Virtual Techniques for Prototype HMI Evaluation

by

Panagiotis Dimitrios Spyridakos

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

Institute for Transport Studies

December 2018

Declaration

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Abstract

The aim of this project was to investigate the behavioural validity of virtual methods, namely driving simulators and computational models, as prototype HMI evaluation tools. A driving study was designed where participants had to perform secondary tasks while driving in a real world and a driving simulator setting. Statistical analysis of the data, along with an in-depth review of related findings was used to identify the levels of behavioural validity that could be achieved by different simulator settings across different metrics. A further analysis was performed to identify behavioural strategies that drivers employ regarding their visual attention sharing while executing HMI tasks concurrently to driving. Finally, two existing computational models were validated and a novel model was proposed that can account for drivers' behavioural phenomena, not previously accounted for.

Acknowledgements

I owe gratitude to all my supervisors: Dr. Hamish Jamson and Dr. Erwin Boer for giving me the opportunity to work on this project in the first place, and Dr. Gustav Marrkula and Prof. Natasha Merat for helping me finish it. A few other people have contributed to the composition and completion of this thesis, directly and indirectly; I am deeply indebted to each and every one of them. My thanks goes to Jean Doerpinghaus for his immense help with the JLR experiments and to Dr. Oscar Giles for demystifying linear mixed effects models for me. A big thank you also to Dr. George Kountouriotis for our lovely collaboration and to the UoLDS team and HFS group. I would also like to thank Dr. Charles Fox for being a constant source of inspiration.

I could not leave out either those who sprinkled enough stardust from the beyond, or my family and friends for always lending an ear.

Leaving best for last, I could not possibly not thank my Annie. Nor could I summon words to describe how much I owe to her for the continuous support, without which I would have never needed to write an Acknowledgements section for a completed thesis.

Contents

Declaration iii

Abstract v

Acknowledgements vii

1 Introduction 1

1.1 Background and Motivation . . . 2
1.2 Project Aims and Research Questions . . . 6
1.3 Thesis Structure . . . 8

2 Literature Review 11

2.1 The Multi-layered Activity of Driving . . . 11
2.2 Performing HMI Tasks while Driving . . . 13
2.2.1 Attentional Resources and Driver Distraction . . . 14
2.2.2 Driver Distraction and Workload . . . 16
2.2.3 Effects of HMI Engagement on Driving Performance . . . 18
2.2.4 Driver Compensatory Behaviours . . . 20
2.2.5 Driver Strategies for Engaging in HMI Tasks . . . 22
2.3 HMI Evaluation Methods . . . 22
2.3.1 Established Guidelines for Procedures and Metrics . . . 23
2.3.2 Computational Models and Modelling Frameworks . . . 24
2.4 Driving Simulator Validity . . . 27
2.4.1 Behavioural Validity . . . 28
2.4.2 Factors Affecting Behavioural Validity . . . 29

2.4.3 Desired Level of Behavioural Validity for HMI Evaluation . 31
2.5 Key Research Gaps 32

3 Empirical Studies and Methods 33
3.1 Participants 33
3.2 Design and Procedure 35
3.2.1 Materials 35
3.2.2 Driving Environment 37
3.2.3 Driving Scenarios 38
3.2.4 HMI Tasks 39
3.2.5 Experimental Design 45
3.2.6 Procedure 45
3.3 Initial Data Reduction 47
3.3.1 Identifying and Extracting Individual HMI Task Executions 47
3.3.2 Defining Where the Drivers Look 48
3.3.3 Data Cleaning 49
3.3.4 Data Loss 51

4 Behavioural Validity of Driving Simulators for Prototype HMI Evaluation 53
4.1 Methods 54
4.1.1 Review of Related Literature 54
4.1.2 Data Analysis 54
4.1.3 Establishing Level of Behavioural Validity 55
4.2 Literature Review Results 55
4.3 Data Analysis Results 60
4.3.1 HMI Task Performance 61
4.3.2 Gaze Behaviour 63
4.3.3 Lateral Control 66
4.3.4 Longitudinal Control 67
4.3.5 Steering Control 69
4.4 The Behavioural Validity Matrix 71

5 Visual Attention Sharing Patterns During HMI Task Execution 73

5.1 Exploratory Analysis . . . 74

5.2 Visual Attention Sharing Structure . . . 78

5.3 Effects of Primary Driving Task Demand on Visual Attention Sharing 82

5.3.1 Glance Durations and Time-to-Line Crossing . . . 83

5.3.2 Glance Durations and Headway Distance Adjustment . . . 90

6 Modelling Drivers' Visual Attention Allocation for Prototype HMI Evaluation 95

6.1 Available Computational Models of Drivers' Visual Behaviour . . 97

6.1.1 IVIS DEMAnD Model . . . 98

6.1.2 The Queuing Network Model Human Processor (QN-MHP) 99

6.1.3 The Salience Effort Expectancy and Value (SEEV) Model 99

6.1.4 Adaptive Control of Thought - Rational (ACT-R) . . . 101

6.1.5 Fitts' and Hick-Hymann Law . . . 101

6.1.6 Distract-R . . . 103

6.1.7 Large et. al Predictive Equations . . . 104

6.1.8 Other models . . . 105

6.2 Predicting Visual Attention Sharing During HMI Engagement . . 106

6.2.1 Distract-R . . . 108

6.2.2 Large et al. Predictive Equations . . . 114

6.2.3 Proposed model . . . 116

6.3 Discussion . . . 123

7 Conclusion 127

7.1 Key Findings and Contribution to Knowledge . . . 128

7.1.1 What type of driving simulator should be used in prototype HMI evaluation related user trials? . . . 128

7.1.2 How do drivers engage with HMI tasks while driving? . . . 129

7.1.3 What types of computational models could predict the observed behaviour? . . . 130

7.2 Methodology and Design Issues and Limitations . . . 131

7.3 Future Work Suggestions 135

Bibliography **139**

A - Participant Briefing Sheets **159**

B - Participant Consent Forms **163**

C - Subjective Questionnaires **165**

D - Counterbalancing **171**

List of Figures

1.1 Thesis Structure. 10

2.1 Michon's conceptual model of the driving task. Figure recreated with permission from Michon (1985) 12

2.2 Multiple resource theory model, reproduced with permission from Wickens (2002). 15

3.1 The University of Leeds Driving Simulator (UoLDS). The pictures are generic and were not taken during the experiment. 35

3.2 Subject car used in the Gaydon experiment. The pictures are generic and were not taken during the experiment. 37

3.3 Proving Ground facilities layout in Gaydon. 38

3.4 Two instances of the Emissions Circuit in Gaydon (left) with their corresponding instances from the simulated version of the test track in University of Leeds Driving Simulator (UoLDS). 39

3.5 In-vehicle HMI setup. 41

3.6 Easy HMI task. 42

3.7 Medium HMI task. 43

3.8 Hard secondary task. 44

3.9 AOIs used for glance coding. 50

4.1 HMI task completion times. 62

4.2 Total off-road glance duration. 64

4.3 Glance frequency. 65

4.4 Mean off-road glance durations. 66

4.5 Standard deviation of lateral position. 67
4.6 Speed variability. 68
4.7 Average speed. 68
4.8 Effects and interactions of SWRRs for gap size of 1^o. 69
4.9 Effects and interactions of SWRRs for gap size of 5^o. 70
4.10 Effects and interactions of SWRRs for gap size of 10^o. 70

5.1 Glance duration distributions during HMI task execution for real world and simulated driving. 75
5.2 Distribution of number of glances per task execution in real world and simulated driving. 79
5.4 A hypothesis of how TLC data would compare against off-road glance durations. 84
5.5 TLC_{off} against subsequent off-road glance duration, for straight and curved road. 85
5.6 TLC_{off} under different driving conditions for straight and curved road. 89
5.7 Visual angle rate of change of lead vehicle at the moment of looking away. 91

6.1 Easy HMI task and corresponding Distract-R representation. . . . 110
6.2 Medium HMI task and its corresponding Distract-R representation. 111
6.3 Hard HMI task and its corresponding Distract-R representation. . 112
6.4 Distract-R model performance. 114
6.5 Large et al. (2018) predictive equations performance. 117
6.6 Proposed model performance. 121
6.7 Best model variants performance. 125

List of Tables

3.1 Participant Demographics 34
3.2 Participant Driving Experience - Annual Mileage 35
3.3 Participant Driving Experience - Years Driving 35

4.1 Results from related literature on driving simulator behavioural validity. 59
4.2 The Behavioural Validity Matrix, based on analysis of the obtained data and existing results from literature. 72

5.1 Visual attention sharing structure during HMI task execution details. 82
5.2 Kolmogorov Smirnov test results for TLC CDFs. 90

6.1 $RMSE$ values for the Distract-R model and variants. 115
6.2 $RMSE$ values for the Large et. al predictive equations. 116
6.3 $RMSE$ values for the proposed model variants. 122
6.4 $RMSE$ values for the proposed model variants, evaluated in fixed base, constant speed, curved road simulator data. 124
6.5 $RMSE$ values for the proposed model variants, evaluated against constant speed, straight road, real word data. 124

7.1 Counterbalancing in the UoLDS experiment. 171
7.2 Counterbalancing 172

Abbreviations

JLR	Jaguar Land Rover
LVc	Lead Vehicle constant speed condition
LVv	Lead Vehicle varying speed condition
Sim. Fix.	Simulator Fixed Base
SWRR	Steering Wheel Reversal Rate
UoLDS	University of Leeds Driving Simulator

Chapter 1

Introduction

In-vehicle interfaces have been gaining popularity since they were first introduced in the beginning of the 21^{st} century and are nowadays present in the majority of production vehicles (Harvey, 2011). They are meant to provide the driver with various types of support or, quite often, entertainment while driving. Consequently, 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 (Meixner et al., 2017).

Such interfaces can be classified into three categories, as proposed by Galer (1995):

1. Interfaces that are fundamentally related to the primary driving task, such as Advanced Driving Assistance Systems (ADAS) (e.g. Automated Driving, Adaptive Cruise Control, Lane Keeping Assistance, etc).

2. Interfaces that provide relevant, but not functionally necessary, information and services to the primary driving task, such as satellite navigation systems or general route-related information.

3. Interfaces that are not directly related to the primary driving task and provide information or entertainment to the driver, such as radio or email accessibility.

The work presented in this thesis revolves around the third type of interfaces, as defined in the categorisation above. Such interfaces will henceforth be referred to by the general term "*Human Machine Interface* (HMI)" and the tasks associated with them as "*Human Machine Interface tasks* (HMI tasks)".

It has been well established that performing HMI tasks while driving can negatively affect driving performance and increase risk (e.g. Lee et al., 2008; Fitch et al., 2013; Kountouriotis and Merat, 2016). Fitch et al. (2013), for instance, found that performing cell phone related visual-manual tasks while driving increases the probability of a safety-critical event almost threefold. Dingus et al. (2016) showed that activities requiring the driver's eyes to be away from the forward roadway, such as interacting with a cell phone or interacting with touch screen menus, all increased crash risk by up to 12 times. Regardless, drivers still choose to perform such tasks, more than half of the time. Namely, in the same study, Fitch et al. (2013) also showed that drivers were "just driving", without engaging with any secondary task, for only 46% of their time.

Although the tasks that could divert the driver's attention from the road could stem from sources either inside the vehicle or outside of it, it has been previously argued that drivers are more susceptible to the former (Lam, 2002). Given that a big portion of those incidents could be potentially caused by in-vehicle HMIs when the tasks associated with them are highly demanding, there is a need to ensure that such tasks would have the minimum impact on driver's safety and performance. In order to achieve that, thorough and reliable evaluation methods of new HMI designs should be in place throughout all the production stages.

This chapter will move on to elaborate on the motivation behind the research conducted during this project, as well as its main aims and areas of focus.

1.1 Background and Motivation

According to the Department for Transport, the number of recorded crashes that were caused by in-vehicle induced distraction in Great Britain through 2015 and 2016 amounted to 4% of the total number of crashes each year and increased in number from one year to the next (Department for Transport, 2015). Although this might already look alarming, there are bodies, like the World Health

Organisation, that suggest this could be only an underestimate, given the difficulties entailed in identifying driver distraction as the main (or sole) cause of a crash (Organization et al., 2011). Indeed, data from naturalistic studies indicates that these numbers are actually substantially higher. For example, according to Klauer et al. (2006), 78 percent of all crash and 65 percent of all near-crash events observed involved some form of driver inattention and distraction, with the majority relating to secondary HMI tasks.

In order to address the matter of safety when it comes to designing and producing new HMIs, a variety of evaluation methods have been proposed over the years to quantitatively assess the distraction potential of a new design, for example: naturalistic studies (e.g. Klauer et al., 2006), Distract-R rapid prototyping (Salvucci et al., 2005*a*), the occlusion method (Senders et al., 1966) and the NHTSA evaluation guidelines (NHTSA, 2012), to name a few. Evaluation in the real world (conducting naturalistic studies and real world testing) can be very costly and time consuming. Using virtual methods, on the other hand, (driving simulator studies, lab based experiments and computational model simulations) can speed up the evaluation process and make it more efficient. From the available virtual methods for HMI evaluation, the work conducted here explores the use of driving simulators and computational models. Moving forward, for the scope of this thesis, the term virtual methods will refer to these two alone.

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. Engström et al., 2005; Klüver et al., 2016; Kountouriotis and Merat, 2016). Computational models have also been explored in different contexts of driving, from simple vehicle control to concurrent secondary task execution, and have been shown to provide a representative account of human behaviour in such settings (e.g. Salvucci, 2001; Salvucci et al., 2005*a*; Liu et al., 2006; Markkula, 2014; Large et al., 2018).

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. Blana, 1996; Classen et al., 2011; De Winter et al., 2012). Given that there is no need for human participants or expensive equipment, it is only intuitive to infer that the same benefits would apply to computational models, too. Delving into more detail as to which reasons could encourage the use of virtual methods in prototype HMI evaluation, the following could be argued:

- *Repeatability*: Both in a driving simulator and in a model simulation, trials can be repeated numerous times under identical conditions. For example, unlike experiments in a test-track or a highway or urban road, traffic conditions can be simulated in detail, avoiding unexpected events that could interfere with the results. Repeatability in a driving simulator environment, though, comes with potential issues as learning effects can become evident when a participant is repeatedly exposed to the same conditions. This problem, however, is non-existent with the use of computational models.

- *Cost effectiveness*: Virtual methods significantly reduce the cost of prototype testing in the long term, due to their repeatability and the lack of need for physical prototype production. That being said, such methods might incur initial higher cost for research and development, especially when it comes to driving simulators. Nevertheless, such methods can still be more cost-effective in the long term.

- *Execution speed*: Driving simulator studies may be conducted faster than real world trials due to the ease of resetting the experimental conditions and the redundancy of vehicle related checks that are necessary in a real world setting (tire pressure, engine temperature, etc.). Moreover, the purpose-developed scenarios in simulator studies, allow for the execution of a large number of dedicated maneuveres per time unit (De Winter et al., 2012). Similarly, model simulations can run a lot faster than real time, thus giving the same number of trials in a fraction of the time. Salvucci et al. (2005*a*) presented such an example through the Distract-R modelling framework, which could simulate 10 minutes of real-world driving in a mere 3 seconds. Although none similar have been reported recently, it is safe to assume that in the age of cloud and distributed computing such figures can have only improved.

- *Increased safety*: Driving simulators provide a safe environment where dangerous events such as crashes and near-misses can be studied with no potential danger to the participant.

- *Scalability*: With computational modelling in particular, there is also no restriction (other than those imposed by hardware systems) as to how many simulations can be run. Hence, an adequate number of repetitions can be obtained, unrestricted by issues such as participant recruiting etc.

It is evident that virtual evaluation methods are overall very efficient, time-, cost- and implementation-wise, features that could greatly benefit both the automotive industry and the end-user drivers. However, virtual methods are not without their drawbacks, with the most significant one being their limited validity, i.e. the degree to which the results obtained from studies employing such methods match the results obtained under real-world conditions. Results obtained using such methods should match real-world driving and, ideally, have minimal deviation from it. In regards to driving simulators, a common issue that can hinder the validity of the results is simulator sickness, where participants can experience symptoms of nausea or disorientation and need to drop out of the study. Considering computational models, on the other hand, since the behavioural and cognitive mechanisms involved in driving and general task performance are not completely understood, they can only approximate human behaviour based on assumptions about what drives that behaviour. Consequently, time and research are still needed to further our knowledge on human behaviour and improve the validity of virtual HMI evaluation methods.

Going back to how these methods could be incorporated in the evaluation cycle, one would first need to define that cycle. One of the views of the development cycle in the automotive HMI context is provided by Pettitt and Burnett (2010) and consists of four discrete stages:

- Stage 1: The early stages of the HMI design. This refers to the time during which the HMI design is being conceptualised and its core components are being decided. At that stage, designers should adhere to established design guidelines and regulations, in order to ensure that their design follows the minimum requirements for a prototype interface design.

- Stage 2: The first (low-fidelity) prototype of the HMI. This is when the designers have a more detailed specification of the interface (e.g. dimensions, positioning within the vehicle, interaction modalities, etc.).
- Stage 3: The revised (medium-fidelity) prototype of the HMI. This would involve creating a first "draft" physical prototype of the interface; this prototype would facilitate the same functions (i.e. menus and tasks) as the intended device but might not be otherwise similar to the intended final product (e.g. appearance, casing or mounting).
- Stage 4: The final (high-fidelity) prototype of the HMI. Here, researchers would have a final (or close to final) prototype of the interface fitted in a simulator cabin or a real vehicle console, at the same position as the intended final product.

Increasing the validity of virtual methods would ensure that they could be employed across all four stages of the evaluation process and help detect design errors further back in the earlier stages, before moving on to costly prototype production and extensive testing.

1.2 Project Aims and Research Questions

The work presented in this thesis aims at providing new insights into the behavioural validity of virtual prototype HMI evaluation methods. Behavioural validity relates to the degree to which the driving behaviour observed through the use of virtual evaluation methods resembles the driving behaviour that would be observed by drivers in the corresponding real world conditions (Blaauw, 1982). As mentioned earlier, the virtual evaluation methods this investigation focuses on are driving simulators and computational models of the HMI interaction.

Moreover, focus is placed on the behavioural strategies that are (consciously or unconsciously) employed by drivers when engaging in visual-manual HMI tasks. It is crucial to properly capture any emerging patterns and understand what drives them, in order to work towards improving the validity of virtual evaluation methods. Having that knowledge can lead to the conceptualisation and implementation of more accurate models, as well as contribute to the enhancement of driving simulator realism.

The work presented in the rest of this thesis addresses the following research questions:

1. **What type of driving simulator should be used in prototype HMI evaluation related user trials?**
   - What levels of realism can different simulator settings achieve?
   - Is this consistent? If not, how does it vary across different settings and evaluated metrics?
   - Is a single simulator setting "good enough" for evaluating different behavioural metrics?
2. **How do drivers engage with HMI tasks while driving?**
   - Are there any distinct behavioural strategies/patterns that drivers employ regarding their visual attention sharing during HMI task execution?
   - Are there any safety or other conceptual thresholds that dictate engagement and disengagement points?
   - Which elements of the overall situation (i.e. driving environment, primary driving task, secondary HMI task and individual driver characteristics) affect the driver's HMI engagement patterns (e.g. onset and duration of engagement) and in which way?
3. **What types of computational models could predict the above observed behaviour?**

- Can any of the existing models capture the behaviour observed when drivers engage to HMI tasks, accurately?
- If not what might be the root cause of that performance?
- Can models be reliably used as a virtual participant for future HMI designs?

1.3 Thesis Structure

The present thesis has been structured in seven chapters, as outlined in Figure 1.1. A brief overview of each one of the following chapters can be found below:

- **Chapter 2** provides a review of the existing literature regarding driving and concurrent HMI task interaction, as well as prototype HMI evaluation methods. In particular, it discusses the nature and demands of both the primary driving task and secondary HMI tasks, their interaction and any effects said interaction has been shown to have on driving performance. It then moves on to discuss the various methods that can be used for evaluation of HMI prototypes as well as for gaining insights into how drivers interact with HMIs in the car, ranging from academic research outcomes to established guidelines and review processes.
- **Chapter 3** describes the two driving studies that were designed and carried out in the scope of this project. The studies involved a combined total of 23 participants, driving in three different environments and two different scenarios, where participants had to interact with a prototype HMI while driving, performing three visual-manual tasks of varying difficulty levels.
- **Chapter 4** presents an investigation on driving simulator behavioural validity. Using data from the aforementioned driving study, as well as results from existing literature, a comparison between driving behaviour in different types of driving simulators and in reality was carried out. The driving behaviour compared here was the one observed during concurrent driving and HMI task execution. The results are presented in the form of a "validity

matrix" that aggregates the level of behavioural fidelity different simulator settings have been found to achieve across different behavioural metrics.

- **Chapter 5** presents the results obtained from a more in-depth analysis of the drivers' visual attention sharing behaviour. Namely, this analysis focuses on the durations, as well as the timing of initiation of glances towards the interface during HMI task engagement. Moreover, an investigation on which factors affect glance onset and duration is presented with a focus on identifying whether such factors could be used to predict the observed behaviour. Insights from these analyses could potentially be used to improve existing or formulate new computational models that can better account for driver behaviour under dual-tasking conditions.

- **Chapter 6** reviews some of the available computational models that can be used as prototype HMI evaluation tools by predicting driver visual attention allocation during HMI task engagement. The two most promising of those models, along with some variations of them (exploring parameters that would be meaningful to be included) were tested and compared against the observed behavioural data. Finally, a novel model is proposed, based on existing HMI task modelling approaches but incorporating new behavioural aspects that have not been directly modelled thus far.

- Finally, **Chapter 7** summarises and discusses all the outcomes resulting from the aforementioned work, proposing improvements and future work.

Although an overall review of the existing relevant research is provided in Chapter 2, more detailed reviews on specific subjects are also presented in individual Chapters. Due to the nature of the work conducted for this project, a detailed review of relevant literature was part of the methodology used in some of the data analyses, hence such reviews are presented as part of the corresponding Chapter, while a higher level review is provided in the following Chapter.

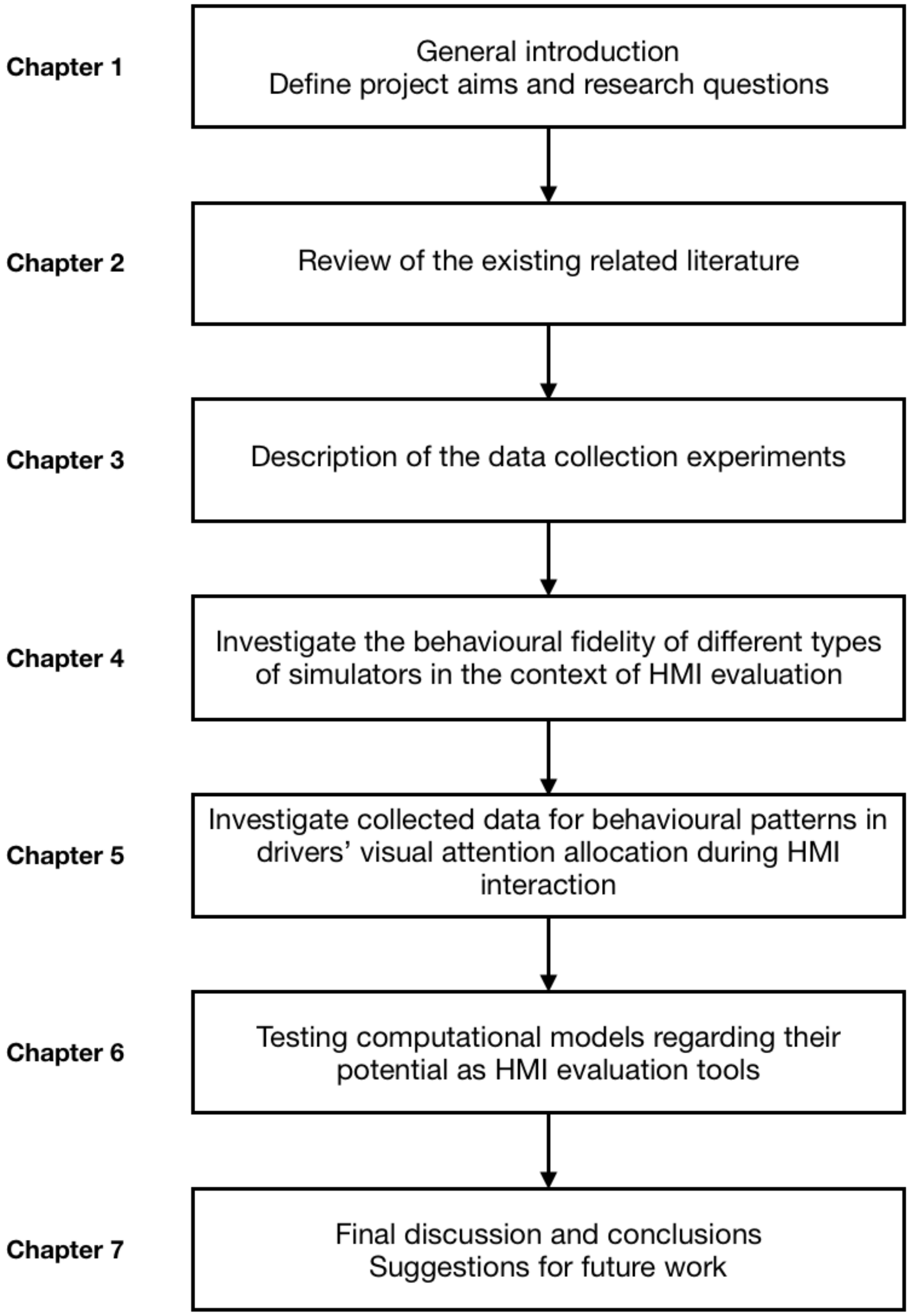


Figure 1.1: Thesis Structure.

