behavioural validity was initially introduced as a more qualitative criterion, only requiring differences in the dependent variable between conditions to be of the same order and direction [14]. For example, if two HMI tasks are compared in terms of the time needed to complete them between real world and simulator conditions, Task 1 should consistently rank lower than Task 2 (or vice versa). However, it is most commonly assessed on the basis that the magnitude of the differences has to be the same, too [15, 19, 20]. When relative

validity also requires the magnitude of differences to be the same, then the differences observed across conditions must have the same numerical value. Revisiting the previous example, the difference in completion time between Task 1 and Task 2 should, in this case, be the same for simulator and real world.

1.2 Desired Level of Behavioural Validity

A variety of factors can affect the behavioural validity of a driving simulator, related either to simulator or user characteristics, such as the motion system of the simulator and the demographic characteristics of trial participants [10]. However, there is no set of rules that defines what level of behavioural validity is needed for different tests, as this is highly situation-dependent and relates to the aim and research questions of the study that investigates it [21]. Relative validity has been advocated as sufficient to address many research questions, as most driving studies examine the effect of different conditions on specific driving parameters [16, 17, 22]. If, however, the study aims at directly comparing absolute numerical values of the examined parameter across different conditions, then absolute validity would be the desired level [18].

For example, a manufacturer interested in conducting comparative testing between different prototype interface designs to identify which one of the interfaces could be associated with longer off-road glances, could make that decision with a simulator that can achieve only relative validity. However, if the aim was to determine the exact glance times associated with executing a task on the interface (e.g. to verify compliance with a set of design guidelines), then absolute validity would be needed to ensure that the behaviour observed in the simulator closely matches what would be observed in the real world.

1.3 Aims of the Present Work

This paper provides, for the first time, a more comprehensive overview of the multitude of simulator types and metrics in use in the area of prototype HMI evaluation. The behavioural validity of different driving simulator types was examined with different types of metrics. Results are combined from the analysis of collected data, as well as from an extensive review of related literature, to ensure that the entire range of simulator types used in HMI evaluation related studies is considered with regards to the behavioural validity levels that they can achieve. The results are presented in the form of a matrix that presents the collective behavioural validity level of different simulator types across different behavioural metrics. Such a matrix could potentially be used as a tool by the automotive industry to identify what type of driving simulator would be more appropriate for a given HMI evaluation test or a desired level of behavioural validity.

2 Methods

An extensive review of related literature was performed to collect previous findings regarding driving simulator validity in the context of prototype HMI evaluation and, thus, identify potential research gaps.

Also, a driving study was carried out, with data collection taking place in a driving simulator, and in a real world setting (test track). The study was approved by the University of Leed Research Ethics Committee. The simulator data collection was conducted in the University of Leeds Driving Simulator (UoLDS), and the real world data collection was conducted in the Jaguar Land Rover (JLR) Emissions Circuit test track in Gaydon, Warwickshire.

2.1 Review of Related Literature

A comprehensive search for related publications was conducted initially over a three month period, from November to January 2016 and was, subsequently, periodically revisited over the next two years, until November 2018. The main techniques used to ensure all relevant references were obtained and reviewed, were search on literature databases (Scopus and Google Scholar), review of reference lists of other relevant publications and review of publications that referenced those (as relevant publication were the ones reporting results of HMI performance studies).

An initial search against publication titles, abstracts and keywords was made on Scopus, yielding 252 results. An initial cleaning was performed based on the relevance of the title and, for the articles that remained, their abstracts were reviewed to validate they were suitable.

The inclusion criteria used to define whether a publication was relevant to the review or not had to be all met and were the following: The reported studies had experiments performed both in a driving simulator and in a real world setting, and the experiments were focused on HMI task execution concurrently to driving.

The above selection process resulted in the following two publications: [10, 16]. Together with [9, 23, 24], which were relevant and previously known to the author form different literature searches, an initial body of five publications was formed. Reviewing the references therein and searching for other publications citing them, the following and final body of ten papers was formed, that was used for the review presented below: [9, 10, 16, 17, 23–28].

Finally, this work was presented on the the 6^{th} International Conference on Driver Distraction and Inattention, where fellow academics were requested to provide feedback regarding related research that might have not already been included. This, however, did not lead to the discovery of any further articles that fit the aforementioned inclusion criteria.

2.2 Participants

A total of 12 participants completed the UoLDS data collection, six of which were females (mean age 37.17 ± 10.42 years). One of the initial participants experienced simulator sickness symptoms and was replaced by a new participant of similar demographics. Potential participants for the simulator study were contacted through the simulator participant database or through the University of Leeds mailing lists. The biggest response came from people in the database, hence all but two participants had prior experience with the simulator. The participants were compensated with $\pounds 15$ for their time.

A total of 11 participants completed the Gaydon data collection, two of which were females (mean age 36.55 ± 11.93 years). Initially, 12 participants were also recruited for the Gaydon data collection, too, but one was unable to attend and could not be replaced due to time restrictions. Potential participants for the Gaydon experiment were contacted through the internal JLR communication network. About half of the participants had previously driven in the test track. None of the participants were in any way involved in the development or evaluation of prototype HMI designs as part of their job specification in JLR. The participants took part in the data collection during their normal working hours, and were not otherwise compensated monetarily for their time.

2.3 Materials

The UoLDS consists of a 4 m diameter spherical projection dome, mounted on an eight-degree-of-freedom moving base. The projection dome provides a 300° field-of-view using a high definition projection system and houses the simulator vehicle cab, a 2005 Jaguar S-type cab with all driver controls operational. The vehicle dynamics model employed for the study was a real-time SimPack model of a Jaguar XF (programme denomination X250).

Two different motion configurations were tested in the simulator during the present study; a setting with no motion (fixed base), where the vehicle handling feedback was provided to the driver through the simulator visual scenery and the steering torque of the vehicle model. This has been the most commonly used motion simulator configuration in the relevant literature. Since hexapod only motion had not been previously used in this context, such a motion configuration, where the simulator dome was moving using the 6 degrees of freedom hexapod, was also used here. The hexapod supplied roll,

Fig. 1: Proving Ground facilities layout in Gaydon.

pitch and yaw movements, providing the drivers with motion cues for perceiving acceleration.