Chapter 2

Literature Review

This chapter provides an overview of existing research related to prototype HMI evaluation and driver behaviour during HMI task performance. Initially, driving is deconstructed as an activity formed by different sub-tasks, discussing their demands and characteristics, before moving on to the effects of concurrent execution of HMI tasks on driving performance. Then, the issue of driving simulator behavioural validity is discussed, followed by modelling approaches used in the context of HMI evaluation.

2.1 The Multi-layered Activity of Driving

Driving is a complex, multi-tasking activity which requires successful acquisition and coordination of various physical, cognitive, sensory and psychomotor skills (Hedlund et al., 2006; Young et al., 2007; Regan et al., 2008; Groeger, 2013). The driver receives visual input of the driving scene, while maintaining the control of the vehicle within a safe margin (Wierwille, 1993; Lansdown, 2000).

One of the most popular descriptive models that conceptualises driving, is the one proposed by Michon (1985) (see Figure 2.1). In this model, three hierarchical levels are identified, each one of which is relevant to different aspects of the driving activity:

1. The *strategic level*, involving any high-level decision regarding the driving journey itself (e.g. determining the route of the trip and the overall goal).

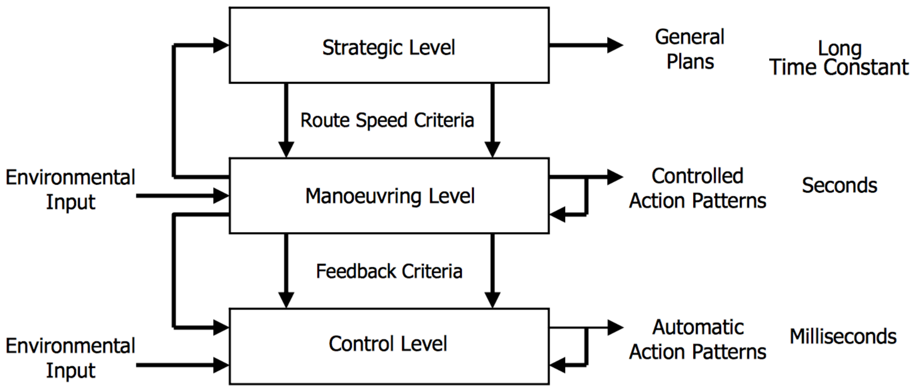


Figure 2.1: Michon's conceptual model of the driving task. Figure recreated with permission from Michon (1985)

2. The *manoeuvring* or *tactical level*, involving tactical decisions (e.g. where to turn and overtake) which should be in line with the goals from the strategic level.

3. The *control level*, involving any actual lateral and longitudinal control of the vehicle by providing inputs such as steering, braking and accelerating, which should be in line with the goals from the tactical level.

Another descriptive account, treating the driving task from a different perspective, is provided by Hedlund et al. (2006), where driving is described as a set of two types of tasks. Steering, accelerating, braking, speed choice, lane choice, manoeuvring in traffic, navigation to destination, and scanning for hazards are considered the *primary* driving tasks. In other words, as primary driving tasks are classified all the activities that are directly relevant to a non-erroneous control of the vehicle. The list of *secondary* tasks includes all other activities into which drivers engage, that are not directly related to controlling the vehicle. As an example, such activities could be conversing with a passenger, viewing the scenery, smoking, using a cell phone or conversing on the phone, to name a few (Hedlund et al., 2006).

As mentioned earlier, in Chapter 1, the work conducted in this thesis focuses on secondary tasks that are related to in-vehicle HMIs. Moving forward, the terms "HMI task" and "secondary task" are being used interchangeably. Such tasks would include tuning the radio, adjusting the climate controls or entering a destination to the in-car satellite navigation system. Drivers have traditionally performed HMI tasks via tactile interfaces, such as buttons and switches, usually located on the central console of the vehicle. Recently, however, the number and variety of secondary functions available within vehicles have increased significantly, from simple radio and climate controls to navigation, visual media, entertainment, communication and network connectivity (Gu Ji and Jin, 2009; Eren et al., 2015; Meixner et al., 2017). Moreover, nowadays, such functions are often integrated in a single, menu-based system and accessed through a touch-screen interface (Harvey, 2011). Kern and Schmidt (2009) investigated 117 car models from 35 different manufacturers, taken at the international automobile exhibition (IAA 2007) in Frankfurt, and found that almost half of them included a touch-screen based interface.

However, the majority of the research related to HMI task performance, since it was conducted more than a decade ago, focused mostly on mobile phone interaction (see Young et al., 2007, for a relevant review). In the cases where alternative visual-manual HMI tasks were investigated, researchers chose to use surrogate HMI tasks (Victor et al., 2005; Engström et al., 2005; Jamson and Merat, 2005; Kountouriotis and Merat, 2016, e.g.). As a result, very limited research has been conducted so far on HMI tasks and production grade prototypes of touch screen based interfaces (for example Large et al., 2015, 2018). Given the fact that such interfaces are becoming more common in production vehicles, it is essential that processes are in place to appropriately assess their distraction potential and ensure that the tasks associated with them are of acceptable complexity.

2.2 Performing HMI Tasks while Driving

One of the primary points of interest when studying HMI task performance, is the type of the task performed by the driver. In particular, it is important to identify the modality of the task and, consequently, what cognitive resources it demands. This information can then help us better understand how drivers interact with the task, as well as what type and magnitude of an effect this task will have on driving performance.

In the following subsections, more detail is provided as to how different types of tasks affect driving performance and why. Moreover, the drivers' ability to self-regulate and maintain a safe driving behaviour when performing HMI tasks is discussed and, finally, an overview of the factors that dictate drivers' engagement to HMI tasks is presented.

2.2.1 Attentional Resources and Driver Distraction

In order to understand what happens when drivers divide their attention between the primary driving task and an HMI task, it would be useful to refer to theories of attention and multitasking. Attention is a broad term that has been used to describe how selection mechanisms operate in the brain and how humans perform complex tasks. Various theories and models of attention have been proposed that aim to explain how information is processed and prioritised.

The Filter Theory, for example, assumes a "bottleneck" stage in between the stages of perceiving an input and analysing it (Broadbent, 2013). In particular, according to this theory, if more than one stimuli are provided as input simultaneously, only one of them is perceptually analysed at that time. The rest remain stored temporarily (similarly to the cache memory of computers) until they can be analysed.

A different descriptive model for divided attention, which does not treat information processing as linear and addresses the execution of multiple concurrent tasks was developed by Kahneman (1973). In that model, Kahneman argues that humans have a limited amount of attentional resources, that can be allocated to any task one might be undertaking. Thus, one is able to perform multiple tasks at once, as long as there are enough attentional resources to be allocated to each one.

One of the most popular attentional resource models, which was partially inspired by Kahnmean's model, is the multiple resource theory introduced by Wickens (2002) (see Figure 2.2). That theory is structured in four distinct di-

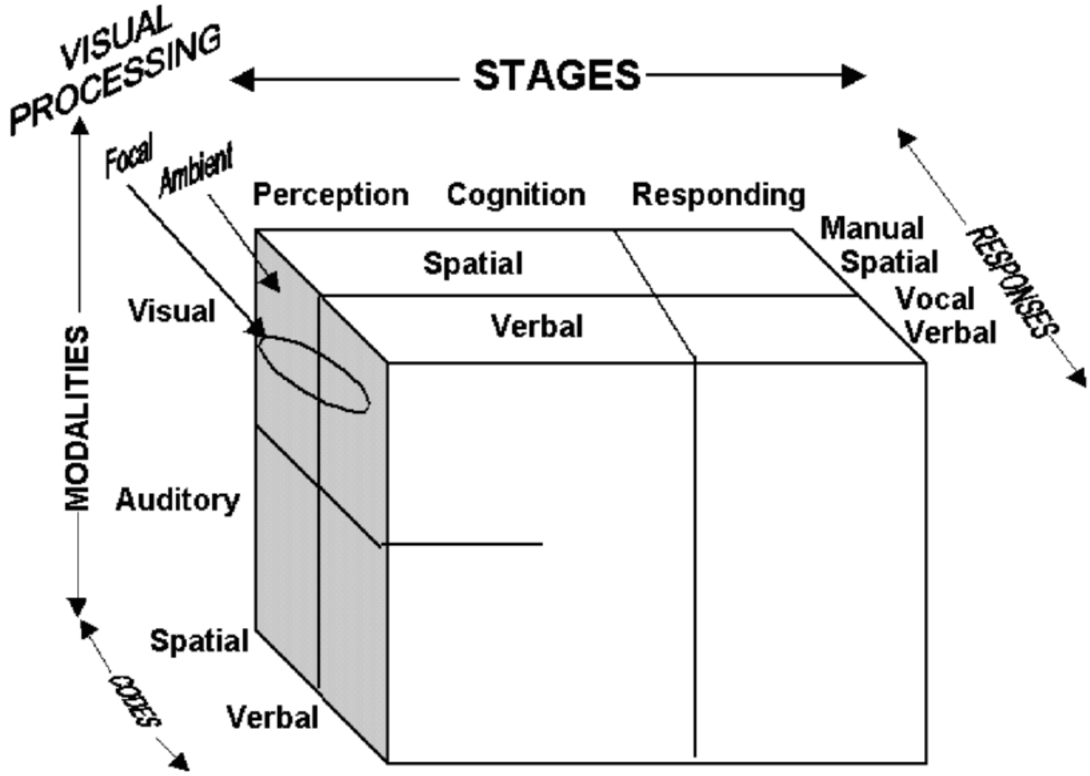


Figure 2.2: Multiple resource theory model, reproduced with permission from Wickens (2002).

mensions, each one of which is related to varied performance in concurrent tasks and each one of which is further divided into two discrete lower "levels". The main notion/rule behind the functionality of this model is the following; "All other things being equal (i.e. equal resource demand or single task difficulty), two tasks that both demand one level of a given dimension (e.g. two tasks demanding visual perception) will interfere with each other more than two tasks that demand separate levels of the dimension (e.g. one visual, one auditory task)" (Wickens, 2002).

However, such models, although explaining secondary task interference, do not account for key phenomena in real-world driving, such as the self-regulation of attention (i.e. how drivers choose to share their attention between the primary driving task and an HMI task) (Engström et al., 2013). Moreover, there is a need for a better way to facilitate how resources are shared between concurrent tasks, regardless of what their conflict might be (Wickens, 2002).

A more recent attentional theory, proposed by Engström et al. (2013), treats attention selection as a form of adaptive behaviour. The driver's adaptive behaviour reflects actions taken to achieve task related goals while remaining inside a subjectively defined comfort zone (Engström et al., 2013). When drivers experience discomfort due to an actual or potential violation of their subjective safety margins, they take actions to return to their comfort zone. In other words drivers are able to dynamically share their attention between the primary driving task and a secondary task, based on the demand of the the former and the effects of the latter. Thus, a key idea behind this model is that attention selection in everyday driving functions towards enabling an appropriate balance between goal achievement and acceptable safety margins maintenance (Engström et al., 2013).

2.2.2 Driver Distraction and Workload

There is a long standing lack of consensus among the academic community on a common definition of distraction, despite the numerous attempts on it (for reviews see Regan et al., 2011; Kircher and Ahlstrom, 2017). Kircher and Ahlstrom (2017) identified the following two statement as being the core of almost all driver distraction definitions:

1. Distraction is assumed to be present when attention is shifted away from targets relevant to driving, and the resulting driver behaviour is, or risks to be, detrimental to safe driving.

2. A shift of attention to anything not relevant to driving results in distraction, regardless of the outcome of the situation.

Although driver distraction is not directly studied here, the second statement will be adopted as a suitable definition in the scope of this thesis. Consequently, it is considered that every time the driver performs an HMI task while driving, they get distracted.

Another common term that is often referenced in driver distraction related research is *workload.* Workload, in the context of driver distraction, has been defined as the competition in driver's perceptual, cognitive, and physical resources between the primary driving task and a concurrent secondary task (Angell et al., 2006). An increased workload can be manifested through poor driving performance (e.g. lane-keeping, longitudinal control, object-and-event detection, or eye- glance behaviour) and an increase in the induced distraction, but can also be associated to an improvement in driving performance (e.g. Cooper et al., 2013).

At this point, it would be interesting to discuss the nature of driver distraction and how it relates to the HMI tasks that lead to it. Most commonly, driver distraction is considered interchangeable with the modality of the associated task and can, hence, be classified as one of the following (Ranney et al., 2001; Young et al., 2007; NHTSA, 2012):

- Visual distraction, which occurs when the driver shift their gaze on a different area than the road ahead (e.g. an HMI).
- Auditory distraction, which occurs when when responding to auditory cues, such as responding to a ringing cell phone.
- Biomechanical, physical or manual distraction, which occurs when the driver removes their hand(s) from the steering wheel to manipulate an object.
- Cognitive distraction, which occurs when the driver is "thinking" away from the primary driving task (e.g. daydreaming or conversing to a passenger without looking away from the road ahead).

Although discretised by definition, there is usually an overlap in the types of distraction a driver is exposed to and most naturalistic tasks performed while driving involve all of the above components (Mehler et al., 2012). Kountouriotis and Merat (2016) have argued that the term "cognitive" is misused as it implies that visual tasks are void of a cognitive component and, instead, proposed the term "non-visual" (for secondary tasks or distraction) to better convey the absence of a visual component and to make the distinction more clear. Given the nature of modern HMIs, as discussed in the opening section of this chapter, it could also be argued that the occurrence of purely visual tasks in the driving context is rather scarce and that usually such tasks also involve a manual element, requiring the driver to use hand gestures to interact with the interface.

2.2.3 Effects of HMI Engagement on Driving Performance

Different types of secondary tasks, based on modality and difficulty, can have different effects on driving performance, as explained earlier. Since the focus of this work is on visual-manual tasks, emphasis is given on their effect on driving performance. However, both for the sake of completeness, as well as due to the overlapping relationship of cognitive and visual distraction, a brief overview of the effect of cognitive tasks on driving performance is also provided.

One very commonly reported effect of visually demanding HMI tasks on driving performance, is poor lane-keeping performance (e.g. Engström et al., 2005; Santos et al., 2005; Liang and Lee, 2010; Kountouriotis and Merat, 2016). This can be manifested through an increase in the vehicle lateral position deviation from the centre of the lane or, in extreme cases, through lane exits. In such cases, lane-keeping is affected by increased control input, which is indicative of an increase in workload (Kountouriotis and Merat, 2016). This degradation is a direct implication of the fact that during periods where additional tasks need to be performed, the driver is unable to respond to errors in lateral control, resulting in periods of fixed steering angle (Wickens and Gopher, 1977; Macdonald and Hoffmann, 1980; Godthelp et al., 1984). Consequently, heading errors build up, resulting in the observed lane weaving (Engström et al., 2005).

Such errors need to be corrected by the driver in order to maintain a safe trajectory for their vehicle. Such corrections are achieved through steering manoeuvres which generally are larger and more disruptive than the ones observed during normal driving (Engström et al., 2005). Indeed, engaging in visually demanding HMI tasks has been found to increase steering activity both when evaluating steering wheel reversal rates (SWRRs - Macdonald and Hoffmann, 1980) and Steering Entropy (SE - Boer, 2000).

As one would expect, sharing visual attention between two distinct areas in space would be evident in gaze location measures and performance metrics that relate to one being fully attentive to a specific visual field. Unsurprisingly, visual HMI task execution has been linked to increased deviation of gaze (Victor et al., 2005; Kountouriotis and Merat, 2016). Moreover, the drivers' ability to detect events happening in the road scene can be impaired and their reaction times can increase. Greenberg et al. (2003), for instance, found that visually demanding tasks such as dialling on a phone, resulted in reduced detection of critical traffic events, while Hibberd et al. (2013) found a delay in brake reaction times. Additionally, data from naturalistic driving studies confirm that looking away from the road for prolonged periods of time is a key contributing factor to crashes and near-misses (Dingus et al., 2006; Klauer et al., 2006; Liang et al., 2012; Victor et al., 2015).

Cognitive tasks, on the other hand, do not have so well defined and consistent effects on driving performance. In fact, effects of cognitive load on driving performance are believed to be strongly selective and task-dependent (Engström et al., 2017). Engström et al.'s (2017) Cognitive Control Hypothesis states that cognitive load impairs performance of non-practiced or inherently variable tasks, which rely on cognitive control, while the performance of well-practiced and consistently mapped (automatised) tasks is unaffected and sometimes even improved.

Based on the description of lane keeping as an automatic task by Michon (1985), Medeiros-Ward et al. (2014) argue that it does not necessarily require a focus of attention, and can actually benefit from diverted attention to a secondary task. Indeed, a reduction in the deviation of the vehicle's lateral position has often been observed (e.g. Engström et al., 2005; Jamson and Merat, 2005; Cooper et al., 2013).

In terms of visual behaviour, cognitive tasks have been associated with an increased gaze focus towards the center of the road ahead (Jamson and Merat, 2005; Victor et al., 2005; Kountouriotis and Merat, 2016; Kountouriotis et al., 2016), while there also seems to be agreement that when drivers perform cognitive tasks while driving (a range of tasks from surrogate measures to conversing on a mobile phone) their event detection capabilities deteriorate leading to increased reaction times (e.g. Greenberg et al., 2003; Horrey and Wickens, 2004; Engström et al., 2017). In terms of crash risk, however, naturalistic driving studies have not found supporting evidence of an increase associated with primarily cognitive tasks (e.g. conversing on mobile phone), with some of them finding a significant reduction in crash risk when engaging to such tasks (Klauer et al., 2006; Fitch et al., 2013; Klauer et al., 2014; Victor et al., 2015). Brookhuis et al. (1991) found that using a mobile phone while driving, negatively affected drivers' sampling of the surrounding traffic when driving on a quiet motorway.

2.2.4 Driver Compensatory Behaviours

Apart from the negative effects described above, there have been changes observed in driving performance during dual-tasking that do not have an adverse effect in the ability of the driver to control their vehicle. While engaging in HMI tasks, drivers have been found to self-regulate their actions, through various conscious or unconscious behaviour, thus, compensating for the effects of divided attention on their driving performance and maintaining an adequate level of safe driving (Young et al., 2007; Haigney et al., 2000).

Such compensatory behaviours can occur at the strategic level of the driving task, for example in terms of the driver's choice to engage or not in a secondary HMI task, as drivers have been shown to adapt the amount of attention they allocate to the secondary task based on the demands of the primary driving task (Chiang et al., 2001). For example, Lamble et al. (2002) showed that older drivers chose not to use a mobile phone while they were driving, as their driving performance was more likely to be negatively affected than that of a younger driver. Later, Funkhouser and Sayer (2012) and Tivesten and Dozza (2015) showed that drivers are more likely to initiate a cell-phone conversations or engage in a visual-manual phone tasks when the vehicle is not moving, compared to driving at high speeds.

Compensatory driving behaviours have also been observed at the operational level, i.e. in adapting the level of engagement to the primary driving task in any way. There is a rich body of literature showing that drivers demonstrate a lower mean speed while engaging to a variety of secondary tasks, including the use of hand-held and hands-free mobile phones, the use of in-vehicle entertainment systems and surrogate visual-manual secondary HMI tasks (e.g. Haigney et al., 2000; Chiang et al., 2001; Rakauskas et al., 2004; Engström et al., 2005; Horberry et al., 2006; Kountouriotis and Merat, 2016). Looking at another aspect of longitudinal vehicle control, an increase in headway distance has been associated with engagement to secondary HMI tasks such as hands-free mobile phone use and email processing (e.g. Strayer et al., 2003; Strayer and Drews, 2003; Jamson et al., 2004).

In terms of adapting HMI engagement in complex road traffic scenarios, some recent research suggests that drivers prioritise safety while distracted by a secondary task. For instance, Charlton (2009) showed that drivers are less likely to overtake slower vehicles ahead of them when they are engaged in a cell-phone conversation, while Cooper et al. (2009) showed that, in a similar scenario, drivers were less likely to change lanes while conversing on the phone. Finally, Oviedo-Trespalacios et al. (2017) recently showed that when distracted by a cellphone conversation, drivers selected a lower speed while driving along a curved road or during car-following situations. At the same study, they found that speed adaptation was negligible in complex environments, concluding that the primary driving task was prioritised over the secondary task.

Although the above could lead to an assumption that drivers are in control of their attention allocation, it could also be argued that such behaviours are only artefacts of cognitive saturation in a dual-tasking setting. More specifically, it is possible that concurrent activities cannot be performed at the same level as when performed individually and the compensatory behaviours observed could be a manifestation of this limitation, i.e. an actual degradation in driving performance. Hence, it is important to note that new HMI designs still need to be properly evaluated to ensure they do not require complex interactions that might distract the driver.

Despite the solid understanding of what risk reducing behaviours the drivers engage in, very little research seems to have focused on what drives and dictates those behaviours (i.e. what factors the drivers take into account to decide how and in which way they will engage to a secondary task). There seems to be a general consensus that the primary driving task complexity along with individual characteristics are primarily driving those decisions, without, however, expanding into specific metrics or perceptual cues that the drivers might use to take those decisions.

2.2.5 Driver Strategies for Engaging in HMI Tasks

Despite the effects of various secondary tasks being well documented and understood, very little research has been conducted with regards to the (conscious or unconscious) strategies drivers employ when engaging to HMI tasks. In particular, there is a need to investigate and identify which environmental or task-related factors the drivers consider to decide how they share their attention between the primary driving task and an HMI task (i.e. when they decide to engage with each task and for how long).

Tijerina et al. (2004) investigated this in the context of car-following and found a strong consistency in the strategy with which drivers looked away from the road ahead during car-following. Drivers generally looked away when the rate of change of distance from the leading vehicle was effectively zero. This could be considered as a clear indication that drivers tend to ensure that they are safe before taking their eyes off the road. However, more research is needed to identify how drivers decide to look away from the road especially in relation to the primary task demand, hence providing further insights to potential risks and how willing drivers might be to take them (Tijerina et al., 2004).

2.3 HMI Evaluation Methods

From what has been discussed so far, it is clear that the effects of secondary HMI tasks on driving behaviour have been extensively studied in a multitude of environments. Results and insights from such studies have been used to formulate computational models and frameworks that can be used to replicate the observed behaviour (e.g. Hankey, Dingus, Hanowski, Wierwille and Andrews, 2000; Salvucci et al., 2005*b*; Liu et al., 2006; Large et al., 2018).

In the context of HMI evaluation, certain guidelines have been published to ensure that the above methods are used in an efficient and structured way to inform decisions about the distracting potential of a prototype HMI design. Such guidelines intend to provide information either on how an HMI should be designed, so that its effects on driving are acceptable (design guidelines), or on how to rigorously, reproducibly and comparably test that those effects are acceptable (evaluation guidelines).

In the two following subsections, some of the established procedures and guidelines for human-centred HMI evaluation are presented, as well as some of the computational models and frameworks that can be used in that context.

2.3.1 Established Guidelines for Procedures and Metrics

Various guidelines have been proposed over the years in attempts to provide robust frameworks of design and usability specifications, that can accommodate HMI evaluation throughout the various design cycles (see Zhang and Smith, 2004; Schindhelm et al., 2004, for extensive reviews on guidelines and standards, still relevant to this day).

The JAMA (2004) guidelines, for example, promote the use of simplified and easily understood displayed images, focusing mainly on navigation interfaces. The "Human Factors Design Guidelines for Advanced Traveller Information Systems (ATIS) and Commercial Vehicle Operation (OVC)", published by the Federal Highway Administration, provide specific guidelines for symbol contrast, colour, font, width-to-height ratio, spacing and number of information units to be put in an interface (Campbell et al., 1997). Both aim at minimising the visual attention demand of an HMI and, hence, its distracting potential.

The Safety and Human Factors Committee of the Society of Automotive Engineers (SAE), approved the 15-second rule (SAE J2364) as a Recommended Practice in 1999 (Green, 1999). The 15-second rule applies mainly to navigation and route-guidance interfaces that require a visual-manual interaction with the driver and is meant to be used as a performance evaluation tool, defining that any navigation task, where the driver interacts with a visual display and manual controls, should take no longer than 15 seconds to be completed.

The Statement of Principles by the Alliance of Automobile Manufacturers covers new information and entertainment technology and devices with visual or manual-visual interfaces, features and functions that are used by drivers when the vehicle is in motion. The statement details a total of 24 principles that need to be followed both during the design stages and measured as performance metrics to ensure the interface is safe enough to go into production (see Group et al., 2006).

Finally, one of the most commonly referenced guidance is the NHTSA guidelines, which prescribe a certain methodology to be followed for a simulator study, as well as which metrics to use to define whether a prototype HMI can be considered safe and fit for production (NHTSA, 2012). Namely, single eye glance, mean eye glance and total eye glance duration away from the road scene need to be measured below predefined thresholds so that the HMI task can be deemed acceptable. Although relatively well established within the Human Factors society, the guidelines have received some scrutiny over the years both in terms of the prescribed experimental procedures as well as the behavioural metrics to be studied (see Aust et al., 2013; Large et al., 2015, for a review). Large et al. (2015), for example, showed that by varying the accompanying prescribed primary driving task complexity, task acceptance can vary, too. In particular, they showed that when drivers are executing a high demand primary driving task, they devote less time on the secondary task (less time looking away from the road) and, as such, tasks that might not be accepted through the conventional testing paradigm were accepted under the increased load scenario.

2.3.2 Computational Models and Modelling Frameworks

As mentioned in the introductory chapter, computational models can be even more cost-effective than the human-in-the-loop methods reviewed above (namely, naturalistic studies and driving simulator studies) and could, thus, prove to be a very valuable tool in the context of prototype HMI evaluation. Different computational models of the interaction during HMI engagement have been proposed over the years, using a variety of methods to predict related performance metrics (e.g. Hankey, Dingus, Hanowski, Wierwille and Andrews, 2000; Salvucci et al., 2005*b*; Liu et al., 2006; Large et al., 2018).

For the rest of this section, such models are presented, grouped by the underlying behavioural approach considered during their development. The following review aims at providing a high-level overview of the available modelling methods and tools. A more in-depth review of these and other relevant models, in terms of their functionality and predictive capacity, is also presented later, in Chapter 6.

Human Computer Interaction Approaches

From the field of Human Computer Interaction, numerous modelling techniques of the interaction of a human with a certain task emerge. Some of the most popular of those techniques are based on the timing of the task itself and are namely Card et al.'s 1983 Goals, Operators, Methods and Selection Rules (GOMS), along with its variations such as the Keystroke Level Model (KLM) (John and Kieras, 1996) and Critical Path Analysis (CPA) model (Lockyer and Gordon, 1991; Baber and Mellor, 2001; Stanton and Baber, 2008).

GOMS and KLM are of particular interest in HMI task modelling as they both represent the interactions between drivers' perceptual, motor, and cognitive systems in terms of individual memories and processes (Card et al., 1983). Both of these modelling techniques are based on the same principle; they model interaction behaviour using a sequential ordering of basis operations (Card et al., 1983). Each basis operation in the sequence that builds the bigger model, is assigned a certain time duration. As a result, the total completion time for a given task can be predicted by the model, given that it is accurately broken down in representative sub-tasks. However, due to this sequential ordering of basis operations in those two modelling techniques, there is no way to represent the overlap between different processing modes (John and Kieras, 1996).

A different approach to modelling interactions using HCI related methods involves the use of Fitts' and Hick's laws. Fitts' Law (Fitts, 1954) is based on Shannon's theorem (Shannon and Weaver, 1949) and can be used to predict the time needed to move to a target using a pointing device. The basic assumption behind Fitts' law is that larger, closer targets require less time to reach than smaller targets, farther away. The Hick-Hyman Law or Hick's Law (Hick, 1952) can be used to predict the time needed to make a decision when presented with multiple options. The two laws have been previously combined to describe tasks as constituting from a *search* and a *point* element (a very accurate representation of visual/manual HMI tasks), with success in predicting static task times (Cockburn et al., 2007) and visual behaviour metrics during HMI task execution (Large et al., 2018).

Control-Theoretic Framework Approaches

Sheridan (2004) suggested a, slightly more complex, framework for modelling driver distraction and, consequently, driver engagement to HMI tasks concurrently to driving, which follows the classic control theory paradigm of a control loop. Considering the primary driving task in a control theory framework (i.e. the driver is the controller and the vehicle is the system), driver distraction can be defined as anything that would affect the control signal between the controller and the system and would, thus, result in less accurate (or noisy) control. Given a transient disturbance in the vehicle feedback (e.g. engagement to a visually demanding HMI task), there is a switch in tasks and control input to the vehicle does not get updated for as long as this disturbance lasts. As a result, since this produces a "blind" open loop, if the disturbance lasts long enough, this could lead to dangerous behaviour (e.g. drifting out of lane or managing speed improperly).

However, despite being qualitatively defined, this approach has not yet been widely utilised towards implementing a computational driver model that can account for dual-tasking while driving (Boer and Spyridakos, 2016, for a sample implementation).

Cognitive Architecture Approaches

Moving towards more realistic representations of the human information processing and task execution mechanisms, different cognitive architectures were created, replicating a range of the human cognitive functions. The Adaptive Control of

Thought-Rational (ACT-R) is a cognitive architecture first introduced in the early 1990s by Anderson (1993). ACT-R in its essence is a framework for understanding and investigating how human cognition works. ACT-R divides human knowledge into two categories; declarative knowledge and procedural knowledge. Declarative knowledge refers to "things that an individual knows and they are aware of knowing" (this is usually manifested by their ability to describe those things to others) (Anderson and Lebiere, 2014). Procedural knowledge on the other hand, refers to "things that an individual possesses and demonstrates knowledge of in their behaviour without being conscious of it" (this is usually demonstrated through tasks we perform in our every day life) (Anderson and Lebiere, 2014). ACT-R has been used in the driving and dual-tasking context, with models of different secondary tasks and under various driving conditions (e.g. Salvucci et al., 2005*a*; Salvucci and Taatgen, 2008). Simpler derivations and implementations of ACT-R have also been created through the years to make the architecture more available and easier to use (e.g. Salvucci and Lee, 2003; Salvucci, 2009).

The Queuing Network - Model Human Operator (QN-MHP) is another cognitive architecture framework, that integrates queuing networks and the procedure / production systems approach, and has been previously integrated with ACT-R (Cao and Liu, 2013). QN-MHP relies on the Natural GOMS Language (NGOMSL) (Feyen, 2002) to describe tasks, which are represented as a set of rules within the architecture (see next section for more details). QN-MHP has been previously used to model driver menu selection and visual search (Lim and Liu, 2004), as well as driver workload (Wu, 2007; Wu and Liu, 2007; Wu et al., 2008).

2.4 Driving Simulator Validity

Since simulators are still widely used for prototype HMI evaluation and all existing models have been based on the behaviour observed under simulated driving, it is important to ensure that driving simulators are reliable and can elicit a realistic behaviour that resembles the one that would be observed in the corresponding real-world conditions.

Driving simulators have predominantly been assessed in terms of their *physical* and *behavioural* validity throughout the relevant literature (Blaauw, 1982; Blana, 1996; Mullen et al., 2011). 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 (Mullen et al., 2011). 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 (Wang et al., 2010). Reed and Green (1999) showed, for example, that increasing the fidelity 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.

When it comes to evaluating performance in different tasks, it has been noted 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 (Blaauw, 1982; Gemou, 2013). Since the work presented later in this thesis (namely in Chapter 4) focuses on the behavioural validity of different driving simulator settings for HMI evaluation, the review hereafter will also focus on behavioural validity alone. Mullen et al. (2011) and Blana (1996) can provide the reader with further details on any of the other types of simulator validity.

2.4.1 Behavioural Validity

Behavioural validity can be further classified into two types; *absolute* and *relative* validity (Blaauw, 1982). 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 (Blaauw, 1982). However, it is most commonly assessed on the basis that the magnitude of the differences has to be the same, too (Godley et al., 2002; Yan et al., 2008; Mullen et al., 2011). As Wang et al. (2010) clarifies, when relative validity is defined as differences in the dependent variable between conditions being of the same order and direction, one refers to an identical rank ordering across conditions. 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). 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.

2.4.2 Factors Affecting Behavioural Validity

The level of behavioural validity of a certain simulator, in contrast to its physical validity as mentioned earlier, is not always proportional to its complexity and characteristics. It is the behaviour observed by the drivers in it that dictates what type, if any, of behavioural validity is achieved. For example, in a study where a medium fidelity, fixed-base driving simulator was used to assess driver interaction with three manual address entry methods (keypad, touch screen and rotational controller), Wang et al. (2010) concluded absolute behavioural validity for the simulator with regards to task completion time. In a different study, assessing five different driving simulators of varying fidelity through four visual/manual tasks, Klüver et al. (2016) concluded relative validity for a high-fidelity, moving base simulator, when considering standard deviation of headway distance. One can infer that both the metric in question as well as the simulator type play an important part in what level of behavioural validity can be concluded. Also, behavioural validity for the same metric may vary between different simulator types as may the behavioural validity of different metrics for a given type of a driving simulator (see for example Klüver et al., 2016).

A multitude of other factors related to the experimental design and methods can also affect the achieved level of behavioural validity. So far, research has focused on two categories of such factors: simulator characteristics and user characteristics (Klüver et al., 2016). The effect of different simulator characteristics on behavioural validity has been previously explored, with regards to horizontal field of view, motion system and the use of a mock-up vehicle cabin.

In general, larger field of view has been found to improve speed choice and lateral control behaviour, hence enhancing behavioural validity for the examined simulator (Kappé et al., 1999; Jamson, 2000, 2001; Rosey and Auberlet, 2014; Klüver et al., 2016). An impaired speed behaviour has been attributed to the use of a mock-up (Rosey and Auberlet, 2014) as well as a motion system (Reymond et al., 2001). Furthermore, employing a motion system has been found to elicit more realistic braking (Siegler et al., 2001) and turning behaviour (Hogema et al., 2012), while using a mock-up alone has a detrimental effect on lateral control, that deviates largely from what is observed under real world driving conditions (Burnett et al., 2007; Rosey and Auberlet, 2014; Klüver et al., 2016). With regards to user characteristics, gender and age had been previously considered by Reed and Green (1999) where they found that older drivers (and particularly older females) demonstrated behaviour in a fixed base simulator that deviated more from reality than that of younger drivers. Klüver et al. (2016) corroborated their findings and also investigated the effect of motion sickness which was found to generally impair driver behaviour and, consequently, behavioural validity.

Surprisingly, there is currently no research addressing the significance of the primary driving task complexity or the relevant complexity of the secondary tasks examined, when evaluating the behavioural validity of a simulator. For example, as discussed earlier, drivers react to the primary driving task complexity by adjusting their behaviour to mitigate the distracting effects of their interaction with a secondary task. Recently, Large et al. (2015) showed that a more complex driving scenario yielded better task acceptance rates by the NHTSA guidelines, offering higher ecological validity. Hence, varying the complexity of the primary driving task could potentially also vary the correspondence of the driving behaviour observed in the simulator to the one observed in the real world.

2.4.3 Desired Level of Behavioural Validity for HMI Evaluation

It is clear that many different factors affect the behavioural validity of a driving simulator. 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 (Allen et al., 1979). 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 (Törnros, 1998; Reed and Green, 1999; Wang et al., 2010). 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 (Gemou, 2013).

For example, a manufacturer interested in conducting comparative testing between different prototype interface designs, in order to identify which one of the interfaces could be associated with longer off-road glances, could make that decision with the simulator used possessing relative validity only. 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.

Some additional review of related literature on behavioural validity will follow in Chapter 4, where the focus is on specific findings from dual-tasking simulator validation studies (i.e. concluded behavioural validity for different metrics across different simulator settings).

2.5 Key Research Gaps

Considering the review presented above, a number of key research topics can be identified that have not been previously explored and would be important to be pursued. When it comes to driving simulators, there has been no structured investigation on how the level of behavioural validity varies with simulator type and metrics in question. Regarding driver behaviour under dual-tasking conditions, although the effects on driving performance are well understood, the research on identifying how drivers decide to share their visual attention between the primary driving task and the secondary task in question is very limited. Finally, despite the fact that multiple computational models have been published, their validation has been limited to data sets associated with their parameter fitting. Consequently, it is difficult to determine whether these models are performing well due to actually capturing the underlying mechanisms of dual-tasking in the driving context, or just overfitting the validation dataset. The work presented in Chapters 4, 5 and 6 aspires to investigate and provide new insights to these issues.

Chapter 3

Empirical Studies and Methods

The present chapter describes the HMI evaluation driving study that was carried out in the scope of this thesis. This study consisted of two data collection experiments, one of which took place in a driving simulator, while the other took place in a real world setting (test track). All testing adhered to the ethical guidelines laid out by the University of Leeds Research Ethics Committee. The first experiment was conducted in the University of Leeds Driving Simulator (UoLDS) and will be henceforth referred to as the UoLDS experiment. The second experiment was conducted in the Jaguar Land Rover (JLR)'s Emissions Circuit test track in Gaydon, Warwickshire and will be henceforth referred to as the Gaydon experiment.

The structure of the rest of this chapter follows the conventional way in which scholars have been outlining their experimental methods in the field of automotive Human Factors (e.g. Engström et al., 2005; Kountouriotis and Merat, 2016).

3.1 Participants

A total of 12 participants completed the UoLDS experiment, six of which were females (mean age 37.17 $\pm$ 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. Previously, they had participated in driving studies related to autonomous vehicles and, in particular, to manual control handover. The participants were compensated with £15 for their time.

A total of 11 participants completed the Gaydon experiment, two of which were females (mean age 36.55 ± 11.93 years). Potential participants for the Gaydon experiment were all JLR employees and were contacted through the internal JLR communication network. About half of the participants had previously driven in the test track. None of the participants was in any way involved in the development or evaluation of prototype HMI designs as part of their job specification in JLR. The participants were not compensated monetarily for their time.

The sample size was initially set at 12 participants for each of the two experiments, with the option to run follow-up experiments later, if needed. This was decided after taking into account specific constraints, such as simulator and test track availability, as well as the difficulty in sourcing large numbers of participants. At this point, it is worth noting that one of the participants in Gaydon did not attend the experiment at their defined timeslot and, due to time constraints, could not be replaced, hence the discrepancy in sample sizes for the two experiments. After the initial analysis of the collected data, the primary results appeared to be generally aligned with the existing literature (see Chapters 4 and 5) and, hence, no additional experiments were conducted.

Tables 3.1 through 3.3 provide details on participant demographics and driving experience.

Table 3.1: Participant Demographics

	Female	**Male**	**Mean Age**	**STD Age**	**Range Age**
UoLDS	6	6	37.17	10.42	24 - 57
Gaydon	2	9	36.55	11.93	23 - 57

Table 3.2: Participant Driving Experience - Annual Mileage

	Mean Mileage	STD Mileage	Range Mileage
UoLDS	7166.67	4217.57	1000 - 15000
Gaydon	11454.55	3251.57	6000 - 15500

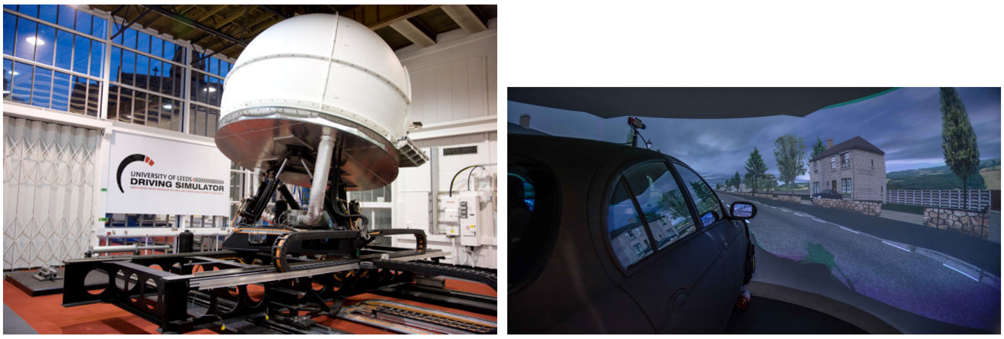
Table 3.3: Participant Driving Experience - Years Driving

	Mean Years	STD Years	Range Years
UoLDS	19.27	11.21	6 - 40
Gaydon	18.55	11.93	5 - 39

3.2 Design and Procedure

3.2.1 Materials

The University of Leeds Driving Simulator (UoLDS) consists of a 4 m. diameter spherical projection dome, mounted on an eight-degree-of-freedom moving base (see Figure 3.1). 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



(a) UoLDS dome. (b) UoLDS interior.

Figure 3.1: The University of Leeds Driving Simulator (UoLDS). The pictures are generic and were not taken during the experiment.

simulator during the present study; a setting with no motion (Simulator Fixed Base (Sim. Fix.)), where the vehicle handling feedback was provided to the driver through the simulator visual scenery and the steering torque of the vehicle model, and a setting with limited motion, where the simulator dome was moving using the 6 degrees of freedom hexapod. The hexapod supplied roll, pitch and yaw movements, providing the drivers with motion cues for perceiving acceleration.

The decision for using those two motion conditions was based on an extensive literature review, presented in Section 4.2, where it was found that no motion has been the most commonly used configuration in HMI evaluation driving studies, while a hexapod only motions has not been previously used in this context (see Table 4.1 for a summarised reference). This way, a common point of reference was established with the existing literature, while also investigating an existing research gap.

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 using a v5 Seeing Machines faceLAB eye-tracker, mounted on the dashboard of the simulator vehicle cab. Data was recorded at a similar frequency of 60 Hz and with a gaze direction measurement accuracy between 0.5° and 1°. Finally, video streams were recorded through 4 cameras with the following configuration:

- One camera inside the vehicle cabin facing the driver.
- One camera inside the vehicle cabin facing over the driver's shoulder.
- One camera inside the Simulator dome, on the roof of the vehicle cabin facing the scenery ahead.
- One external camera.

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 (see Figure 3.2). 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 and with a gaze tracking accuracy of 0.5°. Finally, video streams were

(a) Subject car model. (b) Subject car interior.

Figure 3.2: Subject car used in the Gaydon experiment. The pictures are generic and were not taken during the experiment.

recorded through the eye-tracking glasses camera (located at the binocular focal point) and through 3 VBOX cameras in the following configurations:

- One camera inside the vehicle cabin facing the driver.
- One camera inside the vehicle cabin facing over the driver's shoulder.
- One camera inside the vehicle cabin facing the road ahead.

Video streams from the different VBOX cameras were synchronised and time-stamped.

3.2.2 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 3.3 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. Figure 3.4 illustrates an example of the digital digital environment replicating Emissions Circuit in UoLDS.

Figure 3.3: Proving Ground facilities layout in Gaydon.

3.2.3 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 (Lead Vehicle constant speed condition (LVc)), the lead vehicle was travelling at a constant speed of 50 mph, as per the scenario prescribed by NHTSA (2012). In the second scenario (Lead Vehicle varying speed condition (LVv)), the lead vehicle was travelling at a varying speed between 60 and 70 mph, following a semi-randomised speed profile.

The profiles were obtained from the processing of recorded, real world speed data. This speed data recording took place prior to both experiments and the driver was a JLR employee that did not participate in any experiment. The driver was instructed to drive within the defined speed range in the test track, accelerating and decelerating, as they normally would. Acceleration and deceleration patterns were extracted from the recorded speed data, and were used to define the time needed to accelerate or decelerate from one target speed to the next. This way, a speed profile could be generated for a set of random target speeds. These speed profiles were directly coded in the lead vehicle in the simulator and, for the Gaydon experiment, were provided to the lead vehicle driver to replicate while driving.

The two scenarios were chosen to represent different levels of primary driv-

Figure 3.4: Two instances of the Emissions Circuit in Gaydon (left) with their corresponding instances from the simulated version of the test track in UoLDS.

ing task difficulty, that can elicit a different HMI interaction behaviour form the drivers. For instance, Large et al. (2015) have previously shown that there is a significant effect of primary task demand on secondary task performance. In particular, they showed that increased primary driving task demand led the drivers to dedicate shorter glances towards the HMI and complete HMI tasks faster.

Unlike the usual approach in car-following scenarios, the lead vehicle in this case was not bound to the subject vehicle, i.e. the lead vehicle was moving independently. Consequently, some instances occurred where the distance between the two vehicles grew beyond the desired one. However, there were no instances where the participants completely "lost" the lead vehicle and they were able to recover the distance by accelerating between task executions.

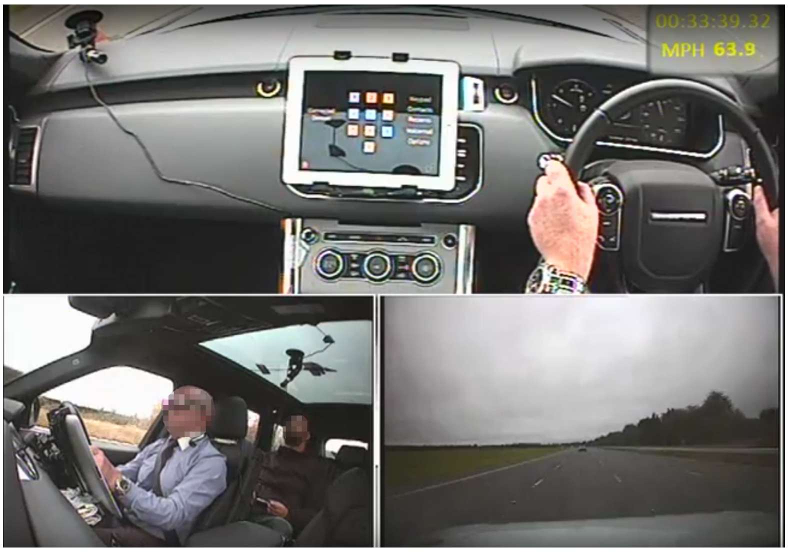
3.2.4 HMI Tasks

Three visual-manual HMI tasks were used in both experiments of this study, with all of comprising touch-screen controller elements. The HMI tasks were implemented so that each one had a different level of difficulty (easy, medium and hard), based on the number and types of interaction they needed to be completed, similar to the approach used in Large et al. (2015). This decision was made in order to facilitate more thorough comparison of driving behaviour between the different conditions (driving environments and scenarios). The particular tasks were chosen as representative of real world HMI tasks, after discussions with JLR and based on an in-house prototype interface.