Vehicle handling and sensor data were recorded through the built-in simulator CAN Bus at a frequency of 60 Hz. Eye-tracking data were recorded (at a similar frequency of 60 Hz) using a v5 Seeing Machines faceLAB eye-tracker, mounted on the dashboard of the simulator vehicle cab. Finally, video streams were recorded through 4 cameras. The recorded video streams had timestamps synchronised with the logging timestamps of the simulator, thus making it easier to extract data segments of interest based on video evidence or refer to the corresponding video segment from simulator data.

The subject car used during the Gaydon experiments was a Range Rover Evoque, fully functional and as in circulation. Vehicle data were recorded from the vehicle CAN using VBOX by Racelogic at a frequency of 60 Hz. Eye-tracking data were recorded using eye-tracking glasses by SMI, at a frequency of 30 Hz. Finally, video streams were recorded through the eye-tracking glasses camera (located at the binocular focal point) and through 3 VBOX cameras. Video streams from the different VBOX cameras were synchronised and timestamped.

2.4 *Driving Environment*

The driving environment that was used for the study was the Emissions Circuit in JLR's Proving Ground test track in Gaydon, Warwickshire, UK. The circuit consists of two straight segments, connected with two elongated curved segments, and has four lanes in a single carriageway configuration. Figure 1 provides an illustration of the test track layout.

A digital replica of the Emissions Circuit test track was created for the UoLDS, preserving all design characteristics of the test track, barring the scenery which was simplified.

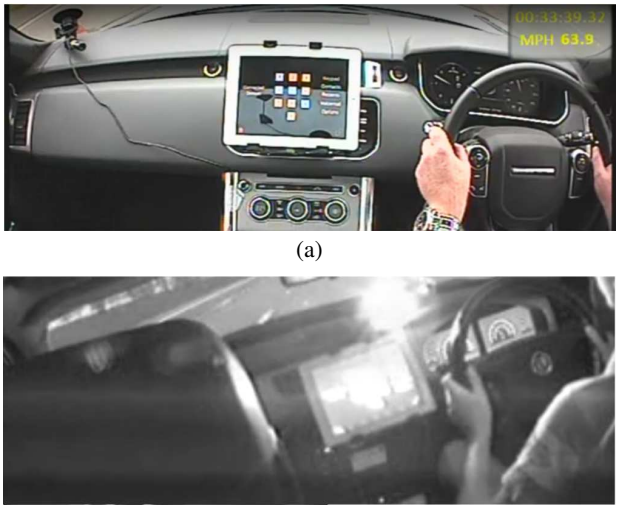
2.5 *Driving Scenarios*

Two different scenarios were tested in both experiments, where a lead vehicle was used. The different scenarios corresponded to different speed profiles for the lead vehicle. In the first scenario, the lead vehicle was travelling at a constant speed of 50 mph, while in the second scenario, the lead vehicle was travelling at a varying speed between 60 and 70 mph, following a semi-randomised speed profile.

2.6 *HMI tasks*

Three visual-manual HMI tasks were used in both experiments of this study. The HMI tasks were implemented so that each one had a different number of interactions, as well as varying in types of interactions that were required to be completed.

Participants were thoroughly trained on how to perform each HMI task both while stationary and while driving. The tasks were implemented as an interactive mobile application that resembled the design of a prototype HMI designed by JLR. An iPad model 2 was used as the HMI and was temporarily mounted on the central console of the vehicle, with its top part aligned with the top arch of the steering wheel. The tablet was mounted approximately to the left of the driver (with an approximate distance of 20 cm between the left edge of the steering wheel and the centre of the tablet) and was tilted 35° back from the vertical. Figure 2 illustrates the tablet configuration and placement within each text vehicle.



(a)

(b)

Fig. 2: HMI iPad tablet tablet placement during the data collection experiments. The top panel (a) corresponds to the Gaydon experiment, while the bottom panel (b) corresponds to the UoLDS experiment.

The HMI tasks, although emulating functionalities one might find in a production vehicle, were independent of any vehicle system and, thus, had no effect in any of its functionalities. For the first task, which emulated massage activation for the driver seat, the drivers had to perform three "Press" interactions. For the second task, which emulated calling a contact from their favourite contacts list, the drivers had to perform four "Press" interactions. Finally, for the third task, which emulating playing a song from their song library, the drivers had to perform two "Press" interactions, followed by a "Scroll" interaction and a further two "Press" interactions.

The tasks were classified as "easy", "medium" and "hard", based on the expected difficulty and complexity level arising from the required interactions, i.e. from the description above, the first, second and third tasks were classified as "easy", "medium" and "hard", respectively. This classification was only used to clearly distinguish the tasks and, despite the terms used, was not meant to make a strong statement about the actual difficulty levels of the tasks. However, after the completion of the experiments, the participants' perceived task difficulty (obtained through relevant subjective questionnaires) appeared to be in agreement with our initial classification.

2.7 *Experimental Design*

A three-factorial design was used, with Environment being a between-subjects factor, while Scenario and HMI task were within-subjects factors. Environment had three levels: Real, Fixed Base and Hexapod. Scenario had two levels: Constant speed and Varying speed. HMI task had three levels: Easy, Medium and Hard.

An additional factor, Road, was originally considered, consisting of two levels; Straight and Curve. Unfortunately, due to safety regulations, participants were prohibited from performing HMI tasks while driving on a curve during the Gaydon experiment. Given that the UoLDS experiment took place before the Gaydon one and that there was no previous knowledge of the aforementioned restriction, only the UoLDS participants were exposed to the Curve level.

As a result, the UoLDS participants were exposed to a total of 24 unique conditions (2 Environment $\times$ 2 Scenario $\times$ 3 Task $\times$ 2 Road), while the Gaydon participants were exposed to a total of 6 unique conditions (1 Environment $\times$ 2 Scenario $\times$ 3 Task $\times$ 1 Road).

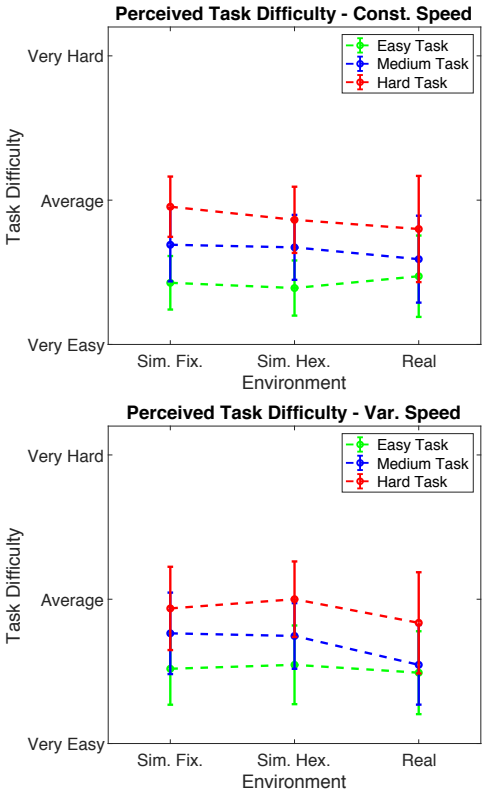


Fig. 3: Subjectively perceived task difficulty for all HMI tasks, across all conditions. Participants scored task difficulty in a 0-100% scale, translated here in a nominal scale where "Very Easy" corresponds to a 0% score, "Average" corresponds to a 50% score and "Very Hard" corresponds to a 100% score.

2.8 *Procedure*

The procedure followed throughout data collection was the same for the Gaydon and the UoLDS experiment, except for minor differences dictated by the environment itself.

For the UoLDS experiment, the participants were briefed, provided informed consent, and trained on the HMI task on the iPad, first outside of the simulator. At this point, they would repeat each task as many times as necessary, until they could confidently declare that they knew how to complete it. The participants were then given a short questionnaire regarding the perceived difficulty of each HMI task. Next, participants took a familiarisation drive with the simulator in full motion (and a normal drive in the test track). This drive aimed at participants acquainting themselves both with the simulator and the concurrent driving and secondary tasks. Next, during the data collection phase, the experimenter was sitting in the back seat of the vehicle, behind the driver. Initially, a full lap of the simulated test track was performed with the participant only driving and not performing any HMI tasks (baseline drive). Then, three full laps of the simulated test track were driven, during which the participants performed the HMI tasks on various instances, when instructed by the experimenter. The experimenter would denote those instanced by saying "Engage now" and, after the participants completed the HMI task, they should indicate so by saying "Done". The experimenter only instructed participants to engage when they were in full control of the vehicle and at least 3 seconds after a previous HMI task execution. This resulted to a high number of task repetitions for each participant. These four laps of the test track constituted a drive, with each drive lasting approximately 25 minutes and corresponding to one Environment $\times$ Scenario combination. Consequently, two drives were required per participant in the real world and four drives in the simulator.

After the completion of each drive, the participants completed a set of questionnaires regarding their subjective experience of the HMI tasks and simulator, where applicable. Since the present paper focuses on the investigation of objective metrics, these questionnaires will not be further reported on here.

As the participants were already "experts" in performing all HMI tasks when moving into data collection, there was no expectations of learning effects becoming evident through repetition. Moreover, since the main focus of the experiments was to investigate differences between simulator settings, only the motion settings in the simulator and the driving scenarios were counterbalanced.

2.9 *Measures*

Based on past research, three types of behavioural metrics were investigated when analysing the collected data: HMI task related measures (namely, task completion time), driver performance measures (namely, mean speed, speed variability and steering wheel reversal rates) and glance behaviour metrics (namely, off-road glance frequency, total off-road glance duration and mean off-road glance duration).

2.10 *Initial Data Reduction*

For both the real world and simulator visual scenes, three major Areas Of Interest (AOIs) were considered: road ahead (all points of the visual scene that intersected the wind shield when the driver was looking through it, focused ahead and without moving their head), HMI and Other. Consecutive fixation points within the same AOI were aggregated to yield glance duration times. Only glances falling within the road ahead AOI (on-road glances) and HMI (off-road glances) were considered for analysis.

Eye-tracking data from the real world were manually annotated, frame by frame, using the BeGaze analysis software. A 2-D model of the driving environment was created and the AOIs discussed above were defined within it. Next, for each fixation point, its location was mapped within one of the AOIs.

Regarding the simulator eye-tracking data, the FaceLab eye-tracker logs eye yaw and pitch based on an initial calibration. Consequently, there is no pre-defined model of the world and fixations points cannot directly be assigned to AOIs. To identify AOIs in the visual scene, fixation points for each driver were visualised and compared between baseline driving and HMI execution intervals. Later, a random sample of task segments was visually compared against video data to ensure the AOIs were properly defined.

Initially, all instances where the drivers made a mistake during the HMI task execution (either due to performing an incorrect action in the context of the HMI task or due to an issue with the HMI itself) were removed from the dataset. The data loss from this operation was minimal, amounting to less than 1% of the total recorded data.

Regarding minimum glance duration, there is currently no agreement in the academic community as to what threshold should be adopted. Salvucci and Goldberg, for instance, defined the minimum required glance duration at 100 ms [35], while Land found the shortest fixation durations to average at 150 ms [38]. For the purposes of this paper and in line with other previously published studies (both from the authors' research group and outside of it - [e.g. 36, 37]) a minimum duration of 200 ms was required for an aggregation of visual data points to be considered as a glance and be included in the analysis.

Moreover, since both eye-tracking systems automatically classify fixation points based on their quality, glances containing more than 50% bad quality fixation points were also not used.

Finally, after the manual annotation of the simulator data, one of the participants was identified to be likely experiencing symptoms of motion sickness, which they had not disclosed to the experimenter at the time of the trial. The conclusion was reached through the recorded video observation, where the participant was found to be drowsy and experiencing what appeared as xerostomia (dry mouth). All analyses were run both with and without that participant's data included and no differences in significance levels were observed in the reported results. Hence, their data were removed to ensure only

valid interactions are represented, reducing the sample size of the UoLDS experiment to 11 participants.

The final dataset after reduction included a total of 617 HMI task executions from the real-world condition, and 1328 from the simulator (with the latter almost evenly split across simulator settings and scenarios).

3 Results

Our initial analyses included the scenario type (constant versus variable speed) as an independent variable, but the conclusions with respect to simulator validity were the same regardless of scenario. Therefore, for increased clarity and readability, we are here only presenting results for the constant speed scenario, which is more closely aligned with the types of scenarios used in previous research on simulator validity for HMI prototype evaluation [e.g. 10, 16, 24]. The dataset to which the results that are presented here correspond consisted of 369 HMI task executions in the real world, 365 task executions in the fixed base simulator and 359 task executions in the hexapod simulator.