The HMI tasks, although resembling functionalities one might find in a production vehicle, were independent of any vehicle system and, thus, had no effect in any of its functionalities. All tasks were implemented so that once finished they automatically returned to their home screen, to facilitate easy repetition while driving (i.e. have no need for resetting). 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 top part of the central console, to the left of the driver (see Figure 3.5). The tablet was positioned to virtually the same position in both vehicles (real world and simulator) and was tilted back at a vertical angle of approximately 35.5°.

Finally, participants were trained on how to perform each HMI task both while stationary and while driving.

The upcoming subsections describe in detail the aspects of each task.



(a) Setup in the Gaydon experiment.

(b) Setup in the UoLDS experiment.

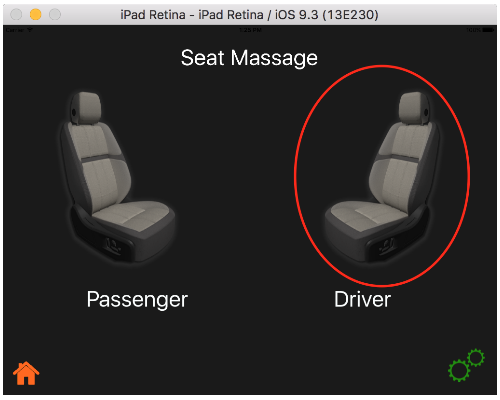
Figure 3.5: In-vehicle HMI setup.

Easy Secondary Task

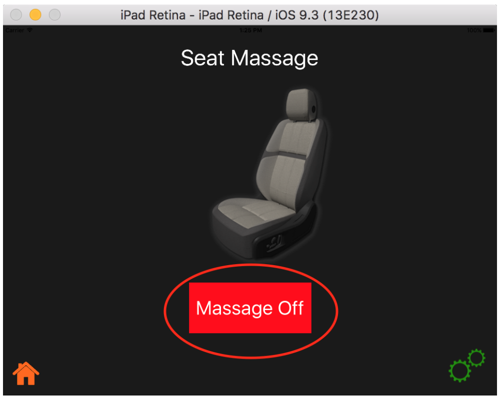
Simulated function: Activate massage for the driver seat.

Interactions: Press, Press, Press.

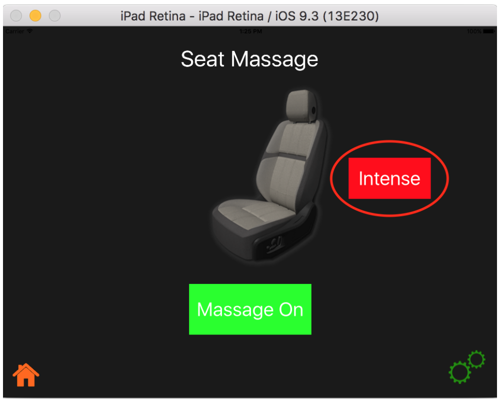
The driver had to press three buttons (see Figure 3.6) to activate the massage on their seat. Initially, they needed to press the driver seat icon on the task home (first) screen (Figure 3.6(a)), then press on the "Massage" button on the second screen (Figure 3.6(b)) and, finally, press on the "Activate" button on the third screen (Figure 3.6(c)).



(a) First screen - Press.



(b) Second screen - Press.



(c) Third Screen - Press.

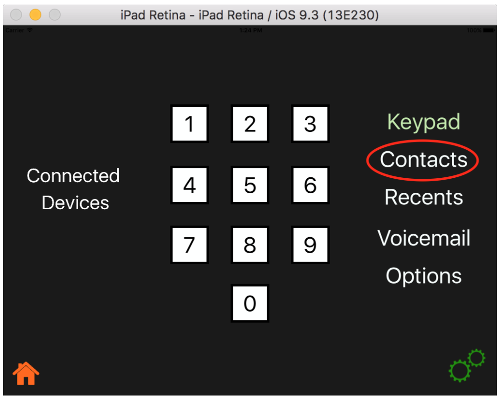
Figure 3.6: Easy HMI task. The home screen is illustrated on top and the succession is from left to right, top to bottom, as denoted by the captions. Red circles denote the areas on the screen (virtual buttons) that the user needed to press in order to move to the next screen and complete the task.

Medium Secondary Task

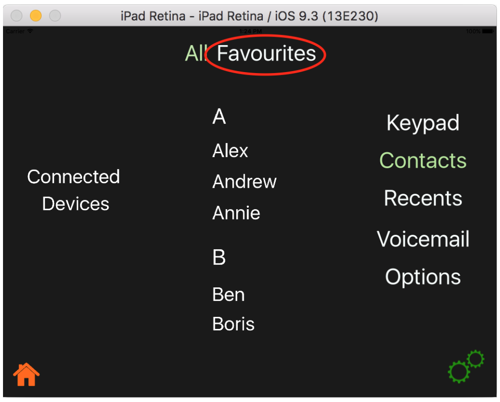
<u>Simulated function</u>: Call a contact on their Home number (the contact was always the same, namely “Mom”).

<u>Interactions</u>: Press, Press, Press, Press.

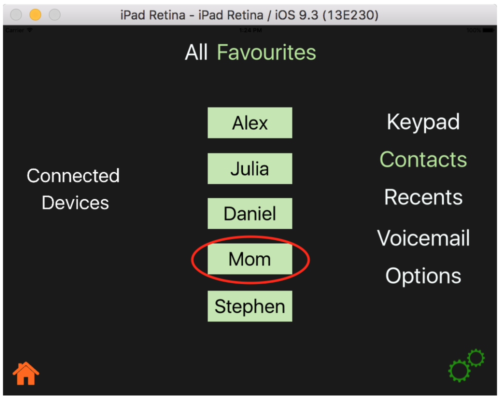
The driver had to press four buttons, representing items in a list, to make a call to a specific contact (namely “Mom”) in their contacts list (see Figure 3.7). Initially, they needed to press the “Contacts” button on the task home (first) screen (Figure 3.7(a)), then press on the “Favourites” button on the second screen (Figure 3.7(b)), then press the “Mom” button on the third screen (Figure 3.7(c)) and, finally, press on the “Home” button on the final (fourth) screen (Figure 3.7(d)).



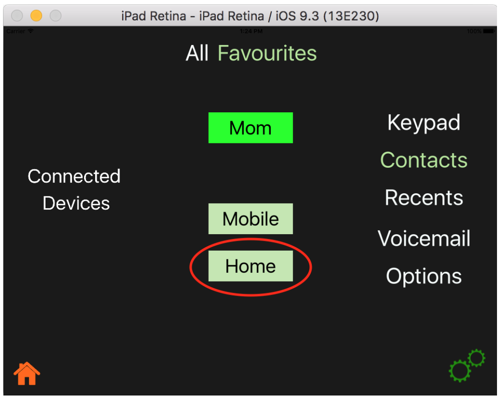
(a) First screen - Press.



(b) Second screen - Press.



(c) Third screen - Press.



(d) Fourth screen - Press.

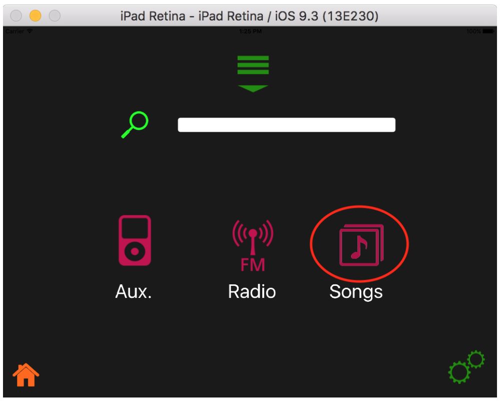
Figure 3.7: Medium HMI task. The home screen is illustrated on top and the succession is from left to right, top to bottom, as denoted by the captions. Red circles denote the areas on the screen (virtual buttons) that the user needed to press in order to move to the next screen and complete the task.

Hard Secondary Task

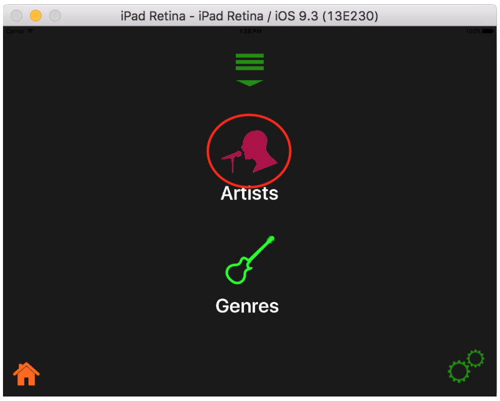
Simulated function: Play a song of a specific artist (the artist was the same every time, namely “Deep Purple”).

Interactions: Press, Press, Scroll, Press, Press.

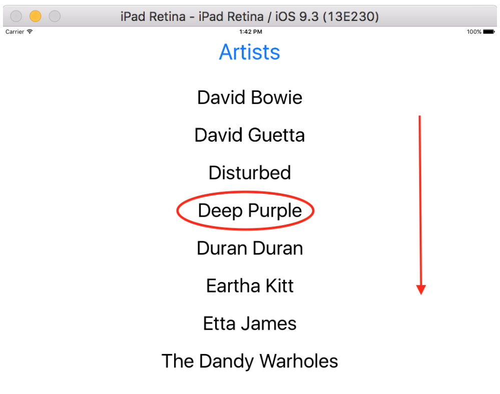
The driver had to select a song from their playlist and make it play (see Figure 3.8). Initially, they needed to press the “Songs” button on the task home (first) screen (Figure 3.8(a)), then press the “Artists” button on the second screen (Figure 3.8(b)), then scroll down until they found “Deep Purple” and press it (Figure 3.8(c)) and, finally, press the play button on the final (fourth) screen (Figure 3.8(d)). At this point, it is worth noting the shared disappointment of the participants’ majority that “Highway Star” did not play on the sound system.



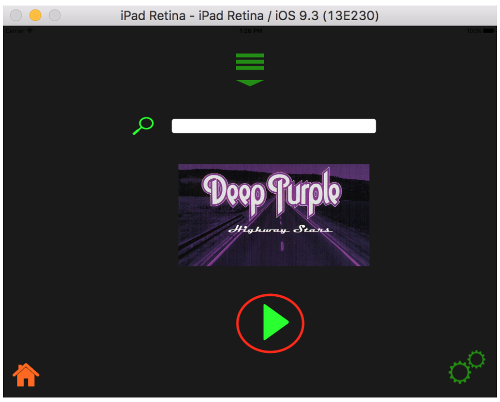
(a) First screen - Press.



(b) Second screen - Press.



(c) Third screen - Scroll and Press.



(d) Fourth screen - Press.

Figure 3.8: Hard secondary task. The home screen is illustrated on top and the succession is from left to right, top to bottom, as denoted by the captions. Red circles denote the areas on the screen (virtual buttons) that the user needed to press in order to move to the next screen and complete the task. The red arrow denotes that the user needed to scroll down to locate the desired target.

3.2.5 Experimental Design

A three-factorial design was used with Environment being a partly between-subjects and partly within-subjects factor, while Scenario and HMI task were within-subjects factors. Environment had three levels: Real, Fixed Base and Hexapod. The same set of participants were exposed to both the Fixed Base and Hexapod levels, while a different set of participants was only exposed to to the Real level. 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. However, 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).

3.2.6 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 (e.g. location of participant briefing).

For the UoLDS experiment, the participants were welcomed and briefed in the simulator briefing area (see Appendix A), before being asked to provide signed consent for their participation in the study (see Appendix B). Participants were then introduced to the HMI tasks and trained on how to perform them while static (still in the briefing area). 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 (see Appendix C). Next, participants took a familiarisation drive with the simulator in the motion setting used throughout the experiment, with the lead car travelling at a constant speed and the experimenter in the vehicle. This drive aimed at participants acquainting themselves both with the simulator and the concurrent driving and secondary tasks. The familiarisation drive was concluded when the participants declared ready to move on to data collection, which seemed to happen after the completion of one lap of the test track..

During the data collection phase, similarly to the familiarisations drive, the experimenter was sitting in the back seat of the vehicle, behind the driver. At this stage, drivers were required to complete four full laps of the test track. 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 each of the HMI tasks on various instances, when instructed by the experimenter. Each one of those three laps was dedicated to the execution of one of the HMI tasks. The experimenter would denote those executions instances 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 initiate a task execution when they were in full control of the vehicle and at least 3 seconds after a previous HMI task execution. These four laps of the test track consisted a drive, each one lasting approximately 6 minutes. Each participant had to complete 4 drives in total, one for each combination of simulator motion and scenario, namely fixed base and constant speed, fixed base and varying speed, hexapod and constant speed, hexapod and varying speed.

After the completion of each drive, the participants were given a subjective questionnaire regarding the perceived difficulty, acceptability and frequency of the HMI tasks, as well as the subjective levels of realism and discomfort (see Appendix C). The results from the subjective questionnaires were not used in the subsequent analyses and, thus, are not reported here.

For the Gaydon experiment, the procedure was identical, with the exceptions that the participant briefing took place in the subject vehicle, while stationary in a parking lot and that each participant had to complete only two drives (one for the scenario of constant speed and one for the scenario of varying speed). Participants in the Gaydon experiment performed an additional drive after a break, where they had to complete a mental arithmetic task while driving (documented in Appendix C). This drive always took place after the familiarisation drive and the two main drives of the HMI experiment and after the participants had a short break. Since driving behaviour during the execution of non-visual tasks was not in the scope of this thesis, the collected data were not analysed and are not presented 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 scenarios were counterbalanced. Appendix D provides information on what the actual condition counterbalancing was for the two experiments. It is worth noting here that, in order to test all possible motion and scenario combinations in the simulator, a total of 24 participants would be needed. Consequently, only half of possible combinations were tested in the UoLDS experiment. Finally, since the task executions were not performed at fixed intervals, different participants performed a different number of task executions, depending on their speed of execution. Overall, a total (across all conditions and tasks) of $1,527$ task executions were recorded in the UoLDS experiment, while 662 task executions were recorded in the Gaydon experiment.

3.3 Initial Data Reduction

3.3.1 Identifying and Extracting Individual HMI Task Executions

The main challenge in the data reduction step was to locate and extract each HMI task execution separately, to allow analysis of individual task executions, and to exclude all non-task data from analysis.

For the UoLDS, the experimenter denoted the initiation of an HMI task execution by raising his hand within the field view of one of the internal simulator cameras when they administered the “Engage now” command and lower it back down after the participant had finished their task. Normally, participants had to declare so by saying “Done”. However, more often than not, participants forgot to do so or used different wording (such as “OK” or “Finished”). After data collection, manual annotation of the recorded videos gave the desired task segments. The raising of the experimenter’s hand was used as the moment of initiation of an HMI task and, in order to maintain consistency between participants, the moment when a participant returned their eyes on the road after completing a task was used as the moment of task completion.

An analogous approach was taken with the Gaydon experiment data. In that case, however, audio signal was present in the SMI recordings and was used instead of the raising hand approach. In particular, as moment of initiation was considered when the utterance “Engage now” was said by the experimenter and, similarly to above, the return of the participant’s gaze to the road ahead defined the moment of task completion. One of the biggest challenges regarding data segmentation was the synchronisation of BeGaze and VBOX data. This was achieved by generating an identical visual scene in both video streams (namely, the lead car braking three times) so that a synchronisation time frame between the two recordings could be found. This way, each eye-tracking data segment could be matched with its corresponding vehicle data segment (unlike the simulator data where they are logged in this manner by default). Unfortunately, due to a time offset between the SMI and VBOX logs, as well as the difference in frame rate, the mapping is not as precise as the one in the simulator data, but sufficient for the type of analysis conducted later.

3.3.2 Defining Where the Drivers Look

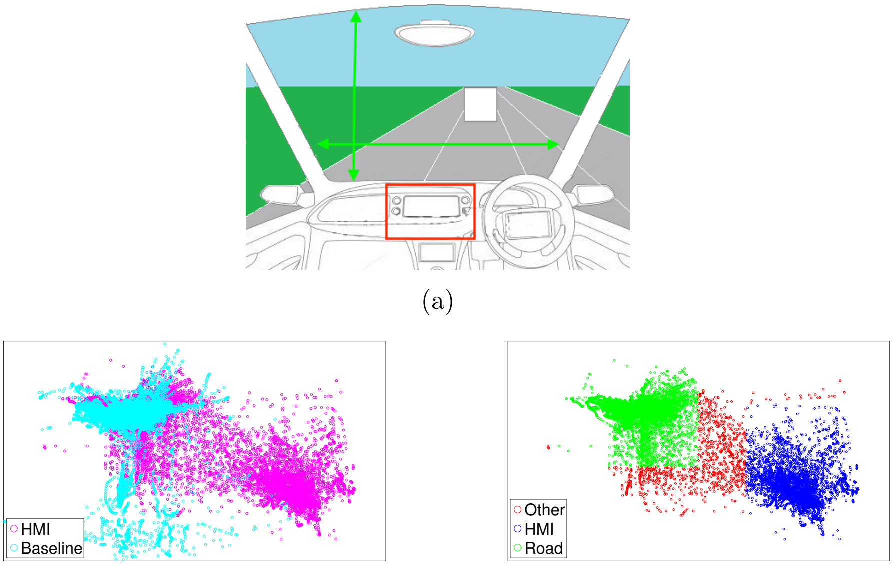
Eye-tracking data for the Gaydon experiment were manually annotated, frame by frame, using the BeGaze analysis software. A 2-D model of the driving environment was created and Areas Of Interest (AOIs) were defined within it. Next, for each fixation point, its location was mapped within one of the AOIs. Figure 3.9(a) shows the schematic used for this mapping of the Gaydon experiment fixation data.

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 (see Figure 3.9(b)). Based on this comparison, AOIs were defined for each driver to help identify where they were looking at each instance (see Figure 3.9(c)). Later, a random sample of task segments was visually compared against video data to ensure the AOIs were properly defined.

For both the real world and simulator experiments, three major 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". Only glances falling within the Road Ahead AOI (on-road glances) and HMI (off-road glances) were considered for analysis. Due to the format the Gaydon eye-tracking data were recorded in (fixation points do not have a global reference point), the ISO glance coding process could not be used here (15007-1:2014, E)). Instead, and in order to be consistent across both datasets, glance durations were calculated by aggregating consecutive fixation points within the same AOI.

3.3.3 Data Cleaning

Initially, all instances where the task executions were incorrect, incomplete or any external factors interfered were removed from the dataset. Such instances were either due to slips, errors (participants pressing slightly off-target) or, in the Gaydon experiments, where participants had to change lanes amidst a task execution. Overall, the number of such instances was very small and the data removed as a result of it amounted to less than 1% of the total recorded data.



(a) (b) (c)

Figure 3.9: AOIs used for glance coding. (a) Schematic used for glance coding of the Gaydon eye-tracking data. The area defined by the green arrows denotes the “Road Ahead” AOI, the red rectangle denotes the “HMI” AOI, while the remaining of the image denotes the “Other” AOI. (b) Fixation points during baseline driving and concurrent HMI task execution driving. The overlap on the top left part indicates where the “Road Ahead” AOI would be defined, while the bottom right cluster denotes the “HMI” AOI. (c) Glance coding AOIs as they were defined for one of the UoLDS participants.

Regarding minimum glance duration, there is currently no agreement in the academic community as to what threshold should be adopted. Salvucci and Goldberg (2000), for instance, defined the minimum required glance duration at 100 ms. Fixation times during reading (which approximates HMI interactions well since drivers need time to register information from the screen), on the other hand, have been found to average around 225 ms (Rayner, 1998). Based on that and in accordance with previous studies from the author’s research group, 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 (Broström et al., 2013; Louw et al., 2017). Moreover, since both eye-tracking systems automatically classify fixation points based on their quality, glances consisting of more than 50% of poor quality fixation points (as those were annotated by the eye-trackers) were also not used.

Finally, after the manual annotation of the simulator data, one of the participants appeared to have been experiencing symptoms of motion sickness, which they had not disclosed to the experimenter at the time of the trial. That participant's data was removed, an action that was later found to have no effect in the reported results.

3.3.4 Data Loss

Due to equipment malfunction, parts of eye-tracking and vehicle data were not recorded during the Gaydon experiment. In particular, eye-tracking data were not recorded for two drives, by two different participants, evenly divided between the two scenarios (i.e. one of the drives was of the constant speed scenario while the other one was of the varying speed scenario). Regarding vehicle data, small segments (affecting single task executions at worst) were not recorded at different stages for different participants. Overall, the portion of data lost in this way amounted to less than 10% of the total amount of collected data for the Gaydon experiment.

Chapter 4

Behavioural Validity of Driving Simulators for Prototype HMI Evaluation

The present chapter focuses on the behavioural validity of different driving simulator types in the context of HMI evaluation[1]. As discussed earlier (in Chapter 2), the degree of behavioural validity can vary both across different driving simulator types, as well as across different evaluated metrics.

The behavioural validity analysis presented in the following sections focuses on the comparison of various performance metrics, related to concurrent driving and HMI task execution, between simulator and reality. The final results are presented in the form of a collective behavioural validity matrix that draws insights from the analysis of the present experiments (see Chapter 3), as well as already reported results in the existing literature. This way, the entire range of simulator types used in HMI evaluation related studies is evaluated with regards to the behavioural validity levels that they can achieve. 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.

[1] A version of this chapter was presented at the *6th* International Conference on Driver Distraction and Inattention and has been submitted for review to the IET Research Journal.

4.1 Methods

4.1.1 Review of Related Literature

An extensive literature review was performed to obtain previously published results, related to driving simulator behavioural validity in the context of HMI evaluation. The focus was on studies that used an HMI task while driving in their experimental design and that conducted tests both in real world and simulated conditions (see the Literature Review Results section below for more details on literature sourcing methodology and inclusion criteria). Based on the reported results, a level of behavioural validity was concluded for each simulator setting, broken down by metric of interest.

4.1.2 Data Analysis

Statistical analysis was conducted on the collected data, to identify main effects of the three factors (Environment, HMI Task and Scenario), as well as the effects of their two-level and three-level interactions (refer back to Chapter 3 for details on the experimental design and procedures). Linear mixed effects modelling (Fisher, 1919) was used as a technique to identify the existence of effects, while Cohen's *d* was also calculated to provide an estimate of the effect sizes (Cohen, 1988).

Although linear mixed effects models have not been as widely used in the psychology and human factors space as other statistical methods (e.g. ANOVA), they do offer some additional advantages over other, well established methods, which is why they were chosen as the analysis method in this case. First, linear mixed effects models are very well-suited for repeated measures data analysis, since they take into account the hierarchical structure of the data. This is particularly important, as distinct observations are not totally independent. Observations across a single participant are usually more similar with one another than they are with observations across different participants. Second, linear mixed effects models are robust in handling missing data, something that, apart from unrecorded data, can also occur with experiments where there is an unbalanced number of repetitions per participant. Traditional statistical methods such as ANOVA, on the other hand, require complete cases to generate results with the most statistical power. This means that the missing cases should ideally be removed by list-wise deletion or replaced with corresponding group mean value. Both of these approaches can impair the model's statistical power and affect the quality of the results. For the present analysis, if such a method were to be used, nearly half of the data would be eliminated.

4.1.3 Establishing Level of Behavioural Validity

To conclude the level of behavioural validity for each simulator setting and metric, the approach described by Wang et al. (2010) was adopted, similarly to Klüver et al. (2016). Relative validity was established when the ranking of the HMI tasks and their main effect was 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.

4.2 Literature Review Results

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 internet search engines (Scopus and Google Scholar), review of reference lists of other relevant publications and review of publications that referenced those (as relevant publication were defined the ones reporting results of HMI performance studies). Moreover, this work was presented on the the 6th 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.

An initial search against publication titles, abstracts and keywords was made on Scopus using the following term: ("behavioural validity" OR "behavioural fidelity") OR (validity OR fidelity) AND "driving simulator". The search results were restricted by excluding articles related to medicine or computer science, which yielded 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. Consequently, meta-analysis studies (e.g. Caird et al., 2008), studies where simulator validity was only evaluated by using a simulator setting as the baseline criterion (e.g. Jamson and Jamson, 2010) and studies where driving simulator validity was not investigated in the context of HMI interaction (e.g. Blaauw, 1982; Jamson, 2001) were not included in the review.

The above selection process resulted in the following two publications: Wang et al. (2010) and Klüver et al. (2016). Together with Santos et al. (2005), Victor et al. (2005) and Engström et al. (2005), 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: Reed and Green (1999); Baumann et al. (2004); Santos et al. (2005); Victor et al. (2005); Engström et al. (2005); Pettitt et al. (2006); Bach et al. (2008); Wang et al. (2010); Knapper et al. (2015); Klüver et al. (2016).

For the papers where behavioural validity was directly investigated by the authors, their conclusions were used directly here. That was the case for the results described in Reed and Green (1999), Wang et al. (2010), Knapper et al. (2015) and Klüver et al. (2016). Reed and Green (1999), used a fixed base simulator, featuring a cabin with narrow field projection, and testing a phone dialling task. The simulator was found to achieve absolute validity for speed but relative validity for lane position, throttle position and steering wheel angle. Wang et al.

(2010) used a similar simulator setting and a surrogate visual/manual HMI to Reed and Green (1999) and found it to achieve absolute validity for initial response time, mean task duration, total glance time, percentage of time looking on the road, as well as standard deviation of speed. The results were inconclusive for average speed and standard deviation of lane position (SDLP), with the authors concluding that the simulator could potentially achieve absolute validity. Finally, the simulator was found to achieve relative validity for glance frequency, i.e. the number of glances employed towards the HMI. Knapper et al. (2015) used the occlusion method in a way-finding task, and found it to achieve relative validity both for mean and standard deviation of speed. Finally, Klüver et al. (2016) used 5 different simulator settings (a desktop simulator, two fixed base simulators featuring a cabin, one with narrow and one with wide field projection, a hexapod and lateral motion simulator and a hexapod and longitudinal motion simulator), measuring three performance metrics of an address entry task. All simulators were found to achieve relative behavioural validity for standard deviation of headway distance. The two simulators employing motion managed to achieve absolute validity for SDLP, while the same simulators along with the higher fidelity fixed base one also achieved absolute validity for task completion times. Finally, the lower fidelity simulators achieved relative validity for task completion times and SDLP.

For the remaining papers, where only analysis results were reported, the level of behavioural validity was inferred by examining statistical significance scores and value plots. Following the methodology proposed by Wang et al. (2010) and described in the previous section, the task order ranking, along with the main effect of task was evaluated to infer relative validity and, then the main effect of environment was considered to conclude absolute validity. Given that in many cases important information were missing (e.g. statistics on environment effects), some of the results were inconclusive.

The results obtained from the aforementioned exercise, are summarised in Table 4.1. Results are grouped by publication and indicate the level of behavioural validity (relative or absolute) concluded for each simulator setting and evaluation metric. When no level of validity could be concluded, the respective cell was classified as “N/A”. For cases were the level of behavioural validity could only be assumed after some assumptions the classifications “possibly absolute” and “possibly relative” were used instead.

		Simulator Type						
Reference	**Metrics**	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 4.1: Levels of behavioural validity achieved by different simulator configurations for different metrics, in the context of HMI evaluation, based on the existing literature. **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.

4.3 Data Analysis Results

Linear mixed effects models were fitted for each metric using MATLAB and the built-in fitlme function and statistical significance for main and interaction effects was reported in the form of p-values, at the $a = 0.05$ confidence level. Initially, different models of varying complexity, in terms of their random effects component, were fit to all dependent variables. The fit of each model was evaluated based on their AIC score (Akaike, 1974) to identify the one that best represented the observed data. The majority of the dependent variables were best represented by the model that included a fully varying slope and intercept per participant, i.e. the maximal random effects structure justified by the data, as suggested by Barr et al. (2013) and also adopted by Klüver et al. (2016).

With y representing the dependent variable (any behavioural metric of interest), $task$ the HMI task (easy, medium and hard), env the driving environment (simulator and real world) and $scen$ the driving scenario (constant and varying speed), the model that was used can be described by the following equation:

$$y \sim task * env * scen + (1 + task * env * scen|part) \tag{4.1}$$

For the remaining models, a comparison of reported effects was performed and no difference in significance levels was found. Hence, the results of the most complex model described above were used. 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 was already cleaned (see Chapter 3) and all observations were valid. However, in the interest of ensuring that the models do not provide inflated results, the same models were also fitted to outlier-free data, with no significant differences in their results.

As Wang et al. (2010) argued, in cases where the values of a dependent variable measured in the field do not differ significantly across conditions, it can be difficult to use the rank ordering criterion. This raises an issue about using statistical significance, alone, as a measure of behavioural validity. Hence, in order to get an estimate of the actual effect for cases like these, Cohen's d was also calculated (as the difference in means divided by the pooled standard deviation) to provide such an estimate (Cohen, 1988). Since the focus of this analyses was to identify differences between simulator and real world, the effect sizes were calculated pairwise for all environments on the average metric value across all three tasks. This averaging over the tasks was conducted so that Cohen's d would provide a more complete insight into the drivers' behaviour, as that was displayed under different conditions. In terms of assessing the actual effect size, an effect of 0.2 was considered small, 0.5 was considered medium, 0.8 was considered large and thereafter was considered very large (Cohen, 1988).

The interaction plots are presented in pairs; one plot for each scenario (constant speed, varying speed). The x-axis shows the three different environments while the y-axis shows the metric in question. The raw metric means for each task are plotted as points connected with dotted lines, along with errorbars representing their 95% confidence interval for the mean. Moreover, effect sizes are overlaid to denote the magnitude of the differences in the metric between environments. For no effect of environment, each task should be represented as a straight line parallel to the x-axis. For no effect of task, task lines should overlap. Finally, for no effect of scenario, the two plots should be identical.

4.3.1 HMI Task Performance

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 4.1 illustrates the total time needed to complete each HMI task, across all driving conditions for all participants.

A significant main effect of task was found ($F(2, 2048) = 47.39$, $p < .001$), with the tasks consistently ordering as easy, medium, hard from the least to the most time required for completion.. However, a significant interaction effect between task and scenario was also observed ($F(2, 2048) = 3.95$, $p = 0.02$). In this case, although the interaction effect hinders the value of the main effect of task observed, since the effect of task is consistent across environments (i.e. there is no interaction effect of environment and task) the main effect of task still holds value for the behavioural validity assessment purposes. In particular, the interaction effect was driven by the medium task requiring more time to be completed in the varying speed scenario ($p = 0.014$). Although absolute values of completion times were slightly higher in the real world, that difference was not statistically significant, hence, there was no effect of environment ($F(2, 2048) = 0.46$, $p = 0.63$). Finally, differences between the two scenarios also appeared to be negligible ($F(1, 2048) = 0.46$, $p = 0.34$).

Given the identical ordering of tasks across all three environments (from smallest completion time to largest: easy, medium, hard with no interaction effect of environment and task $F(4, 2048) = 1.19$, $p = 0.314$), the absence of an effect of environment ($F(2, 2048) = 0.46$, $p = 0.63$), and the small effect sizes observed, absolute validity can be concluded for both the fixed base and the hexapod simulator for task completion times, for both scenarios.

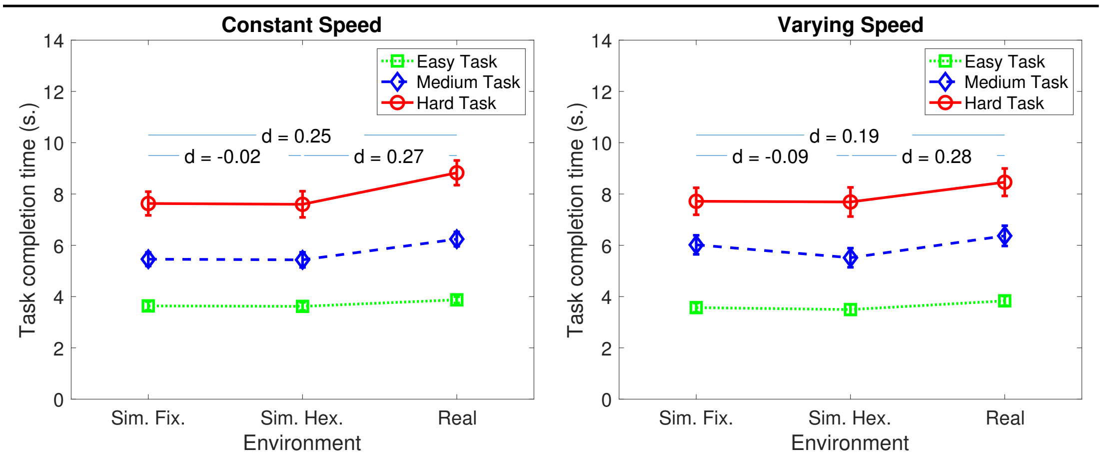


Figure 4.1: HMI task completion times.

4.3.2 Gaze Behaviour

Total Off-road Glance Duration

This metric was calculated as the sum of the durations of all glances towards the HMI during a task execution. Figure 4.2 shows the glance durations towards the HMI for each task across conditions for all participants. After initial model fitting, the model residuals were found to be non-normally distributed. A log transformation was applied to both datasets and models were re-fitted, this time showing no deviation from homoscedasticity or normality for the residuals.

A significant main effect of task was observed ($F(2, 1834) = 49.39$, $p < .001$), which was consistent across environments, with no interaction effect observed ($F(4, 1834) = 0.46$, $p = 0.765$). Additionally, no effect of environment or scenario was found ($F(2, 1834) = 1.64$, $p = 0.195$ and $F(1, 1834) = 0.12$, $p = 0.734$ respectively). Drivers appear to devote less visual attention to the HMI in the real world, something that can be justified by the increased driving demand due to higher potential risk. This is an indication, as discussed earlier, in Chapter 2, that drivers tend to self regulate and adjust their attention sharing strategies according to the demands of the primary driving task. Also, this points to an interesting potential explanation, that drivers perhaps develop a strategy where they learn to perform the tasks with as little visual attention sharing as possible, to ensure they remain within safety vehicle control margins.

Glance times for the medium and hard tasks are almost identical in the fixed base simulator for both scenarios. The two tasks differ more noticeably in the hexapod and real world, with those differences slightly more amplified in the varying speed scenario. Since the two tasks are practically indistinguishable, their relative ordering cannot be considered. Nevertheless, since the ordering of the easy task with either the medium or the hard is consistent across environments, given the absence of an effect of environment, possibly relative validity can be concluded for both the fixed base and hexapod simulators, for both scenarios. Given inconsistent ordering of all three task across all conditions the effect sizes between real world and either simulator setting are in the medium range, absolute

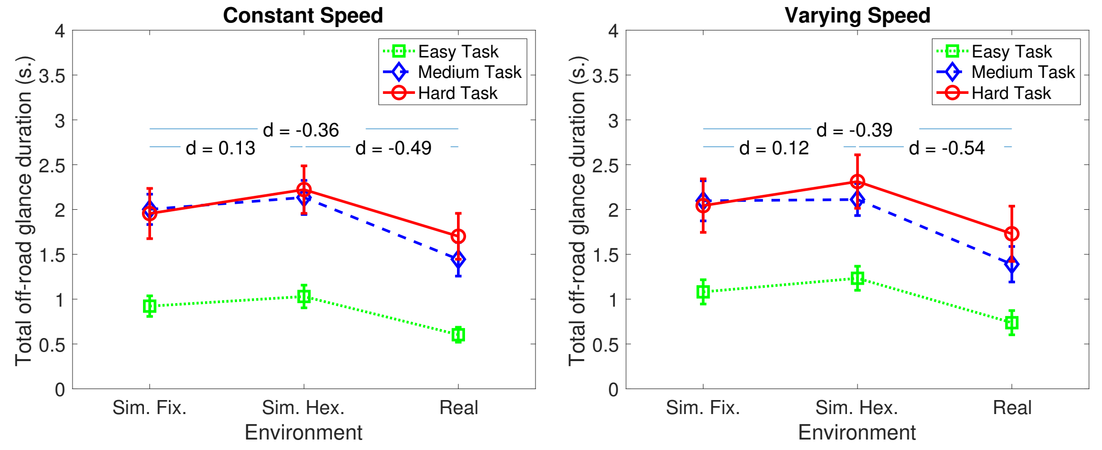


Figure 4.2: Total off-road glance duration.

validity could not be established in this case. This is in agreement with the findings from Victor et al. (2005) where the possibility of relative validity could be concluded.

Frequency of Off-road Glances

Frequency of off-road glances refers to the total number of glances towards the HMI during a task execution. A significant main effect of task was observed ($F(2, 1969) = 38.22$, $p < .001$), which was consistent across environments ($F(4, 1969) = 0.71$, $p = 0.579$). There was no effect of environment or scenario ($F(2, 1969) = 0.09$, $p = 0.918$ and $F(1, 1969) = 0.21$, $p = 0.65$ respectively). Drivers appear to be employing slightly more glances to the HMI when driving in real world settings. 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, the consistent ranking of tasks and the small effect sizes, indicate that absolute behavioural validity can be concluded for both the fixed base and the hexapod simulators, for both scenarios.

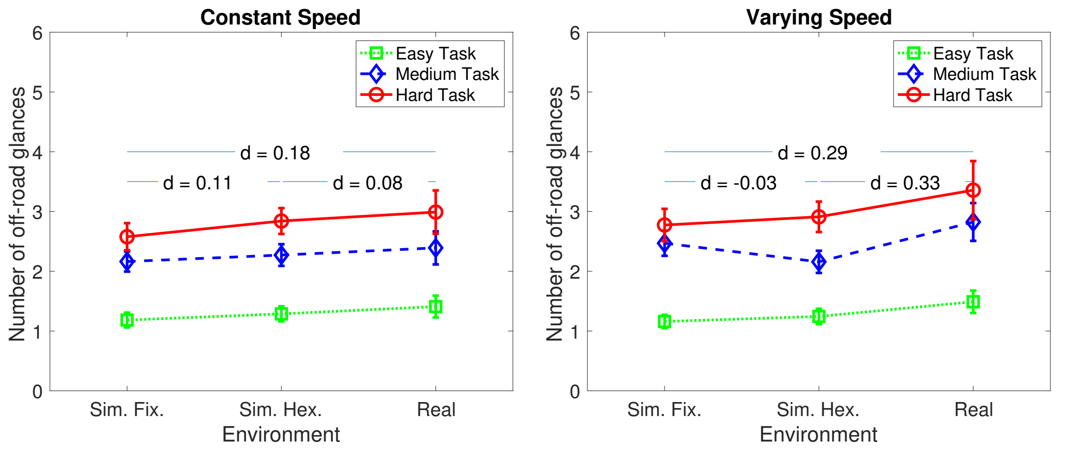


Figure 4.3: Glance frequency.

Mean Off-road Glance Duration

Mean off-road glance duration was defined as the average duration of all glances towards the HMI within a single task execution, or, in other words as the ratio of total off-road glance duration time over the number of off-road glances. After initial model fitting, the model residuals were found to be non-normally distributed. A log transformation was applied to both datasets and models were re-fitted, this time showing no deviation from homoscedasticity or normality for the residuals. A significant main effect of task was observed ($F(2, 1834) = 15.13$, $p < .001$), consistent across environments ($F(4, 1834) = 0.71$, $p = 0.585$), along with a marginally significant effect of environment ($F(2, 1834) = 3.42$, $p = 0.033$), while scenario had no significant effect on mean off-road glance durations ($F(1, 1834) = 0.44$, $p = 0.507$). Given the consistent ranking of the tasks across all conditions, relative validity could be established for both simulator settings, for the constant speed scenario. Due to the inconsistencies in ranking for the varying speed scenario, no level of validity can be concluded. Moreover, due to the effect of environment (although marginal) and the large effect sizes, absolute validity cannot be concluded for either simulator setting and driving scenario.

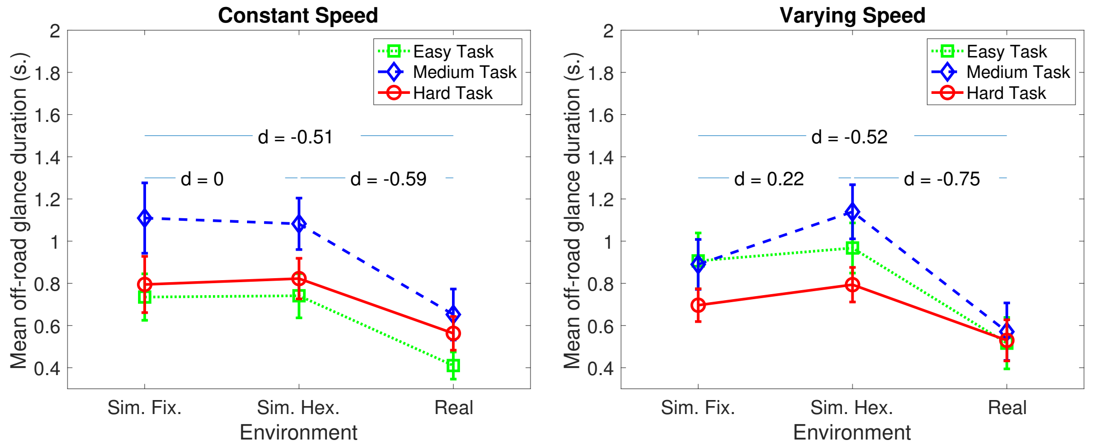


Figure 4.4: Mean off-road glance durations.

4.3.3 Lateral Control

Standard Deviation of Lateral Position (SDLP)

For the sake of consistency, since actual lane position measurements were not available from the real world test track, displacement was calculated using the same method for both real and simulated data (displacement was derived from the recorded lateral acceleration signal). First a simple median filter was applied to the signal for noise reduction and, then, it was integrated twice (using the method of cumulative trapezoidal numerical integration) to yield lateral displacement.

Figure 4.5 illustrates lateral displacement variability across conditions. An immediate observation is that SDLP was higher in the real world, with more variability during the varying speed scenario, something that is not in agreement with previous findings in the literature (e.g. Wang et al., 2010). This observation could, in this case, be attributed to the method used for calculating the metric. In particular, sudden bursts in lateral acceleration could cause increased SDLP values that might not necessarily reflect the actual lateral displacement. Moreover, numerical integration is an approximation method, hence the inferred lateral displacement is also an approximation here. Finally, the way lateral acceleration was calculated by the simulator vehicle dynamics model in this case is not necessarily representative of what was recorded in the real world vehicle, hence results from the integration might not be directly comparable. No significant main effect of task, environment or scenario was found ($F(2, 2007) = 0.74$, $p = 0.478$ for task, $F(2, 2007) = 0.84$, $p = 0.43$, for environment and $F(1, 2007) = 1.95$, $p = 0.162$ for scenario). Moreover, no significant interaction effect between environment and task was found ($F(4, 2007) = 0.16$, $p = 0.956$). Due to the lack of difference between tasks, ranking order cannot be used and, hence, relative validity cannot be concluded. However, exactly due to the negligible differences and small effect sizes observed, a possibility of absolute behavioural validity could be argued in this case.

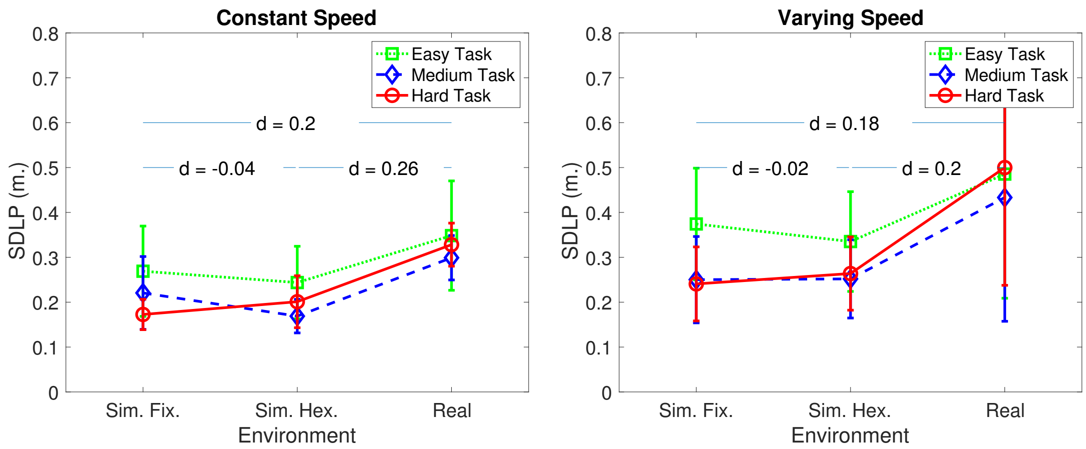


Figure 4.5: Standard deviation of lateral position.

4.3.4 Longitudinal Control

Speed Variability

Speed variability was calculated as the standard deviation of longitudinal velocity during a task execution being significantly higher in the simulator conditions. No significant main effect of task was observed ($F(2, 2007) = 0.2$, $p = 0.821$), consistently across environments ($F(4, 2007) = 0.15$, $p = 0.961$). Environment and scenario, on the other hand were found to significantly affect speed variability ($F(2, 2007) = 18.61$, $p < .001$ and $F(1, 2007) = 12.06$, $p < .001$, respectively). Since there is no effect of task, ordering could not be considered and, consequently, neither absolute nor relative behavioural validity can be concluded for either of the simulator types and driving scenarios.

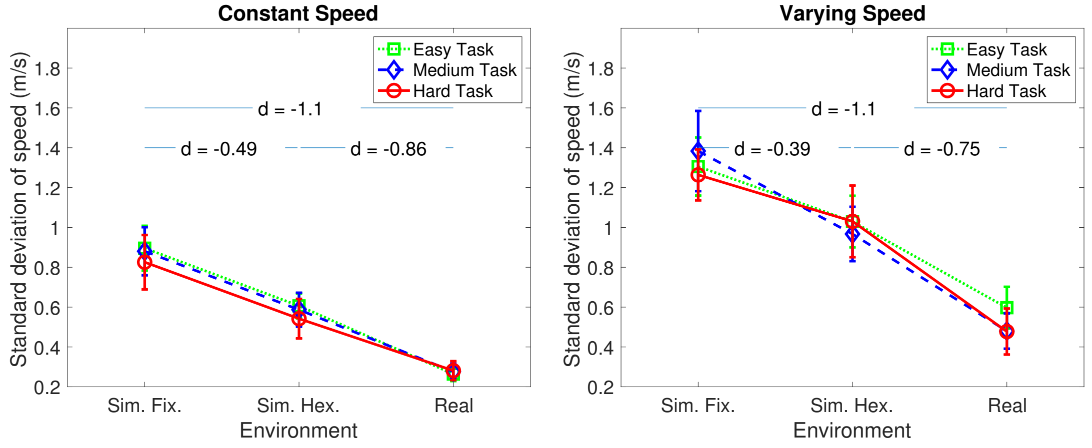


Figure 4.6: Speed variability.

Average Speed

Average speed was calculated as the average value of longitudinal velocity during a task execution (see Figure 4.7). A significant effect of scenario was observed ($F(1, 2007) = 440.39$, $p < .001$), as was expected due to the differences in speed, while task and environment and their interaction appeared to have no effect on drivers' speeding behaviour ($F(2, 2007) = 0.48$, $p = 0.62$ and $F(2, 2007) = 0.58$, $p = 0.559$ and $F(4, 2007) = 0.31$, $p = 0.87$, respectively). Since there is no effect of task, ordering could not be considered, hence relative validity cannot be concluded here. However, due to the absence of an effect of environment and the low effect sizes observed, a possibility for absolute validity could be argued in this case.