Similar to [10], the technique used to analyse the data was linear mixed effects modelling [29]. Linear mixed effects models take into account the hierarchical structure of the data, which makes them very well-suited for modelling and analysis of repeated measures experiments. Moreover, linear mixed effects models are very robust when it comes to handling missing data, which makes them invaluable in cases of unbalanced studies.

Models were fitted for each metric using MATLAB and the built-in fitlme function. In line with the modelling approach taken in [10], the models included a fully varying slope and intercept per participant, i.e. the maximal random effects structure justified by the data, as suggested in [30]. After an initial fitting, the residuals of the models were visually inspected to identify whether their distribution approached normality. Where the normality assumption was violated, data were log-transformed and models were refitted. No treatment was taken for outliers since the data were already cleaned and all observations were valid.

Following the paradigm in [16] (also adopted in[10]), the following approach was adopted to conclude the level of behavioural validity for each simulator setting and metric:

- Relative validity was established when the ranking of the HMI tasks and their main effect were consistent across conditions (i.e. no interaction effect of Environment $\times$ Task observed).
- Absolute validity was established by the presence of relative validity and the absence of a main effect of environment.

3.1 Task Completion Time

Task completion time was defined as the time elapsed from the moment the experimenter instructed an HMI task execution initiation, until the moment when the participant returned their gaze to the the road ahead, after completing the task. Figure 4 illustrates the average task completion values for the three HMI tasks across all environments.

A significant main effect of task was observed ($F(2, 2048) = 47.39$, $p < .001$). Although completion times were slightly higher in the real world, that difference was not statistically significant, in other words, there was no main effect of environment observed ($F(2, 2048) = 0.46$, $p = 0.63$).

Given the identical ordering of tasks across all three environments (from shortest completion time to largest: easy, medium, hard) and the absence of an effect of environment, absolute validity can be concluded for both the fixed base and the hexapod simulator.

3.2 Frequency of Off-road Glances

Off-road glance frequency was define as the number of glances the drivers employed towards the interface during a task execution. Figure 5 illustrates the average number of glances needed for

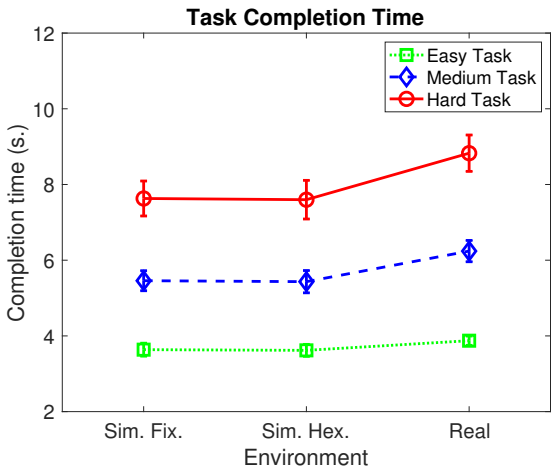


Fig. 4: Task completion times.

each of the three HMI tasks across all environments. A significant main effect of task was observed ($F(2, 1969) = 38.22$, $p < .001$), while there was no effect of environment ($F(2, 1969) = 0.09$, $p = 0.918$). The task ordering is the same in all environments (from the one requiring the fewest glances to the one requiring the most - easy, medium, hard). The absence of an effect of environment, along with the consistent ranking of task across environments, indicates that absolute behavioural validity can be concluded for both the fixed base and the hexapod simulators.

3.3 Total Off-road Glance Duration

This metric was calculated as the aggregate duration of all glances towards the HMI during a task execution. A significant main effect of task was observed ($F(2, 1834) = 49.39$, $p < .001$), while no effect of environment was found ($F(2, 1834) = 1.64$, $p = 0.195$).

Glance times for the medium and hard tasks are almost identical in the fixed base simulator, while the two tasks differ more noticeably in the hexapod and real world (see Figure 6). Hence, their relative ordering cannot be considered in this case. However, since the easy task ranks lower against both the medium and the hard task across all environments, the possibility of relative validity can be concluded for both the fixed base and hexapod simulators.

At this point, it would be interesting to also investigate the observed behaviour through the relevant NHTSA guidelines [11], as an additional consideration, since the use of an overly complex HMI task that would consistently fail the prescribed benchmarks, could potentially render any behavioural validity conclusion inaccurate. Here, no task execution required more than a total of 12 seconds

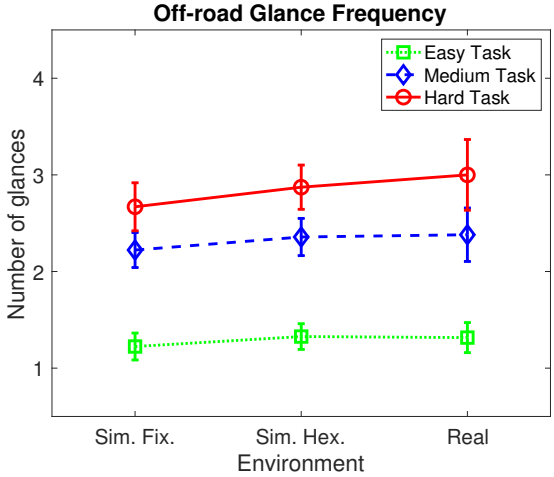


Fig. 5: Off-road glance frequency.

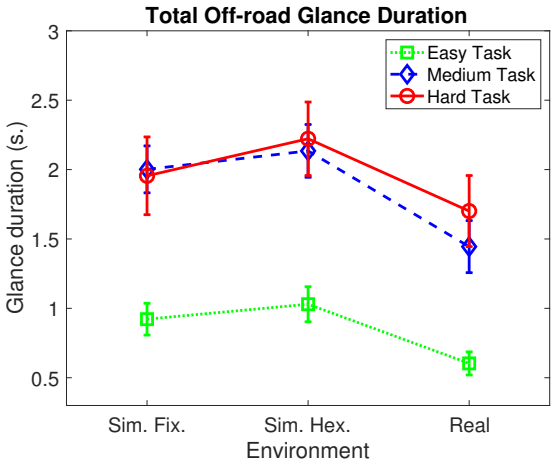


Fig. 6: Total off-road glance duration

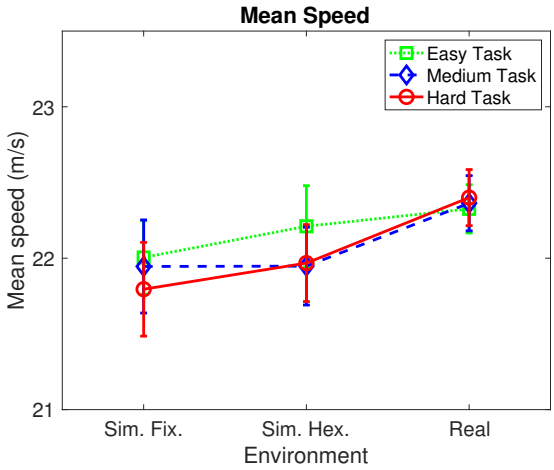


Fig. 8: Mean speed.

of visual attention focused on the HMI, for the corresponding task to be completed successfully .

3.4 Mean Off-road Glance Duration

Mean glance duration was defined as the average duration of all the glances towards the HMI in a single task execution. A significant main effect of task was observed ($F(2, 1834) = 15.13$, $p < .001$), with the medium task requiring the longest glances on average and the easy task requiring the shortest (see Figure 7). Participants in all setting were often observed to look on the road ahead while performing the scrolling action of the hard task, while the easy task only required brief glances towards the HMI. Consequently, the medium task became the most visually demanding one to perform, in terms of average off-road glance duration. This ranking of tasks was consistent across all environments, with relative differences being consistent, too, as indicated by the absence of an interaction effect between environment and task ($F(4, 1834) = 0.71$, $p = 0.585$). Finally, a marginally significant effect of environment was observed ($F(2, 1834) = 3.42$, $p = 0.033$), which points towards concluding relative validity with a possibility of absolute validity for both the fixed base and the hexapod simulator settings.

Regarding compliance to the NHTSA guidelines, it is required that for 85% of test participants, the mean duration of all their individual eye glances towards the HMI, while performing a secondary task, be less than 2 seconds. In this case, for the simulator setting 8 out of the 11 participants employed at least one glance but not more than 10% of the total number of their glances towards the HMI, that lasted more than 2 seconds. One of the participant had 43% of their glances towards the HMI lasting more than 2 seconds. In the real world on the other hand, 8 out of the 11 participants used no glances towards the HMI, while the remaining 3 looked away from the road for more than 2 seconds less than 15% of the time.

3.5 Mean Speed

This metric was defined as the average vehicle speed during HMI task executions. Average speed was marginally higher for the easy task and in the real world, yet no significant effect of task or environment were observed ($F(2, 2007) = 0.48$, $p = 0.48$ and $F(2, 2007) = 0.58$, $p = 0.559$, respectively). As can be seen in Figure 8, there is no consistent ranking of tasks across conditions, but since the differences between tasks are so minimal, ranking would not be meaningful, regardless. Consequently, it is difficult to make a conclusion about absolute and relative validity, given the small differences between the tasks. However, since no main effect of environment was observed, the possibility of absolute validity could be argued in this case.

3.6 Speed Variability

This metric was calculated as the standard deviation of vehicle speed during HMI task executions. Figure 9 points to speed variability being significantly higher in the simulator conditions, manifested through a significant effect of environment($F(2, 2007) = 85.2$, $p < .001$). Since no effect of task was observed ($F(2, 2007) = 0.2$, $p = 0.821$) ,and since ordering could not be considered, no level of behavioural validity can be concluded for either simulator setting.

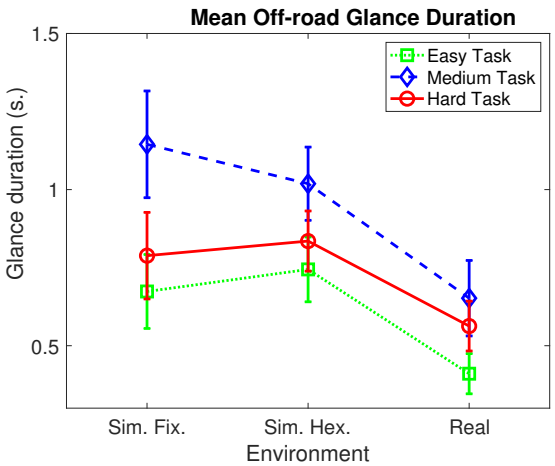


Fig. 7: Mean off-road glance duration

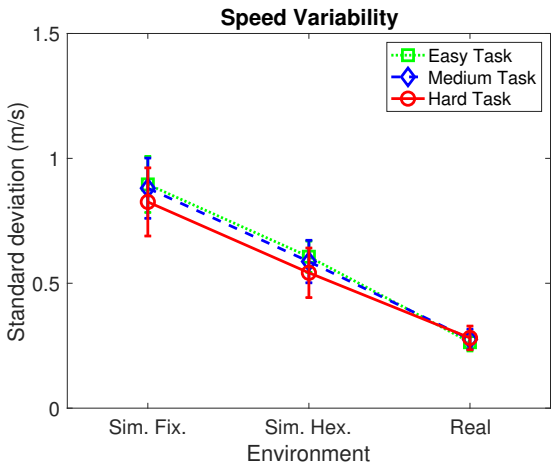


Fig. 9: Speed variability.

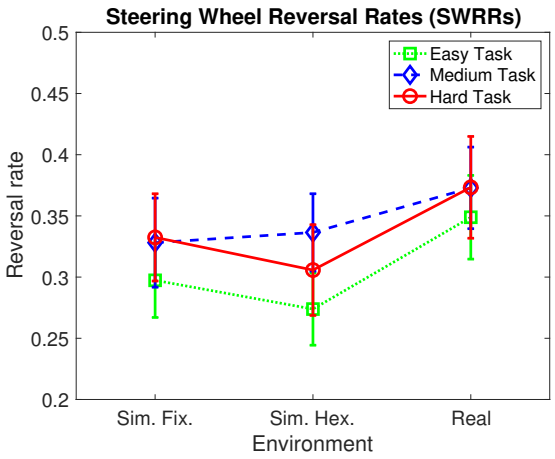


Fig. 10: Steering Wheel Reversal Rates (SWRRs).