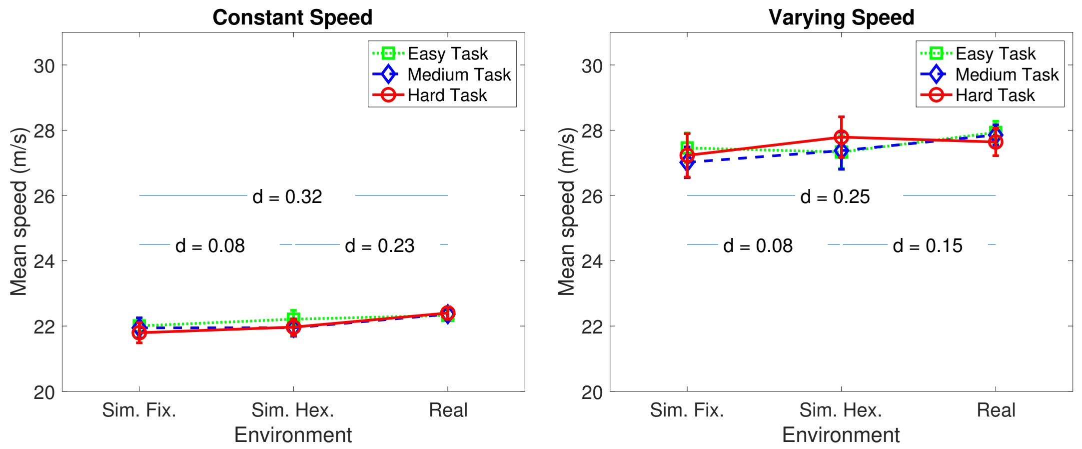


Figure 4.7: Average speed.

4.3.5 Steering Control

Following, the Steering Wheel Reversal Rate (SWRR) for the HMI tasks are illustrated. The reversal rates were calculated using the method described in Markkula and Engström (2006). The rates were calculated for gap sizes of 1, 5 and 10 degrees. The results and conclusions for all three gap sizes were similar, hence detailed statistical analysis results are only reported for the 1 degree gap size, while visualisation of all gap sizes is provided (see Figures 4.8, 4.9 and 4.10). None of the main factors in question appeared to significantly affect steering behaviour ($F(2, 2007) = 1.12$, $p = 0.326$ for task, $F(2, 2007) = 0.94$, $p = 0.392$, for environment and $F(1, 2007) = 0.05$, $p = 0.831$ for scenario). Moreover, the effect of task did not appear to vary in different environments ($F(4, 2007) = 0.26$, $p = 0.906$). Task ranking appears to be somewhat consistent throughout all conditions for gap sizes of 1 and 5 degrees. Due to the lack of an effect of task, however, ordering cannot be considered and relative validity cannot be established. Again, though, a case of possible absolute validity could be argued, given that the effect sizes observed are relatively small.

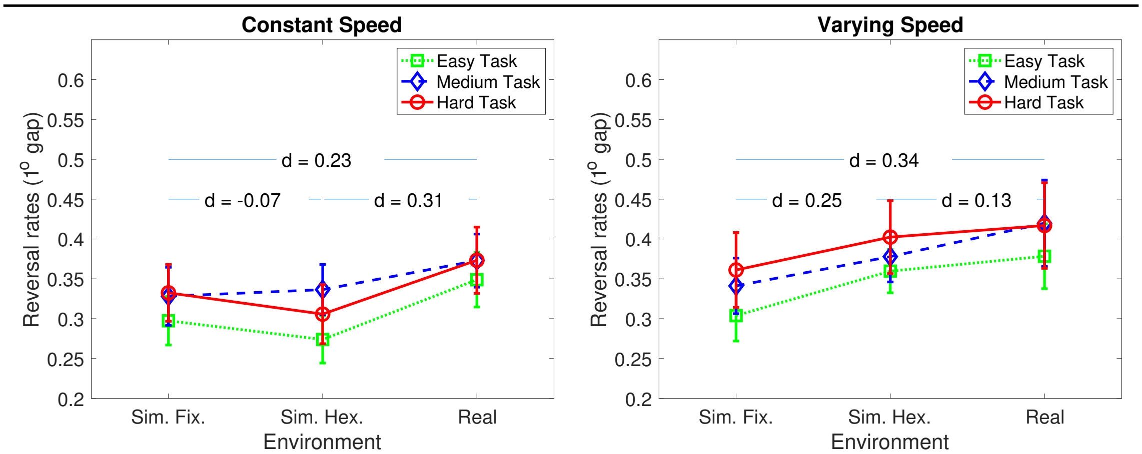


Figure 4.8: Effects and interactions of SWRRs for gap size of 1^o.

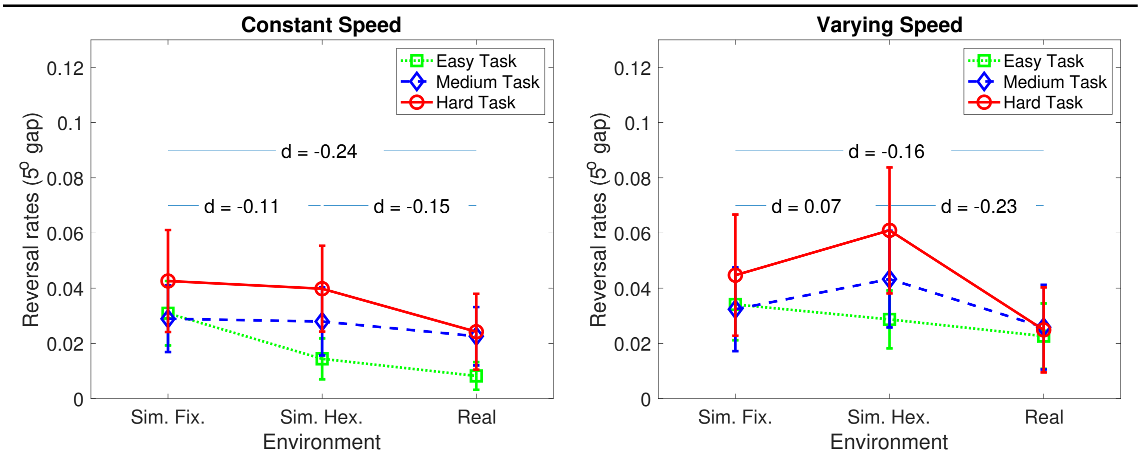


Figure 4.9: Effects and interactions of SWRRs for gap size of 5^o.

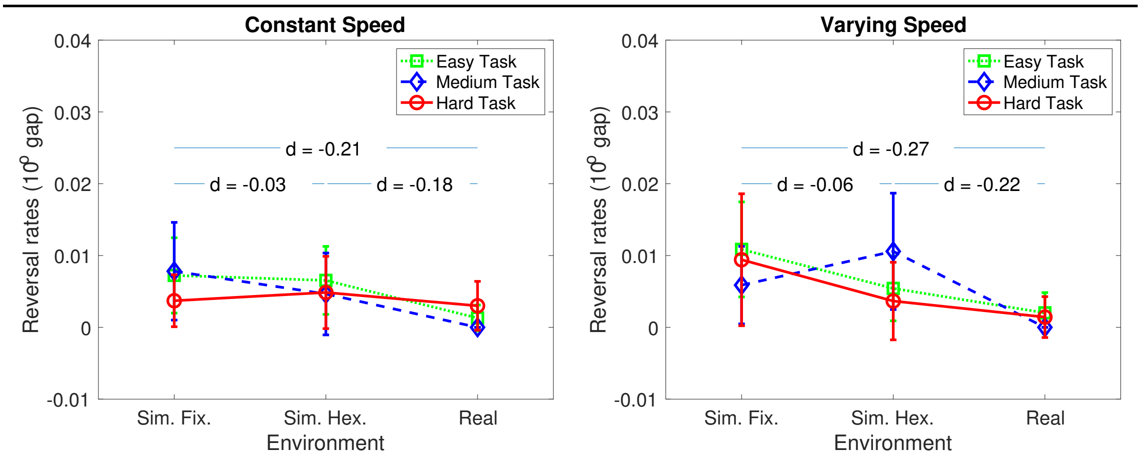


Figure 4.10: Effects and interactions of SWRRs for gap size of 10^o.

4.4 The Behavioural Validity Matrix

A Behavioural Validity Matrix was constructed based on the results obtained from the present analysis as well as on existing results in the literature, as previously noted. The matrix was filled using a simple colour-coding scheme that is illustrated in the top part of Figure 4.2 and is associated with the level of behavioural validity each setting seems to be able to achieve. "Possibly" Absolute or Relative validity refers to either an inconclusive result in literature or to an interpolation of existing results towards the higher level of validity observed. For example, in a case where for the same simulator setting, investigating the same metric, there were two different verdicts for behavioural validity, relative and absolute, the final verdict here would be possibly absolute validity. Moreover, every simulator setting that had a lower fidelity simulator ranked with a higher level of validity for the same metric, was also considered to possibly achieve that level. For example, if for a certain metric, the desktop simulator was ranked with absolute validity, and the hexapod and lateral motion simulator was ranked with relative validity or had no ranking, the latter would be finally ranked as possibly absolute. The red rectangle in the matrix indicates the contributions of the work conducted in the scope of this thesis. In particular, the behavioural validity rankings for the *Cabin with wide field projection* are presented in combination with existing results from the literature. The behavioural validity rankings for the *Hexapod*, on the other hand, are a novel contribution of this thesis.

The behavioural validity matrix in its current state could prove to be a useful tool towards answering the question of what type of simulator one needs to conduct reliable HMI evaluation tests. It is important to note at this point, however, that the behavioural matrix presented here was formed using results from a variety of studies where different experimental choices were made. Since different HMI tasks and different driving scenarios could potentially yield different levels of behavioural validity for the same metric, tools like this behavioural matrix are in need of updates and revisiting.

When it comes to choosing a driving simulator setting for prototype HMI

Legend
Absolute
Possibly Absolute
Relative
Possibly Relative
N/A

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	Relative	Possibly Absolute	Possibly Absolute	Absolute	Absolute	Absolute	Absolute
Gaze behaviour	Total off-road glance duration	Relative	Possibly Absolute	Absolute	Possibly Absolute	Possibly Absolute	Possibly Absolute	Possibly Absolute
	Mean off-road glance duration	N/A	Possibly Relative	Possibly Relative	Possibly Relative	Possibly Relative	Possibly Relative	Possibly Relative
	Off-road glance frequency	Possibly Relative	Possibly Relative	Relative	Possibly Absolute	Absolute	Possibly Absolute	Possibly Absolute
Lateral Control	SDLP	N/A	Relative	Relative	Relative	Possibly Absolute	Absolute	Absolute
Longitudinal Control	Speed - StD.	N/A	Possibly Relative	Possibly Relative	Relative	Possibly Relative	Relative	Relative
	Speed - Mean	N/A	Possibly Relative	Relative	Relative	Possibly Absolute	Possibly Absolute	Possibly Absolute
Steering Control	SWRR	N/A	Possibly Relative	Possibly Relative	Relative	Possibly Absolute	Possibly Absolute	Possibly Absolute

Table 4.2: The Behavioural Validity Matrix, based on analysis of the obtained data and existing results from literature.

evaluation, as is evident from the matrix, there is no “one-size-fits-all” solution to that problem; instead, it is highly dependent on the types of tests that require consideration (i.e. evaluated metrics), as well as on the level of behavioural validity that needs to be achieved. Different metrics require different simulator types to be evaluated with higher behavioural validity. Only simulators with motion cues can achieve near absolute or absolute validity for lateral and longitudinal control measures (namely SDLP and mean speed). This is easily explained by the fact that humans cannot comprehend movement as well without vestibular cues. Task completion times, on the other hand, can be examined in very low fidelity simulator settings, with informative results, approximating the corresponding real-world behaviour very well.

Finally, researchers and human factors specialists could also find the behavioural validity matrix useful in interpreting the results of a driving simulator study. When no real-world data collection can be conducted, for example, the matrix can be used to help researchers infer how well the observed results in the examined simulator setting would generalise to the real-world context.

Chapter 5

Visual Attention Sharing Patterns During HMI Task Execution

The present chapter investigates which factors affect drivers' visual attention allocation, when engaging in visual-manual HMI tasks, and in which way. Such factors could point towards behavioural patterns that can be utilised for the specification and implementation of computational models, capable of replicating the observed behaviour. Drawing from existing techniques in the literature, as well as through some novel pieces of analysis, the glance behaviour of the drivers was analysed on the individual glance level, and is presented in the following sections.

Initially, a general exploratory analysis of the glance duration data was conducted, aiming at comprehending their distribution and understanding which factors might affect glance durations across different conditions. Next, the structure of drivers' visual attention sharing was investigated, focusing on the relationship between off-road glance frequency and glance duration. Finally, the effect of the primary driving task demand on glance duration and glance onset timing was investigated, with regards to lateral and longitudinal vehicle control.

5.1 Exploratory Analysis

A crucial part of exploratory data analysis is to visually inspect the data for preliminary pattern identification (Tukey, 1977). With glance durations, the typical assumption is that a log-normal distribution fits the data well (e.g. Morando et al., 2019). In this case, since the aim was to visualise and inspect the raw data, a non-parametric representation of their probability density function (PDF) was used. A smoothing function was applied to the data to obtain a kernel density estimation of the probability distribution (Parzen, 1962; Peter D, 1985). Similar to a histogram, the kernel density estimation method creates a function to represent the probability distribution of the data in question. But unlike a histogram, which places the values into discrete bins, a kernel distribution sums the component smoothing functions for each data value to produce a smooth, continuous probability curve.

After an initial inspection, glance durations did not differ significantly between simulator types (fixed base and hexapod) and driving scenarios (constant and varying speed). Hence, the data for both the simulator and the real word was collapsed across the two scenarios. Moreover, the simulator data was also collapsed accordingly along the two simulator types. Consequently, the data is presented here categorised by HMI task (easy, medium, hard) and road type (straight, curve) for the simulator and by HMI task only for the real world.

For each task, a kernel density estimate was computed for each driver and for the aggregated data of all the drivers for that condition, based on a normal kernel function, evaluated at 1000 equally spaced points. Figure 5.1 illustrates on- and off-road glance duration distributions for each HMI task in the different conditions defined above, along with their respective medians.

In agreement with existing research (e.g. Horrey and Wickens, 2007; Birrell and Fowkes, 2014), glance durations appear to be right skewed (i.e. having more data points concentrated in the low values). Apart from the shape of the distributions, another preliminary observation can be directly made from a simple visual inspection of the plots: different drivers appear to employ “personalised”

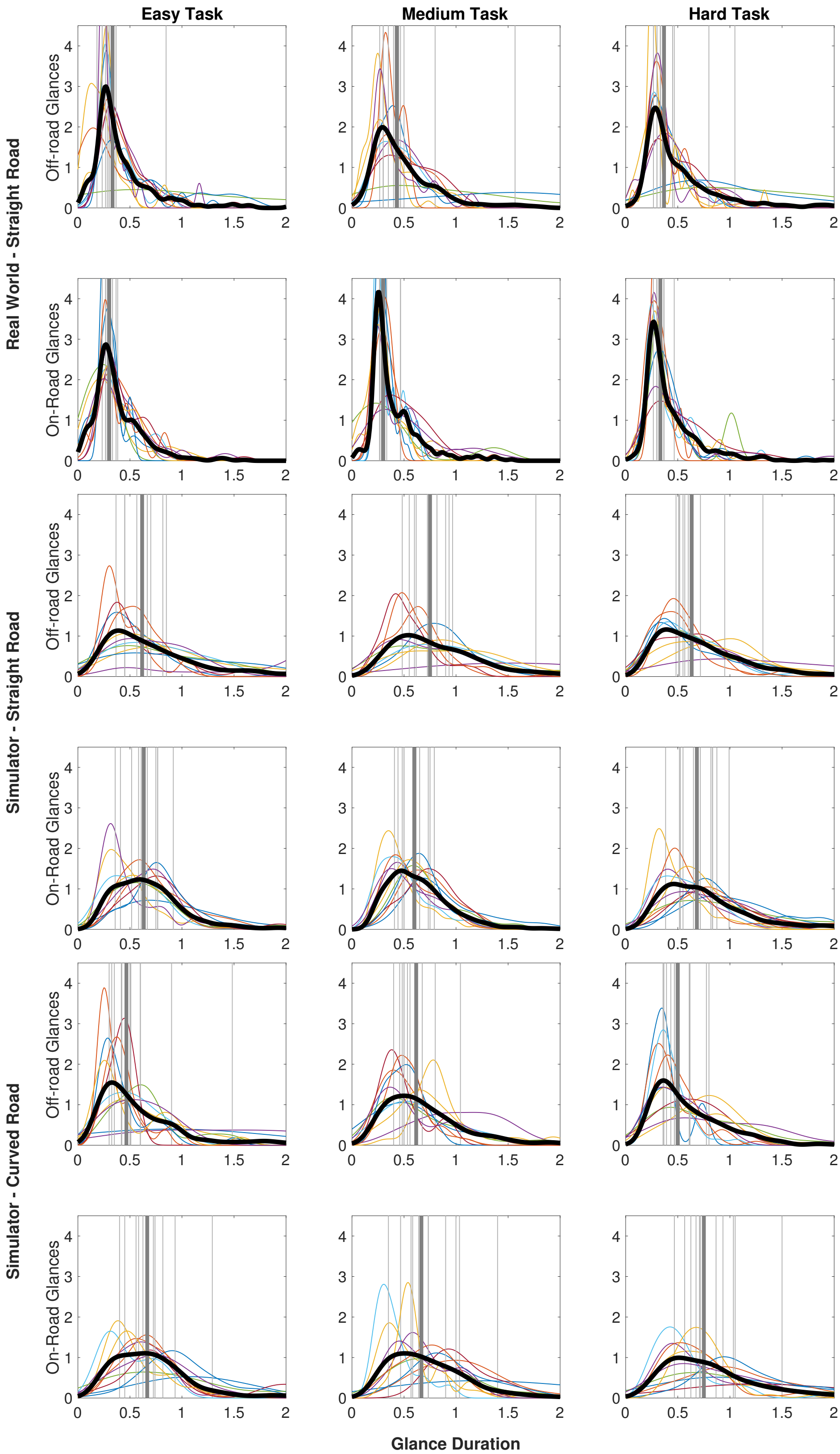


Figure 5.1: Glance duration distributions during HMI task execution for real world and simulated driving. In each panel, the coloured lines correspond to individual driver distributions, while the thick black line corresponds to the overall aggregate data distribution. The thin, light grey, vertical lines correspond to individual drivers' median glance durations, while the thick, light grey, vertical line corresponds to the aggregate data median glance duration.

strategies as to how they share their visual attention between the road and the HMI. In particular, different drivers appear to employ glances of different durations when looking towards the HMI. The overall distributions (thick black curves) in the various conditions do not seem to always capture individual driver distributions (coloured curves), something that becomes visually evident by the misalignment of the curves, as well as by the distance between the glance duration medians (light grey vertical lines). Initially, a singe Kruskal-Wallis test (Kruskal and Wallis, 1952) was performed on the data and verified that individual driver glance durations do not come from the same distribution (results for all tasks and conditions rejected the null hypothesis in the $a = 0.05$ confidence interval, with $p < .001$). To eliminate the increased risk for type I error associated with multiple comparisons, additional pair-wise Kruskal-Wallis tests were performed for each driver and the corresponding overall glance distribution in each condition, yielding the same results. Moreover, visual differences in the individual observed distributions between different conditions, also indicate some within-driver variability for different HMI tasks and road type.

In order to get a more detailed view on drivers' visual attention sharing behaviour, some further quantitative analysis was also performed. Consistent with the analysis presented in Chapter 4, linear mixed effects models were used to analyse individual glance durations here, too. Statistical analysis was performed on the two datasets independently, to investigate the effects of glance target (road, HMI - for both datasets) and road (straight, curve - only for the simulator data) on glance duration.

The rationale for not performing the statistical analysis on the aggregate dataset, was to avoid real world data affecting any significance of the curved road condition in glance durations (since curves could not be tested in real world). Moreover, the detailed analysis that follows in the next sections of the present chapter, focuses on simulator data only, hence it would be more meaningful to make that distinction here, too.

Following the recommendation by Barr et al. (2013), both models included a random effects component that corresponds to the maximal random structure as that can be justified by the design. For the simulator data, main effects and all their two-way interactions were investigated for glance target (road ahead, HMI), HMI task (easy, medium, hard) and road type (straight, curve). The model included a random intercept and random slopes for the three factors varying per participant. For the real world data, main effects and all their two-way interactions were investigated for glance target (road ahead, HMI) and HMI task (easy, medium, hard). The model included a random intercept and random slopes for the two factors, varying per participant. After initial model fitting, the residuals for both models were visually investigated using QQ (quantile-quantile) plots (Wilk and Gnanadesikan, 1968). Both generated non-normally distributed residuals. A log transformation was applied to both datasets and models were refitted, this time showing no deviation from homoscedasticity or normality for the residuals.

In the simulator, a significant interaction effect between road and glance target on glance duration was found ($F(2, 8841) = 4.07$, $p = 0.017$). In particular, glances towards the road were significantly longer when driving in curves ($F(1, 8841) = 53.42$, $p < 0.001$), with drivers spending 0.82 seconds on average looking at the road ahead and 0.65 seconds looking at the HMI. This could be explained by the increased workload and primary driving task demand associated with driving in a curve. It is worth noting at this point that participants were, in general, a lot more conservative in their HMI engagement tactics when driving in curves in the simulator, often postponing performing the task until they felt safe to do so. This observation seems to be falling in line with the findings by Oviedo-Trespalacios et al. (2017), where they concluded that in complex driving environments, the primary driving task gets prioritised over the secondary task. One participant, in particular, did not perform any HMI task executions while driving on a curve, since, as they pointed out after the experiment, they did not feel it was safe.

In the real world setting, drivers' individual glances towards the road ahead were shorter in duration than the ones towards the HMI, showing a main effect of glance target on glance duration ($F(1, 2696) = 5.25$, $p = 0.021$ - see top two rows of Figure 5.1). In particular, drivers spent 0.41 seconds on average looking on the road ahead and 0.52 seconds looking at the HMI. This difference could be attributed to the fact that drivers had a better control of the vehicle in the real world, due to more acute vestibular and tactile feedback, hence being able to "afford" to spend more time looking away from the road.

Considering the above, it can be argued that both the HMI task and the primary driving task demand appear to affect the way in which drivers share their attention between the road and a secondary task while driving. Moreover, different drivers tend to employ individual strategies of visual time sharing, that result in glances of different durations towards the road and towards the HMI.

5.2 Visual Attention Sharing Structure

After identifying that different drivers employ off-road glances of different durations, it is logical to investigate how, in detail, drivers structure their visual attention sharing between the road and the HMI and if similar differences are observed there, too. The present section investigates this very question, along with if and how such differences in structure might be associated with the differences observed in glance durations.

Since all three HMI tasks that were used in this experiment consist of discrete sub-tasks (the individual interactions that drivers needed to complete), a simple first step towards understanding drivers' visual attention sharing structure would be to verify if that matched the structure of the HMI tasks they are performing. The simplest form of such structural match would be to allocate one single glance for every single interaction required. For example, in the present experiment, a driver employing such a strategy would need three glances to execute the easy task (three Press interactions), four glances to execute the medium task (four Press interactions) and at least five glances to execute the hard task (four Press and one Scroll interaction).