3.7 Steering Wheel Reversal Rates

Steering whel reversal rates (SWRRs) were calculated for each task execution using the method by [31], for gap sizes of 1, 5 and 10 degrees. Since there were no differences in reported significance levels between the three gap sizes, only results of 1^o reversal rates are presented here.

As illustrated in Figure 10, the difference of 1 degree reversals between the simulator and reality were not significant, as manifested by the absence of a main effect of environment ($F(2, 2007) = 0.94$, $p = 0.392$). Moreover, no significant effect of task was observed ($F(2, 2007) = 1.12$, $p = 0.326$), thus, rendering task ordering uninformative. Consequently, relative validity could not be concluded for either simulator setting. However, given the absence of effect of environment, if relative validity were to be concluded, absolute validity could also be concluded on those grounds.

4 Discussion and Conclusions

A key objective of this work has been to provide an overview of the multitude of simulator types and metrics in use for prototype HMI evaluation, and their respective validities. This will be addressed in a first subsection below. Thereafter, two additional subsections will discuss limitations of the present work, as well as potential future work.

4.1 The Behavioural Validity Matrix

In order to evaluate the potential of driving simulators as a tool for prototype HMI evaluation, it is important to identify the degree of behavioural validity they can achieve, i.e. to what extent they are eliciting the same driving behaviour as what would be observed in real world conditions. Behavioural validity can be classified as absolute (when performance metrics have the same values for each task in reality and simulator) or relative (when performance metrics have the same relative differences between each task for reality and simulator). Analysing collected data from a driving study conducted both in the real world and in a driving simulator, and combining them with previously published results, a matrix was created that can provide insights on the level of behavioural validity a certain driving simulator type can achieve for different behavioural metrics, in the context of prototype HMI evaluation.

For the papers discussed in Section 2.1, when behavioural validity was directly investigated by the authors [i.e. 10, 16], their conclusions were used directly here (since they used the same approach as here). For those papers where only analysis results were reported [e.g. 9, 24], the level of behavioural validity was inferred by examining statistical significance scores and task rankings across conditions. Table 1 illustrates a summary of the different levels of behavioural validity concluded during the literature review.

Relative (R) and absolute (A) validity were concluded using the same approach as in the collected data analysis (see introduction of Section 3). When a certain level of validity could not be directly concluded due to insufficiently reported statistical analysis results, possibly relative (PR) or possibly absolute (PA) validity was reported where other evidence was present. Finally, when no level of behavioural validity could be established, the corresponding fields were marked with N/A.

One of the most common methods for simple HMI Evaluation regarding task completion times and gaze behaviour, the occlusion test, was added to the behavioural validity matrix, although it is not strictly a simulator configuration. However, it has been demonstrated that occlusion can effectively be used to predict such metrics with relative validity (see [32] and [33] for an overview and further references to earlier literature).

The Behavioural Validity Matrix (see Table 2) was filled by combining the results obtained from the present analysis as well as from the review of relevant literature. In this case, the fields noted as "Possibly" Absolute or Relative valid, refer to either an inconclusive result in literature or to a "logical guess" based on existing results from simpler and/or more complex simulator configurations. The matrix was filled using a simple colour-coding scheme, presented in the top panel of Table 2.

Evidently, there is no single answer as to which simulator should be used for HMI evaluation. The same simulator can achieve different levels of behavioural validity across different behavioural metrics, while also for the same metric, different simulator settings provide a different level of behavioural validity. Instead, the exact purpose of the study should dictate the level of behavioural validity needed and, therefore, drive the decision of which simulator setting should be used in each case.

Testing task completion times or glance duration metrics, for instance, can be achieved with relative validity with a simple desktop simulator or with the even simpler occlusion method. For the same tests, absolute validity can easily be achieved with no motion system, using a static cabin simulator with narrow field projection. Steering control behaviour, quantified through SWRR, can be captured with relative validity in a wide field, fixed base simulator, while there is no setting that can currently achieve absolute validity. Finally, when longitudinal control is in question (usually tested through speed or HW variations) it has been found that no setting achieves absolute validity, although the validity significantly improves when using a full motion simulator.

At this point, it is important to note that, although the behavioural validity matrix in its current form can help researchers and human factors specialists decide which simulator setting is appropriate based on the evaluated metric and level of behavioural validity required, further tests still need to be carried out to verify the collectively concluded validity levels. Given the heterogeneity of the studies reviewed here (different HMI tasks and analysis methods), a more thorough and consistent investigation of driving simulator behavioural validity would help establish them even further as HMI evaluation tools and use them more efficiently.

Moreover, there are cases where statistical significance alone could be questioned as a measure of behavioural validity, especially in small sample sizes. This is due to the fact that absence of evidence for an effect does not necessarily provide solid evidence for the absence of an effect. Consequently, in cases like this, it is important to also consider additional metrics (e.g. effect sizes) to more accurately quantify the magnitude of the difference between conditions.

4.2 Limitations

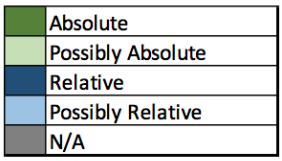
As every piece of research, the present one also comes with its limitations. Initially, some of the experimental choices might appear sub-optimal, such as the participants sample size. Comparing driver behaviour in simulator against real world conditions would ideally require a within-subjects design, i.e. the same participants to be used in both settings. For a between-subjects design as the one used here, it could be argued that a larger sample size would be

Reference	Metrics	Simulator Type						
		Occlusion	Desktop	Cabin with narrow field projection	Cabin with wide field projection	Hexapod	Hexapod and lateral motion	Hexapod and longitudinal motion
Bach et. al 2008	Number of Off-Road Glances, Number of Off-Road Glances > 2 sec.		PA					
	Task Completion Time		A					
	Long. Control Errors, Lat. Control Errors, Interaction Erros		R					
Baumann et. al 2004	Total Task Time, Mean Error	A						
Engström et. al 2005	Task Correct Reponses, Speed (Mean & SD), SWRRs (1 deg.)				R		R	
	SDLP, TLC minima, LANEX				PR		PR	
	Skin Conductance, Heart Rate				N/A		R	
	Self Reported Driving Performance				A		A	
Klüver et. al 2016	SDLP		R	R	R		A	A
	SD Headway		R	R	R		R	R
	Task Completion Time		R	R	A		A	A
Knapper et. al 2015	Speed (Mean & SD)	R						
Petitt et. al 2006	Total Off Roaf Glance Duration, Task Completion Time	R						
Reed et. al 1999	Lane Position, Steering Wheel Angle, ThrottlePosition			R				
	Speed			A				
Santos et. al 2005	Speed, HW (Mean & SD), SDLP, LANEX, SWRRs		PR		PR			
	Self Rep. Performance		PA		A			
	Response Time		N/A		A			
Victor et. al 2005	Mean Glance Dur., Perc. Glances > 2 sec., SD Glance Duration, Glance Freq., Total Glance Dur.				PR			
	Percent Road Centre				A			
Wang et. al 2010	Initial Response Time, Mean Task Duration, Total Glance Time, Percent Eyes Forward, SD Forward Velocity			A				
	Glance Frequency			R				
	Mean Forward Velocity, SDLP			PA				

Table 1: A summary of the levels of behavioural validity that can be concluded for different simulator settings, across metrics that are relevant in the context of HMI evaluation. **A** refers to absolute behavioural validity, **R** refers to relative behavioural validity, **PA** refers to possibly absolute behavioural validity, **PR** refers to possibly relative behavioural validity, while **N/A** was used for the cases where no level of behavioural validity could be concluded, even with additional assumptions.

more appropriate to eliminate as much as possible the effect of individual differences (as the one used, for instance in [10]). In the study presented here, the sample size was relatively limited (eleven participants in the real world and in the simulator, respectively).

An additional issue that arises from the limited number of participants in the UoLDS experiment is relevant to the condition counterbalancing. Given that there were a total of 4 different combinations of simulator motion setting and driving scenario in the



Test Target	Typical Metrics	Simulator Type: Occlusion	Desktop	Cabin with narrow field projection	Cabin with wide field projection	Hexapod	Hexapod and lateral motion	Hexapod and longitudinal motion
HMI Task Execution	Task completion time							
Gaze behaviour	Total off-road glance duration							
	Mean off-road glance duration							
	Off-road glance frequency							
Longitudinal Control	Speed - StD.							
	Speed - Mean							
Steering Control	SWRR							

Table 2: The Behavioural Validity Matrix, based on analysis of the obtained data and existing results from literature. The red border indicates the direct contribution of the present work to the matrix.

simulator, a total of 24 participants would be needed for full counterbalancing. Moreover, the absence of counterbalancing for the tasks could have also created some ordering effects in the collected data.

However, given that the results presented here, along with the conclusions drawn thereafter are in agreement with the existing research, it can be argued that the aforementioned issues did not have a damaging effect on this study.

An additional limitation of the presented results would be related to the behavioural validity matrix, itself. Given the heterogeneity of the studies used to construct it, considering everything together when assessing the level of behavioural validity of different simulator types can only happen under certain assumptions. In particular, considering the differences in experimental design and analysis methodologies, as well as the materials used (from the HMIs and their tasks to the actual simulators) between the existing studies, it could be argued that the various results are not directly comparable.

In this case, too, however, the existing literature seems to be in agreement, which enhances the motivation for the approach taken here.

It is important that such issues are addressed in similar future studies, to ensure results and consequent inferences are valid and reliable.

4.3 Future Work

In further understanding and accurately reporting the level of behavioural validity that can be achieved by different driving simulators in the context of HMI evaluation, a more thorough meta-analysis of already published studies is still required. There, experimental scenarios and types of HMI tasks used should be compared to expand the current behavioural validity matrix into more dimensions and provide more detailed guidelines as to which simulator should be used under which conditions.

The current body of related published research is rather heterogenous, since different experimental and analysis methodologies have been used. Consequently, additional assessment of the published results is required to determine if and to what extent they can be used to infer the behavioural validity level of different driving simulators.

As a next step, additional metrics that have been widely used in distraction studies (e.g. reaction times under dual-tasking conditions) could potentially be considered in the context of driving simulator behavioural validity evaluation. However, such tasks are not always easy to replicate and test in real world conditions due to various safety implications.

Finally, as all vehicle related technology advances, new HMI types, as well as new driving simulator types need to be evaluated in such context to ensure the behavioural validity assessments remain up-to-date and relevant.

5 Acknowledgements

This work was supported by Jaguar Land Rover and the UK-EPSRC grant EP/K014145/1 as part of the jointly funded Programme for Simulation Innovation (PSI).

6 References

1 Hedlund, J., Simpson, H.M., Mayhew, D.R. 'International conference on distracted driving: Summary of proceedings and recommendations: October 2-5, 2005'. (CAA, 2006).

2 Young, K., Regan, M., Hammer, M.: 'Driver distraction: A review of the literature', in: 'Distracted driving', 2007, pp. 379–405

3 Regan, M.A., Lee, J.D., Young, K.: 'Driver distraction: Theory, effects, and mitigation'. (CRC Press, 2008)

4 Groeger, J.A.: 'Understanding driving: Applying cognitive psychology to a complex everyday task'. (Routledge, 2013)

5 Meixner, G., Häcker, C., Decker, B., Gerlach, S., Hess, A., Holl, K., et al. 'Retrospective and future automotive infotainment systemsâĂŤ100 years of user interface evolution'. In: Automotive User Interfaces. (Springer, 2017. pp. 3–53

6 Lee, J.D., Young, K.L., Regan, M.A.: 'Defining driver distraction', *Driver distraction: Theory, effects, and mitigation*, 2008, **13**, (4), pp. 31–40

7 Fitch, G.M., Soccolich, S.A., Guo, F., McClafferty, J., Fang, Y., Olson, R.L., et al. 'The impact of hand-held and hands-free cell phone use on driving performance and safety-critical event risk'. (, 2013.

8 Kountouriotis, G.K., Merat, N.: 'Leading to distraction: Driver distraction, lead car, and road environment', *Accident Analysis & Prevention*, 2016, **89**, pp. 22–30

9 Engström, J., Johansson, E., Östlund, J.: 'Effects of visual and cognitive load in real and simulated motorway driving', *Transportation Research Part F: Traffic Psychology and Behaviour*, 2005, **8**, (2), pp. 97–120

10 Klüver, M., Herrigel, C., Heinrich, C., Schöner, H.P., Hecht, H.: 'The behavioral validity of dual-task driving performance in fixed and moving base driving simulators', *Transportation research part F: traffic psychology and behaviour*, 2016, **37**, pp. 78–96

11 NHTSA (2012), âĂŸVisual-manual nhtsa driver distraction guidelines for in-vehicle electronic devicesâĂŹ, Washington, DC: National Highway Tra c Safety Admin- istration (NHTSA), Department of Transportation (DOT) .