Figure 5.2 illustrates the number of off-road glances employed per HMI task execution, in each condition, for all drivers. The "percentage of executions"

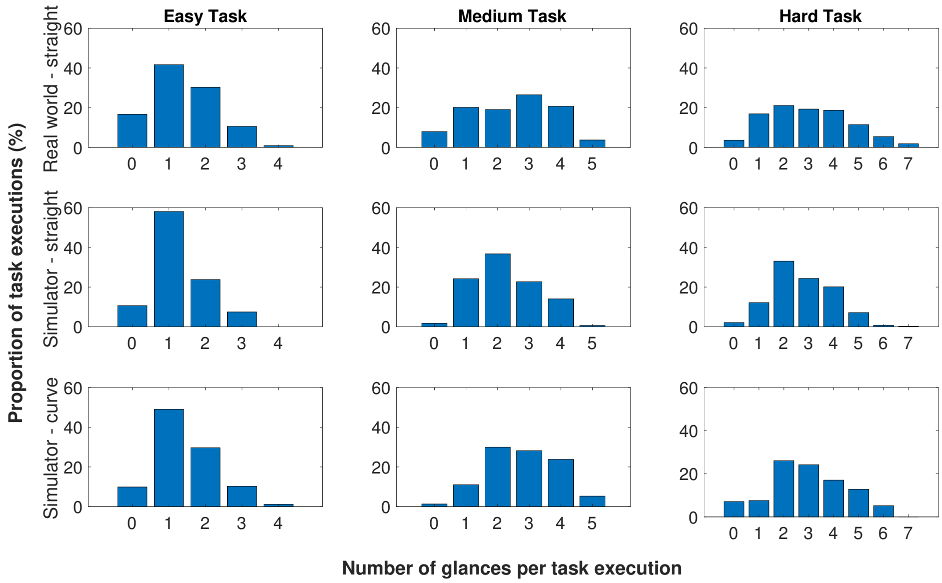


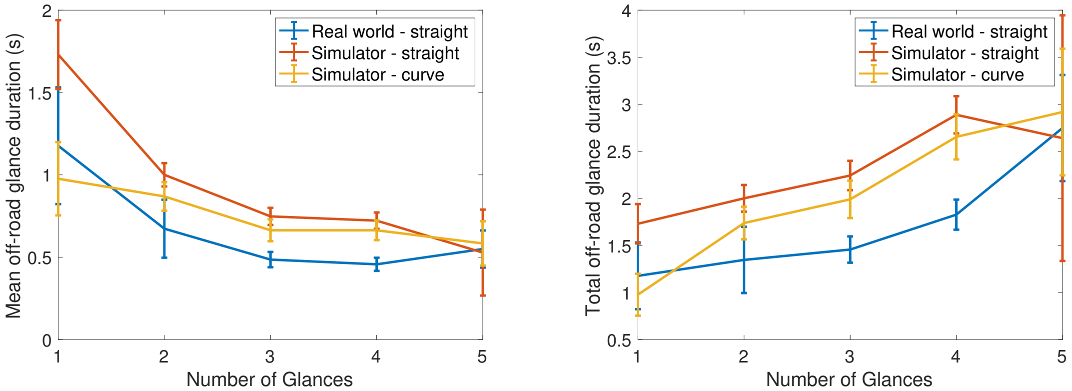
Figure 5.2: Distribution of number of glances per task execution in real world and simulated driving.

breakdown corresponds to the normalised number of executions of each particular task for each particular condition (the number of executions for each task in each condition was normalised over the total number of task executions in that condition). It is immediately evident that the above assumption regarding visual attention structure is not in agreement with what drivers actually did. In particular, more than 70% of task executions (for all HMI tasks, across all conditions), were performed with the drivers employing fewer glances than the individual required interactions. This chunking of the HMI task is driven by drivers performing multiple interactions in a single glance, or performing some of the interactions without looking, by mapping the relative position of the next target on the screen. Salvucci et al. (2005*a*), for instance, also noticed that drivers grouped button presses together when typing a phone number while driving, by either looking longer towards the keypad or typing by tactile detection of the appropriate buttons. Another interesting observation is that some task executions were performed without the drivers looking towards the HMI at all. After inspection of the video footage this could be attributed to some drivers being able to perceive the position of the initial target on the HMI using their peripheral vision and mapping the relative positions of subsequent targets.

To better understand the implications of this phenomenon, it is necessary to investigate chunking in the context of glance durations. The relationship between number of off-road glances and glance duration has been previously discussed in connection to drivers' risk taking behaviour. Donmez et al. (2009) found that riskier drivers tended to use few and longer glances, while more conservative, safe drivers tended to opt for more in number and shorter in durations glances away from the road. Here, a detailed quantitative analysis of this relationship was conducted, through statistical analysis of the observed data. Using linear mixed effects models, the effect of number of off-road glances on mean off-road glance duration and total off-road glance duration was investigated. Similarly to the analysis in the previous section, two models were fitted, one for the real world data and one for the simulator data. After initial model fitting, both models generated non-normally distributed residuals. A log transformation was applied to both datasets and models were re-fitted, this time showing no deviation from homoscedasticity or normality for the residuals. It is worth noting at this point that although there were changes in the absolute values of the statistical metrics, no shift in statistical significance levels was observed.

Figure 5.3 illustrates the investigated relationship for the medium HMI task across all conditions. The respective plots for the easy and hard HMI tasks were similar and, thus, not presented here. For the medium task, more than 90% of task executions across all conditions employed 1 to 5 glances, hence those are presented in Figure 5.3.

For both the real world and the simulator, an inversely proportional relationship between number of off-road glances and mean off-road glance duration was observed, i.e. fewer number of glances were associated with larger mean glance duration and vice versa (see Figure 5.3(a)). For the real world, this relationship was not consistent, as number of glances did not have a significant effect on mean glance duration ($F(1, 518) = 0.78$, $p = 0.377$). For the driving simulator, on the other hand, the number of glances towards the HMI significantly affected the mean glance duration during an HMI task execution ($F(1, 1973) = 14.09$, $p < 0.001$).



(a) (b)

Figure 5.3: Number of off-road glances

Considering that each HMI interaction requires some time, performing multiple interactions in a single glance could result in an increase in total glance duration. Indeed, a consistently proportional relationship between number of off-road glances and total off-road glance duration was observed for both the real world and the simulator, with more glances resulting in larger durations ($F(1, 518) = 26.5$, $p < .001$ and $F(1, 1973) = 36.09$, $p < .001$ respectively - see Figure 5.3(b)).

Table 5.1 collectively presents all the above reported results. For each task, in each condition, the Table provides a grouping of the number of off-road glances used per HMI task execution, with their associated mean off-road glance and total off-road glance durations.

Two distinct strategies can be observed on how drivers engage with HMI tasks: they either use fewer and longer glances or more and shorter glances. Moreover, the number of glances employed in an HMI task execution, also proportionally affects the total time drivers look towards the interface. Studying Figure 5.2 and Table 5.1 reveals that the number of glances scales with the number of interactions needed to complete a task, but, typically, the number of glances is lower than the number of interactions. As a result, it could be argued that drivers' tendency to use fewer glances could be associated with them, eventually, trying to spend less time looking away from the road.

Easy Task									
Number of glances	Real world - straight			Simulator - straight			Simulator - curve		
	Proportion of executions (%)	Mean glance duration	Total glance duration	Proportion of executions (%)	Mean glance duration	Total glance duration	Proportion of executions (%)	Mean glance duration	Total glance duration
0	16.67	0.00	0.00	10.66	0	0	9.89	0	0
1	41.67	0.65	0.65	58.08	1.10	1.10	49.05	0.85	0.85
2	30.26	0.44	0.89	23.80	0.62	1.24	29.66	0.54	1.09
3	10.53	0.41	1.24	7.46	0.55	1.66	10.27	0.47	1.40
4	0.88	0.52	2.06	0	0	0	1.14	0.43	1.72

Medium Task									
Number of glances	Real world - straight			Simulator - straight			Simulator - curve		
	Proportion of executions (%)	Mean glance duration	Total glance duration	Proportion of executions (%)	Mean glance duration	Total glance duration	Proportion of executions (%)	Mean glance duration	Total glance duration
0	7.94	0.00	0.00	1.73	0	0	1.32	0	0
1	20.11	1.18	1.18	24.19	1.73	1.73	11.01	0.98	0.98
2	19.05	0.67	1.35	36.72	1.00	2.00	29.96	0.87	1.74
3	26.46	0.49	1.46	22.68	0.75	2.24	28.19	0.66	1.99
4	20.63	0.46	1.83	14.04	0.72	2.89	23.79	0.66	2.65
5	3.70	0.55	2.75	0.65	0.53	2.64	5.29	0.58	2.92

Hard task									
Number of glances	Real world - straight			Simulator - straight			Simulator - curve		
	Proportion of executions (%)	Mean glance duration	Total glance duration	Proportion of executions (%)	Mean glance duration	Total glance duration	Proportion of executions (%)	Mean glance duration	Total glance duration
0	3.61	0.00	0.00	2.12	0	0	7.11	0	0
1	16.87	0.37	0.37	12.17	0.74	0.74	7.58	0.61	0.61
2	21.08	0.80	1.61	33.07	0.90	1.81	26.07	0.69	1.38
3	19.28	0.63	1.89	24.34	0.80	2.39	24.17	0.63	1.89
4	18.67	0.51	2.05	20.11	0.70	2.78	17.06	0.55	2.20
5	11.45	0.46	2.30	7.14	0.69	3.46	12.80	0.66	3.30
6	5.42	0.53	3.16	0.79	0.76	4.53	5.21	0.64	3.86
7	1.81	0.52	3.63	0.26	0.75	5.25	0	0	0

Table 5.1: Visual attention sharing structure during HMI task execution details. For each HMI task, in each condition, the number of glances towards the HMI is provided as a proportion of the total number of task executions in that condition, along with the corresponding mean and total off-road glance duration.

5.3 Effects of Primary Driving Task Demand on Visual Attention Sharing

Returning to the exploratory analysis presented earlier in this chapter, it was shown that the demand of the primary driving task affects glance durations, as manifested by differences between simulator and reality and between straight and curved road within the simulator. The analysis presented in this section aims to define whether drivers condition their off-road glances (both duration- and onset-wise) on safety perception and primary driving task demand, related to lateral and longitudinal control. The analysis hereafter was only performed on the simulator (UoLDS experiment) data, since the relevant metrics explored were not recorded or could not be computed for the real world.

5.3.1 Glance Durations and Time-to-Line Crossing

Time-to-line crossing (TLC) is defined as the time needed for any part of the vehicle to reach one of the lane boundaries (Godthelp et al., 1984). TLC has been previously used to identify and model steering behaviour (Godthelp et al., 1984) as well as to identify cognitive distraction (Li et al., 2018), among others. The relationship between TLC and off-road glance duration, as well as the onset of off-road glances in an HMI evaluation context, has not been previously explored.

A hypothesis is proposed here that suggests drivers might take into account TLC values (or their perception of it) before engaging in an off-road glance and deciding for how long to do so. The hypothesis is qualitatively illustrated in Figure 5.4. Let TLC_{off} be the TLC at the moment a driver looks away from the road (and in this case towards the HMI) and Off_{max} the maximum time a driver can look away before exiting the lane. Assuming constant steering when looking away (as suggested for example by Godthelp et al., 1984), it should hold that $Off_{max} = TLC_{off}$ (represented by the diagonal line in Figure 5.4). As discussed previously in Chapter 2, drivers have been found to self-regulate when being distracted by a secondary task (e.g. by adapting their speed), to remain within safe driving margins. Consequently, a minimum "safety" threshold is also assumed in this case for TLC, i.e. a value of TLC, below which drivers do not take their eyes off the road (represented by the vertical line in Figure 5.4). Finally, off-road glance durations are expected to be bound by an upper limit threshold, regardless of the TLC_{off} value. This can be justified intuitively, by the fact that drivers cannot indefinitely look away from the road as they would eventually need to return their gaze to adjust for errors and control the vehicle. It has, however, also been previously argued that drivers limit the duration of their off-road glances to maintain safety. Wierwille (1993), for example, argued that drivers generally employ glances to an HMI that are one second or less in duration and not longer than 1.5 seconds. Following, this hypothesis, all observed data should fall within the light cyan area, as indicated in Figure 5.4. It is worth

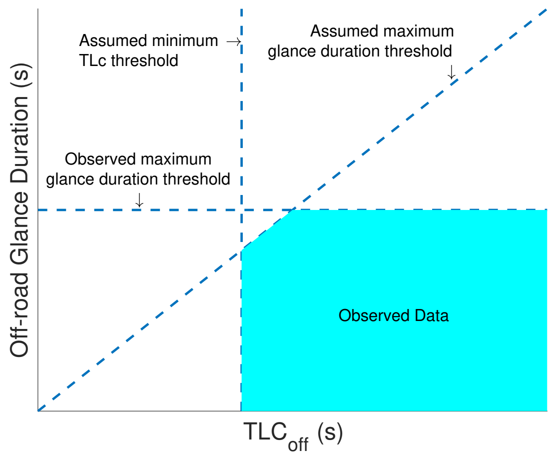


Figure 5.4: A hypothesis of how TLC data would compare against off-road glance durations. A maximum threshold of off-road glance duration is assumed that equates to the TLC at the time of looking away. Glance durations are bound above, and a minimum TLC threshold is assumed to initiate an off-road glance.

noting at this point, that the diagonal line representing the maximum look away time does not need to cross the $(0, 0)$ point, as such an instance would not be observed in the recorded data. It is illustrated as such in the following plots only for demonstration purposes, mainly to provide an indication that it is the $x = y$ diagonal line.

In order to test the above hypothesis, for each off-road glance, TLC_{off} was calculated using the following formulas (Godthelp and Konings, 1981; Mammar et al., 2006; of Automotive Engineers, 2015):

$$\frac{1}{2} \cdot LA \cdot TLC_{left}^2 + LV \cdot TLC_{left} = LP_{left} \tag{5.1}$$

$$\frac{1}{2} \cdot LA \cdot TLC_{right}^2 + LV \cdot TLC_{right} = LP_{right} \tag{5.2}$$

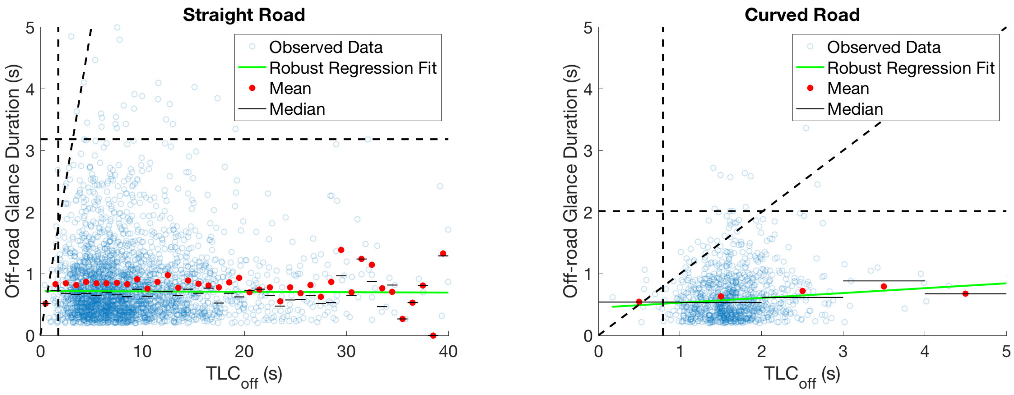
where:

$TLC_{left/right}$ stands for TLC to the left or right lane boundary, respectively,

LA stands for lateral acceleration,

LV stands for lateral velocity and

$LP_{left/right}$ stands for the lateral distance from the left or right lane boundary, respectively.



(a) (b)

Figure 5.5: TLC_{off} against subsequent off-road glance duration, for straight and curved road. Data are binned by TLC_{off} value, with the edges of each bin shown by a horizontal black line through the bin's median glance duration value and the bin's mean glance duration value shown in red. A robust regression fit of the data is shown in green.

The minimum positive root of the two equations is used as TLC_{off}. The above is an approximate method, assuming constant lateral acceleration. It was found to reliably replicate results of more complex analytical methods and be sensitive to cognitive distraction (Li et al., 2018).

Scatter plots of TLC_{off} against subsequent off-road glance duration are illustrated in Figure 5.5. They display aggregated data for the UoLDS experiment, separated by road type only (i.e. for each road type, all simulator type, task and scenario data are collapsed). TLC_{off} was also plotted as a function of individual drivers, task, driving scenario and simulator type, and, although demonstrating some variability, all plots looked effectively the same as those in Figure 5.5.

Regarding the previously formulated hypothesis (Figure 5.4), an initial look at Figure 5.5 seems to validate it. The horizontal line representing the maximum glance duration value observed, was drawn for the 99^{th} percentile and fell at a value of 3.18 s for the straight road and 2 s for curved road. Regarding the minimum TLC_{off} line (vertical), it was drawn at the 1^{st} percentile value and fell on 1.75 s for straight road and 0.79 s for curved road. For the straight road condition, 0.18% of the glances exceeded the hypothesised upper limit (diagonal limit), while for the curved road that proportion was at 3.8%. Since the focus of this analysis was to investigate the relationship between TLC_{off} and off-road glance durations, and given the small portion of data violating this hypothesis, no further analysis was conducted on those glances. An interesting first observation is that the minimum TLC_{off} was shorter when driving on curve than on straight road (see Figure 5.5(b)). This is an artefact of the road geometry, since TLC values are in general shorter in curves, as verified by the plot. Moreover, it is interesting to notice that the upper bound for glance duration is a lot shorter in the curved road, something that should be expected based on the exploratory analysis results presented at the beginning of this chapter, as well on the overall known tendency of the drivers to adapt their visual attention based on the demands of the primary driving task.

Investigating the relationship between TLC_{off} and subsequent off-road glance duration, was the next step taken in this analysis. As can be inferred by visually inspecting the scatter plots, no linear or semi-linear relationship appears to exist between the two measures. Nonetheless, the relationship between off-road glance duration and TLC_{off} was evaluated using robust regression (Rousseeuw and Leroy, 2005) to verify it. The regression model failed to capture a relationship between TLC and glance duration for either the straight or the curved road ($R^2 < .001$ and $R^2 = 0.02$ respectively).

As no direct relationship was found through regression analysis, a different assumption to be tested here would be that instead of using specific values of TLC_{off} to predict off-road glance durations, a range of TLC_{off} values could be used to predict a range of glance durations. In order to test this hypothesis, the data was binned as shown in Figure 5.5. Initially, a log-normal distribution was fit to the collective glance durations for each road type. Next, for each TLC_{off} bin, a different log-normal distribution was fit to the glance durations that fell within that bin.

Two simple probabilistic models were considered to quantitatively verify the assumption. Both models used a TLC_{off} value as an input and generated an off-road glance duration, drawn from a corresponding distribution. The first model (simple model) drew glance durations from the overall distribution fitted to the aggregate data. The second model (complex model) drew glance durations from the fitted distribution that corresponded to the TLC_{off} value provided as input. For good measure, a variety of bin approaches were used (varying bin number and bin size). The likelihood of each model was then calculated, to obtain the Akaike Information Criterion (AIC) (Akaike, 1974), which was then used as a performance comparison metric between the models.

Generally, models that have a lower AIC score, provide a better representation of the observed data. The simple model achieved an AIC score of 493 for the curved road and 3272 for straight road. From all the complex models tested (by varying the binning approach), the best scoring variants achieved 500 and 3285, respectively. This means that the best performing complex models performed marginally worse than the simple ones and, thus, offer no additional value to modelling visual attention allocation based on TLC_{off}. From all the above, it is conclusively proven that TLC values at the moment of looking away from the road cannot be used to predict subsequent off-road glance durations.

Revisiting the initial hypothesis and the observed results in Figures 5.5(a),(b), the next logical question to pose is whether there is a connection between TLC value and off-road glance onset, i.e. whether TLC plays any role in when drivers choose to look towards the HMI, rather than for how long. Drivers are known to adapt their behaviour to the demands of the primary driving task. The absence of glance duration data below certain TLC values (0.42 and 0.18 seconds for straight and curved road, respectively) denotes either a margin that drivers use to decide when it is safe to look away, or that drivers never actually attain such low TLC values when driving. To verify which assumption is true, TLC was investigated on the moment of look-away and during baseline driving (when no HMI task was performed). As discussed in Chapter 2, control errors build up during periods of no visual input from the road scene ahead until drivers feel they need to focus their attention back on the road and correct those errors. It could be hypothesised that this error build-up continues during intermittent periods of looking back on the road. Particularly, the assumption here is that drivers look back on the road using short, check glances inbetween interactions with the HMI and only when those errors have surpassed a certain threshold do they devote more time to looking ahead. Hence, the TLC values at the moment of looking away were further divided into the values associated with the first glance of the task execution (TLC_{Off_F}) and all the consecutive ones (TLC_{Off_C}).

Empirical cumulative distributions were calculated for the data (see Figure 5.6) and a Kolmogorov-Smirnov (KS) test was performed to identify if they belong to the same distribution (see Table 5.2). The data seem to stem from different underlying distributions as the KS test rejected the null hypothesis (of data coming from the same distribution) at the 0.05 confidence level. Table 5.2 provides the D statistic that describes the maximum pairwise difference between all the curves.

Investigating the CDF curves closer in Figure 5.6(a), the vast majority of TLC values when driving on straight road segments is below 10 seconds. Given that no glances were observed to be longer in duration than that value, too, investigating beyond that point would not provide any additional information. Hence, the focus was shifted to low TLC values (see Figure 5.6(c)). It is immediately evident that TLC_{Off_F} is consistently higher than TLC_{Off_C} (mean value of 3.52 and 3.2 seconds, respectively), indicating that drivers compensate for control error build ups in the beginning of the task. Another interesting observation is that up to the value of 4.3 seconds, TLC_{Off_F} is also higher than TLC values during baseline driving (mean value of 3.52 and 3.11 seconds, respectively). This indicates that drivers ensure they are within a safe margin before looking away, even more so than when their focus is constantly on the road, as, in the latter case, they have more attentional resources available to account for and correct control errors. The crossover at that point, where TLC_{Off_F} starts to become shorter than baseline, indicates a point beyond which drivers feel confident to look away regardless of the TLC value.

For curved road driving (see Figures 5.6(b) and 5.6(d)), it appears that, up to the value of 2 seconds, TLC_{Off_F} is consistently higher than TLC_{Off_C} (mean value of 1.55 and 1.49 seconds, respectively), verifying the previous argument, and both of them consistently higher than TLC values during baseline driving (mean values of 1.55, 1.49 and 1.37 seconds, respectively). This reinforces the previous

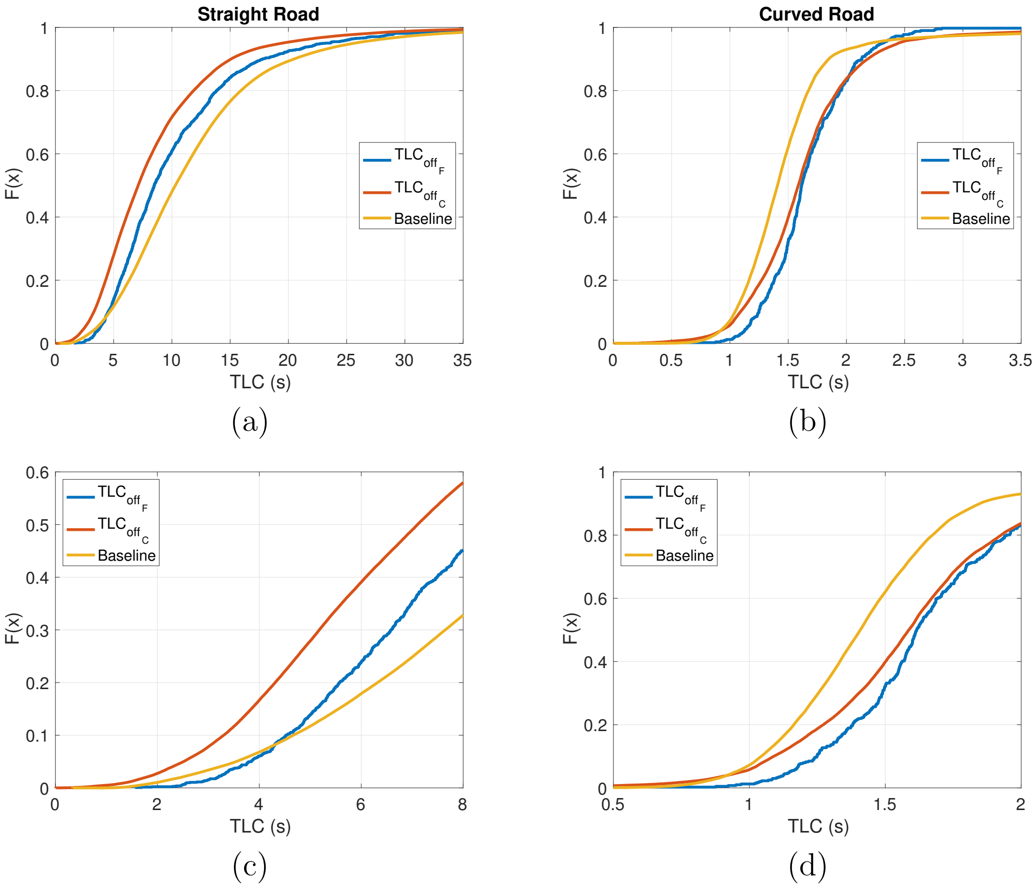


Figure 5.6: TLC_{off} under different driving conditions for straight and curved road. Plots (c) and (d) are enlarged versions of (a) and (b), respectively.

findings, as it shows that drivers took added precaution when the demand of the primary task increased. During baseline driving, drivers tended to drive closer to the lane boundaries (i.e. "cutting" the curves), while during HMI task execution they moved their vehicle closer towards the center of the lane, to allow for a bigger safety margin.