12 Blana, E.: 'Driving simulator validation studies: A literature review.', , 1996,

13 Classen, S., Bewernitz, M., Shechtman, O.: 'Driving simulator sickness: an evidence-based review of the literature', *American journal of occupational therapy*, 2011, **65**, (2), pp. 179–188

14 Blaauw, G.J.: 'Driving experience and task demands in simulator and instrumented car: a validation study', *Human Factors*, 1982, **24**, (4), pp. 473–486

15 Mullen, N., Charlton, J., Devlin, A., Bedard, M.: 'Simulator validity: Behaviors observed on the simulator and on the road', , 2011,

16 Wang, Y., Mehler, B., Reimer, B., Lammers, V., D'Ambrosio, L.A., Coughlin, J.F.: 'The validity of driving simulation for assessing differences between in-vehicle informational interfaces: A comparison with field testing', *Ergonomics*, 2010, **53**, (3), pp. 404–420

17 Reed, M.P., Green, P.A.: 'Comparison of driving performance on-road and in a low-cost simulator using a concurrent telephone dialling task', *Ergonomics*, 1999, **42**, (8), pp. 1015–1037

18 Gemou, M.: 'Transferability of driver speed and lateral deviation measurable performance from semi-dynamic driving simulator to real traffic conditions', *European Transport Research Review*, 2013, **5**, (4), pp. 217–233

19 Godley, S.T., Triggs, T.J., Fildes, B.N.: 'Driving simulator validation for speed research', *Accident analysis & prevention*, 2002, **34**, (5), pp. 589–600

20 Yan, X., Abdel.Aty, M., Radwan, E., Wang, X., Chilakapati, P.: 'Validating a driving simulator using surrogate safety measures', *Accident Analysis & Prevention*, 2008, **40**, (1), pp. 274–288

21 Allen, R.W., Klein, R.H., Ziedman, K.: 'Automobile research simulators: a review and new approaches', *Transportation research record*, 1979, **706**, pp. 9–15

22 Törnros, J.: 'Driving behaviour in a real and a simulated road tunnelâĂŤa validation study', *Accident Analysis & Prevention*, 1998, **30**, (4), pp. 497–503

23 Santos, J., Merat, N., Mouta, S., Brookhuis, K., De.Waard, D.: 'The interaction between driving and in-vehicle information systems: Comparison of results from laboratory, simulator and real-world studies', *Transportation Research Part F: Traffic Psychology and Behaviour*, 2005, **8**, (2), pp. 135–146

24 Victor, T.W., Harbluk, J.L., Engström, J.A.: 'Sensitivity of eye-movement measures to in-vehicle task difficulty', *Transportation Research Part F: Traffic Psychology and Behaviour*, 2005, **8**, (2), pp. 167–190

25 Baumann, M., Keinath, A., Krems, J.F., Bengler, K.: 'Evaluation of in-vehicle hmi using occlusion techniques: experimental results and practical implications', *Applied ergonomics*, 2004, **35**, (3), pp. 197–205

26 Pettitt, M.A., Burnett, G.E., Bayer, S., Stevens, A. 'Assessment of the occlusion technique as a means for evaluating the distraction potential of driver support systems'. In: IEE proceedings-Intelligent transport systems. vol. 153. (IET, 2006. pp. 259–266

27 Bach, K.M., Jæger, M.G., Skov, M.B., Thomassen, N.G. 'Evaluating driver attention and driving behaviour: comparing controlled driving and simulated driving'. In: Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction-Volume 1. (British Computer Society, 2008. pp. 193–201

28 Knapper, A., Christoph, M., Hagenzieker, M., Brookhuis, K.: 'Comparing a driving simulator to the real road regarding distracted driving speed.', *European Journal of Transport & Infrastructure Research*, 2015, **15**, (2)

29 Fisher, R.A.: 'Xv.âĂŤthe correlation between relatives on the supposition of mendelian inheritance.', *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 1919, **52**, (2), pp. 399–433

30 Barr, D.J., Levy, R., Scheepers, C., Tily, H.J.: 'Random effects structure for confirmatory hypothesis testing: Keep it maximal', *Journal of memory and language*, 2013, **68**, (3), pp. 255–278

31 Markkula, G., Engström, J. 'A steering wheel reversal rate metric for assessing effects of visual and cognitive secondary task load'. In: PROCEEDINGS OF THE 13th ITS WORLD CONGRESS, LONDON, 8-12 OCTOBER 2006. (, 2006.

32 Burnett, G., Neila, N., Crundall, E., Large, D., Lawson, G., Skrypchuk, L., et al.: 'How do you assess the distraction of in-vehicle information systems? a comparison of occlusion, lane change task and medium-fidelity driving simulator methods', *Proceedings of DDI2013*, 2013,

33 Large, D., Burnett, G. 'An overview of occlusion versus driving simulation for assessing the visual demands of in-vehicle user-interfaces'. In: International Conference on Driver Distraction and Inattention, 4th, 2015, Sydney, New South Wales, Australia. 15309. (, 2015.

34 Hart, S.G., Staveland, L.E. 'Development of nasa-tlx (task load index): Results of empirical and theoretical research'. Advances in psychology, 52:139âĂŞ183, 1988.

35 Salvucci, D. D. and Goldberg, J. H. (2000), 'Identifying fixations and saccades in eye-tracking protocols', in Proceedings of the 2000 symposium on Eye tracking research & applications, ACM, 71âĂŞ78.

36 Louw, T., Madigan, R., Carsten, O. and Merat, N. (2017), âĂŸWere they in the loop during automated driving? links between visual attention and crash potentialâĂŹ, Injury Prevention 23(4), 281âĂŞ286.

37 Broström, R., Ljung Aust, M., Wahlberg, L. and Ka ÌĹllgren, L. (2013), âĂŸWhat drives oâĘt-road glance durations during multitasking: capacity, practice or strategy?âĂŹ, in 3rd International conference on driver distraction and inattention.

38 Land, M. F. âĂIJEye movements and the control of actions in everyday life. Progress in Retinal and Eye Research, 25(3), 296-324, 2006.