At this point it is worth noting that since baseline was always the first part of each drive, the observed behaviour might contain some artefacts of ordering effects. Based on the consistent results across conditions, however, at this point it can be concluded that, although TLC cannot be used as a metric to predict off-road glance durations, it is indeed a measure that drivers take into account to decide when it is safe to divert their gaze away from the road.

	TLC_{off_C}	Baseline
Straight Road		
TLC_{off_F}	$*D = 0.19$	$*D = 0.14$
TLC_{off_C}		$*D = 0.28$
Curved Road		
TLC_{off_F}	$*D = 0.08$	$*D = 0.32$
TLC_{off_C}		$*D = 0.25$

$*p < .001$

Table 5.2: Kolmogorov Smirnov test results for TLC CDFs.

5.3.2 Glance Durations and Headway Distance Adjustment

Having discussed how lateral control affects visual attention sharing, it is crucial to also investigate the relationship between longitudinal control and visual driver glance behaviour. Hence, the other main factor to investigate in terms of how drivers adjust their visual time sharing, would be headway (HW) distance from the lead vehicle. In particular, similarly to what was done for TLC, the analysis in this section focuses on whether there is a relationship between HW distance adjustment and onset.

Given that the drivers were instructed to maintain a certain HW distance, an initial approach would be to classify HW changes compared to target distance. However, given that each driver eventually used a different reference than the one directed, the above approach would not be very robust. Thus, an analysis regarding lead vehicle visual angle rate of change was conducted, instead. Tijerina et al. (2004) previously investigated closing gaps with a lead car, in a naturalistic driving setting. In their study, however, the glances away from the road were not necessarily related to HMI task execution. As a similar analysis in the context of HMI task execution has not been published since, it would be meaningful to validate their findings in the present study. Here, the investigation was performed both for opening and closing gaps with the lead car during HMI execution. Following the logic in Tijerina et al. (2004) and Maddox and Kiefer (2012), visual angle rate of change $\dot{\theta}$ was computed for each look-away moment and compared

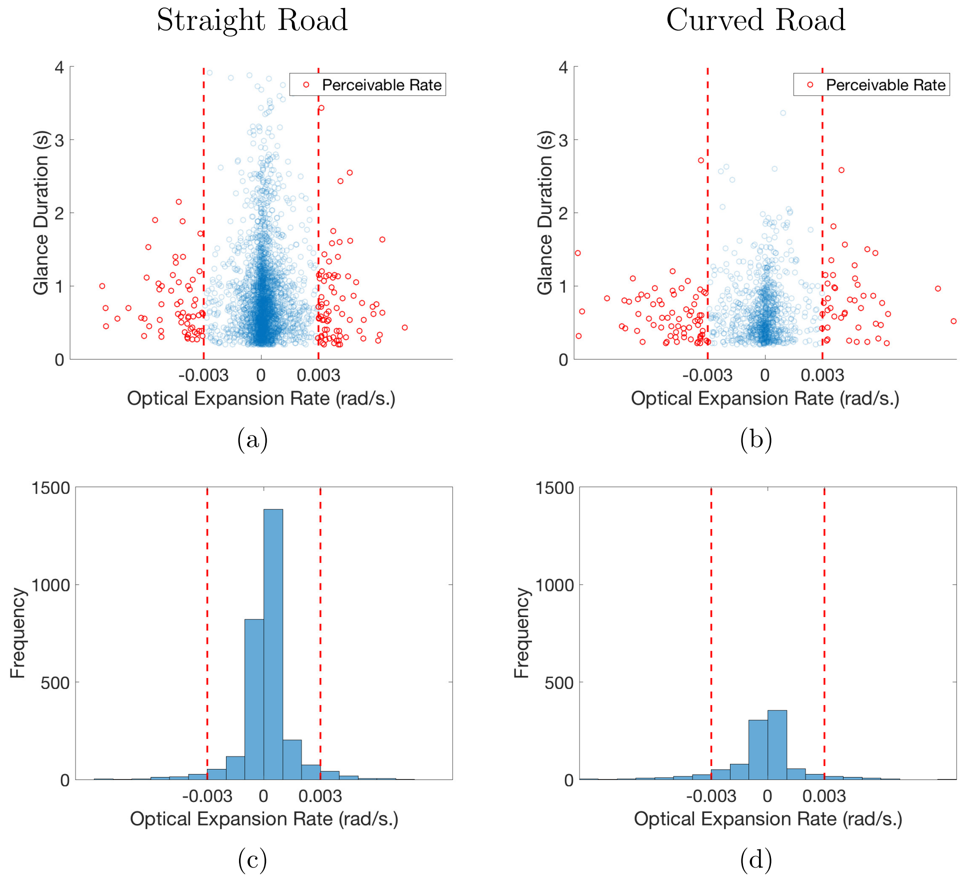


Figure 5.7: Visual angle rate of change of lead vehicle at the moment of looking away. The left column illustrates data corresponding to straight road driving, while the left column illustrates data corresponding to curved road driving. The top row (Figures (a) and (b)) illustrates the correlation between the optical expansion rate of the visual angle rate of change of the lead vehicle at the moment of look-away, against the subsequent off-road glance duration. The bottom row (Figures (c) and (d)) illustrate the distribution of said visual angle rate of change values, i.e. the number of off-road glances performed at each value.

against glance duration, using the following formula:

$$\dot{\theta} = \frac{-W \cdot \dot{R}(t)}{R(t)^2 + W^2} \tag{5.3}$$

where W denotes the width of the lead vehicle, $\dot{R}(t)$ denotes the rate of change of the distance of the two vehicles at time t and $R(t)$ denotes the distance of the two vehicles at time t. Similarly to TLC scatter plots, data was collapsed across all conditions for the two road types as that is where the biggest differences lied.

According to Maddox and Kiefer (2012), $\dot{\theta}$ is perceivable when below the

0.003 threshold for closing gaps and, thus, above -0.003 for opening gaps. In line with findings from Tijerina et al. (2004), the vast majority of glances away from the road were initiated at times where the visual angle rate of change was not perceivable by the driver (see Figure 5.7). In particular, at the moment of look-away, only 5.1% of the time was $\dot{\theta}$ perceivable by the participants. For curved road, this percentage was higher (13%) as drivers in general tended to drive closer to the lead vehicle when in curves. Consequently, as Tijerina et al. (2004) noted, it can be argued that drivers look away when $\dot{\theta}$ is essentially zero. Similarly, Tijerina et al. (2004) reported a total of 19% of the glances occurring when $\dot{\theta}$ was at a perceivable value. The lower percentage of perceivable $\dot{\theta}$ instances in the present study could be justified by the longer HW distances involved compared to the study by Tijerina et al. (2004). To begin with, drivers were instructed to maintain a distance of 70 meters from the lead vehicle. Due to the fact that the lead vehicle was not bound to the subject vehicle, however, there were a few instances where the participants almost "lost" the lead vehicle to a larger gap. As is evident in from Equation 5.3, the distance between the two vehicles has an inversely proportional relationship with $\dot{\theta}$, i.e. the larger the distance, the smaller the $\dot{\theta}$. Consequently, in cases where the distance between the subject and the lead vehicle grew substantially large, changes in HW would be more likely to not have been perceivable by the driver.

At this point, however, an interesting question arises; is there a difference in how perceivable $\dot{\theta}$ is between different driving conditions (namely driving while executing HMI tasks and driving with no concurrent tasks)? Looking into the respective baseline data, for the straight road, only 6.8% of the time was $\dot{\theta}$ perceivable during baseline driving, while for curved road this number was higher, at 20.3%. This indicates that, for the vast majority of the time, drivers would have been unable to perceive changes in HW at all, even during periods of driving without executing any concurrent task.

Consequently, although the findings here, that drivers tend to initiate off-road glances when $\dot{\theta}$ approaches zero are in support of the findings by Tijerina et al. (2004), this could be a by-product of the fact that this is simply where $\dot{\theta}$ values lie the majority of the time. In the present study, however, the above observed behaviour could also be an artefact of baseline always being the first part of each drive and, hence, participants showing different behaviour due to order effects.

Chapter 6

Modelling Drivers' Visual Attention Allocation for Prototype HMI Evaluation

The present chapter investigates the potential of using computational simulations as prototype HMI evaluation tools. Consistent with analyses presented earlier, throughout Chapter 5, the focus here will remain on the ability of models to replicate the drivers' visual attention sharing behaviour. For almost three decades now, a consistent effort has been made from the academic community to develop models that can capture driver behaviour under dual-tasking conditions and aid in better understanding, analysing and predicting it (e.g. Hankey, Dingus, Hanowski, Wierwille and Andrews, 2000; Horrey et al., 2006; Lee et al., 2016; Large et al., 2018, to mention a few). As discussed in Chapter 1, such models and simulation tools could be employed in the early stages of the production cycle to grant human factors specialists and researchers with insights regarding the usability and distraction potential of a prototype HMI design under concurrent driving conditions. Such insights could then facilitate informed decisions regarding design modifications and improvements prior to physical prototyping.

Briefly revisiting the HMI prototyping cycle, it can be described as a four-stage process, where the first two stages include non-physical conceptualisations of the new interface design (general), while Stages 3 and 4 involve the development of physical prototypes of varying detail. Computational models could be incorporated across all stages of the evaluation cycle as a means of exploring the viability of a new HMI design before moving into production testing. Such models, if they were to be used for prototype evaluation, they should be HMI task- and driving task-agnostic or, in other words, lend themselves well to a variety of devices, tasks and interaction scenarios.

Removing human participants from the loop, however, raises questions about the behavioural validity of the generated results. As discussed in Chapters 2 and 4, driving simulators can be evaluated in terms of their behavioural validity, i.e. the extent to which they can elicit the same behaviour from the drivers as the one that would be observed in corresponding real world conditions. Computational models of driver/HMI interaction have been quantitatively evaluated throughout the literature, predominantly in terms of how closely they can match the observed behaviour (see Lim and Liu, 2004; Salvucci et al., 2005*a*, for instance). It has been argued, however, that utility of such models might exist in their ability to predict relative differences in measures for different prototype HMI designs (Salvucci, 2009; Large et al., 2018). Hence, the concepts of relative and absolute validity can be also extended to such computational methods of evaluation, too. In particular, models of absolute behavioural validity could be used to check compliance with design and performance guidelines even before prototypes are built and human participants are employed, while models of relative behavioural validity could facilitate the comparison between alternative HMI designs in terms of their distracting potential.

The rest of this chapter is divided into three sections. The first one provides an overview of existing HMI interaction models that have been or can be used in the context of prototype HMI evaluation. In the second section two of those models are evaluated against data collected in this project. A novel model is also proposed, that accounts for some of the behavioural phenomena described in Chapter 5 that are not directly addressed by the existing models. Finally, the third section provides a comparison of the best performing model variants along with some discussion of the results.

6.1 Available Computational Models of Drivers' Visual Behaviour

A multitude of different dual-tasking models in the driving context have been developed over the years, either in an attempt to study human behaviour under such conditions, or to be used as prototype HMI evaluation tools. The computational model review in Chapter 2 provided a general overview from the perspective of the underlying behavioural assumptions and modelling approaches used in this context. The vast majority of such models and frameworks, however, (with the exception of ACT-R, which has seen wider use over the years, by various authors) have seen limited use in the academic community and have rarely undergone additional validation. Consequently, it would be meaningful to evaluate their performance on previously unseen data and, thus, their ability to generalise and their potential to be used as prototype HMI evaluation tools.

In the scope of this thesis, a set of such existing models was reviewed and a small subset of those was then validated on the collected data. In order to narrow down candidate models and decide on which ones would be further validated here, the following set of criteria was used:

- The model should be able to produce visual behaviour metrics (glance duration related metrics).
- The model should be already implemented and freely available for use, or be straight-forward to implement, without requiring intimate knowledge of specific behavioural theories or modelling frameworks.
- The model should allow the user to define a variety of visual-manual HMI tasks, without being restricted from the interface design, i.e. be interface-agnostic and provide a modular design environment in which the user can define different element combinations to define displays and tasks.
- The model should be easy to use.
- The model should be fast in simulations.

Following, all the models that could be candidates for use in virtual testing and prototype evaluation are revisited, to discuss aspects of the models' functionalities, that would be of interest to a human factors specialist when utilising them, before finalising the ones to be used in the upcoming evaluation. Namely, the following information is provided for each model:

- A slightly more in-depth presentation, than the one presented in Chapter 2, of how the model works.
- What type of performance metrics it can produce.
- How easily accessible it is to researchers and Human Factors specialists; e.g. Does one need a specialised license or is it open source? Is there a functional / ready-to-use implementation available? If not, is it easy to implement and test the model?

6.1.1 IVIS DEMAnD Model

The IVIS DEMAnD system was introduced by the Federal Highway Administration division of the United States Department of Transportation (Hankey, Dingus, Hanowski, Wierwille and Andrews, 2000). It is a software tool where the user can virtually replicate and evaluate prototype HMI designs. The software interface provides the user with a library of 198 pre-defined sub-tasks (e.g. adjust temperature and radio tuning), grouped by the resources needed to perform each one (namely, visual, auditory, manual, supplemental information processing (SIP) and speech). The user can use the existing sub-tasks, modify them or create new ones through the interface menu to represent a candidate HMI task. Additionally, the user is given the ability to define various other related parameters such as driver age, the position of the driver in the vehicle, position and size of interface to be evaluated and more. The model modifies default values, predefined for each task, based on user inputs about driver characteristics and primary driving task demands, to generate predictions on metrics such as mean single glance duration, mean number of glances, mean total visual task time, mean total task time and mean hand at task time. A summary of the results is then presented in the interface, highlighting where the driving performance is likely to be affected relative to driving without a secondary task.

Unfortunately, due to its age the task library is geared towards traditional panels where the presence of physical buttons and knobs is prevalent. Consequently, the representation of more modern interface tasks (e.g. hierarchical menus through touch screens) is not thorough. The model has not seen wider use since its creation and nowadays is not freely available for researchers to use.

6.1.2 The Queuing Network Model Human Processor (QN-MHP)

QN-MHP relies on the Natural GOMS Language (refer back to Chapter 2) to describe tasks by modelling the individual actions (or subtasks) that form them (Feyen, 2002). A set of rules needs to be generated to represent the task within the architecture. The inputs into QN-MHP, for example, in order to model a steering task, would be road and position related information, whereas the output of the model would be the driver's hand movement on the steering wheel. During driving, it needs to be determined whether the vehicle is within the lane boundaries, by using lane position as a measure. If the vehicle is not within the lane, a subtask of steering back into lane will be activated. QN-MHP has been previously used to model driver menu selection and visual search (Lim and Liu, 2004), providing a range of metrics regarding driving and visual performance during simulated HMI execution. However, substantial effort is required to develop models in QN-MHP, which restricts the potential users to researchers with extensive experience with the framework.

6.1.3 The Salience Effort Expectancy and Value (SEEV) Model

The Salience, Effort, Expectancy, and Value (SEEV) computational model is a model of selective visual attention and was introduced as a means of predicting visual scanning behaviour in different dynamic environments (Horrey et al., 2006; Wickens and Horrey, 2008). SEEV was initially developed and applied in the context of aviation and later on in driving related tasks, simulating how visual attention is allocated to different Areas of Interest (AOIs) during concurrent task execution. As the model name indicates, the attention switching is driven by four factors:

1. *Salience*, which refers to physical properties of events; the more salient an event, the easier it captures attention, e.g. a sudden brake light from the vehicle ahead.

2. *Effort*, which is an inhibitory factor discouraging one from switching attention between areas that are far apart.

3. *Expectancy*, which reflects one's tendency to look at sources that provide a large amount of information in short time (high event rate) more frequently and, finally,

4. *Value*, which represents the fact that one tends to allocate attention to sources providing information that are highly task relevant (hence having a higher value).

These factors are the components of an additive model that calculates the probability of attending a specific AOI (see Equation 6.1).

$$P(AOI) = s \cdot S - ef \cdot EF + ex \cdot EX + v \cdot V \tag{6.1}$$

The terms in capital represent the factors as listed earlier, while the coefficients s, ef, ex, and v, represent the weights (or relative influence) of each one of those four factors on visual scanning behaviour. The SEEV, as a stochastic model, can easily be evaluated in a Monte Carlo Simulation, to generate visual scanning trajectories between the different AOIs that represent a dual tasking scenario. Distributions of visual attention allocated in the different AOIs can be extracted and, hence, on- and off-road glance durations for different HMI tasks can be calculated. However, regarding model evaluation, all weighting factors need to be fit empirically and there is no objective method of defining them to generalise to previously unknown conditions.

6.1.4 Adaptive Control of Thought - Rational (ACT-R)

Salvucci and colleagues have created a computational implementation of ACT-R with versions in Lisp and Java programming languages, as well as a stand-alone software application with a Graphical User Interface (GUI) [1]. These provide a framework where the user, similarly to utilising a programming language, can develop models that represent certain tasks. Due to its customisable nature, ACT-R can be used to generate a wide range of metrics, such as task completion times, visual attention-related metrics and many more. The main components of ACT-R are modules, separated in perceptual-motor and memory modules, buffers, which are used to access the modules and the pattern matcher, which searches for a piece of knowledge about the execution of a task (production) that matches the state of a buffer at a given time.

ACT-R has been previously used to model driver steering control, lateral and longitudinal control as well as dual-tasking while driving (Salvucci, 2005, 2006; Brumby et al., 2007). The associated software and all related resources to the architecture are open source, freely available and thoroughly documented. However, it has been previously noted that building an ACT-R model for a new HMI is time-consuming, and requires extensive experience with the architecture (Salvucci, 2005).

6.1.5 Fitts' and Hick-Hymann Law

These two laws provide models of the human behaviour with regards to visual search and localisation, as well as target pointing, both of which are highly relevant and applicable in modelling HMI task interactions. Although not definite models of this interaction themselves, these two laws have been used in that

[1] Accessible at `http://act-r.psy.cmu.edu/software/`

context and are, hence, presented here for the sake of completeness but also to provide an introduction to the models described later, that make use of them.

Fitts' Law (Fitts, 1954) is based on Shannon's theorem (Shannon and Weaver, 1949) and can be used to predict the difficulty of a target selection task (known as Fitts's *Index of Difficulty (ID)*) and, consequently, the time needed to move to a target using a pointing device. It assumes a logarithmic relationship between movement time and target-distance and width, generally expressed as:

$$T = log_2(\frac{D}{W}),$$

where D is the distance from the target and W is the width of the target. This essentially translates to larger, closer targets requiring less information processing and, thus, less time to reach than smaller targets that are farther away.

The Hick-Hyman Law or Hick's Law (Hick, 1952) can be used to predict the time needed to make a decision when presented with multiple options. Hick's law assumes a linear relationship between reaction time and the number of available options, which for a set of equally probable options can be expressed as:

$$T = b \cdot log_2(n + 1),$$

where b is a constant usually set empirically and n the number of available options.

Both laws have been used independently used in Human Computer Interaction studies, however it has been argued that Hick's law complements Fitts' law and their combined use can yield enhanced predictions of human behaviour (e.g. Cockburn et al., 2007; Large et al., 2018). Cockburn et al. (2007), for example, reported near perfect levels of accuracy when using a model combining Fitts' and Hick's laws to predict static task times. Given the high ecological validity of such models, as well as the low complexity involved in implementing such equation-based models, this approach could be a great candidate for prototype HMI evaluation methods.

6.1.6 Distract-R

Salvucci and colleagues (Salvucci et al., 2005*b*; Salvucci, 2009) developed a software package called "Distract-R", based on the ACT-R framework, aimed at providing researchers with an easy to use tool for rapid prototype HMI evaluation. Distract-R is a user friendly interface that allows the experimenter to replicate HMI designs, using a range of virtual components such as buttons, rotary controls and voice commands. Moreover, the user can define a secondary task associated with the interface by defining the sequence in which actions need to be performed (button presses, text entries, etc.) to complete a task. After the interface and tasks are defined, the experimenter can also modify other aspects of the experiment, such as driver age (choosing between young and old), driving style (by defining steering aggressiveness and stability levels) and driving scenario (such as straight or curved road and presence or absence of a leading vehicle). Distract-R uses the ACT-R driver model and cognitive functions (Salvucci and Gray, 2004) along with Fitts' law for hand movement to target (Fitts, 1954) to simulate secondary HMI task interaction while driving. After the simulation is run, a variety of driver performance measures, such as mean secondary task completion time, total off-road glance duration, lateral deviation and velocity, heading error and reaction time in case of leading vehicle braking are provided to the user through the interface. The experimenter can then view some basic visualisations (bar charts) of said measures through which they can compare and evaluate different interfaces/tasks. Although there is no direct way to export the generated data to a file, the user can copy the data from the interface in text form to conduct further analyses.

Similarly to the IVIS DEMAnD software, Distract-R was created with static interfaces in mind (i.e. older panels with hardware buttons and knobs). Consequently, newer types of interfaces such as touch screen based ones with different screens between and within tasks, cannot be replicated as easily. However, given the ease with which the user can create and evaluate interfaces, as well as the speed of simulation, Distract-R provides a great tool for quick prototyping and evaluation of different HMI designs, without compromising the quality of generated data.

6.1.7 Large et. al Predictive Equations

Recently, Large et al. (2018) used data collected from simulator experiments to develop a set of equations that can predict visual demand of a new HMI task (Large et al., 2018). The equations were formed through linear regression fit of the data and combine Fitts' law for hand movement to target (Fitts, 1954) and Hick's law for visual localisation of target (Hick, 1952), following the "decision/search and pointing" approach proposed by Cockburn et al. (2007). Two sets of similar equations were created, representing *structured* and *unstructured* interfaces or task displays. In a unstructured display, one needs to search visually one item at a time, whereas in a structured display one can take shortcuts based on pre-existing knowledge about the structure. However, as the authors note, every interface and task is learnable by the user, hence rendering it eventually structured under their equation definition. The equations can predict the total off-road glance duration (TGT), the number of off-road glances (NG) and the mean off-road glance duration (MGD) associated with completing the task. Employing the equations for both structured and unstructured interfaces, the researcher can predict the range of performance that may be achieved from novice to expert users. Moreover, since repeated exposure would constitute an interface structured, the predictions can be used to infer design quality. As the authors explain, if the structured equations predict the observed behaviour better, that would be an indication that the interface or interactions associated with that task afford anticipation (Large et al., 2018). If, on the other hand, the observed behaviour is more similar to the one predicted by the unstructured equations, that could be indicative of the interface or task lacking learnability and, hence, being in need of design improvements (Large et al., 2018). All equations are rather straightforward and can be implemented in any programming environment (or even by hand, using a calculator - see Equations 6.2 through 6.7).

$$TGT_{st} = \left(\frac{log_2 N}{log_2(N + t_i)}\right)(0.029N + 0.44) + 0.11log_2 N + 0.11log_2\frac{D}{W} + 0.35 \quad (6.2)$$

$$TGT_{un} = \left(\frac{log_2 N}{log_2(N+t_i)}\right)(0.1N - 0.028) + 0.045 log_2 N + 0.11 log_2 \frac{D}{W} + 0.17 \quad (6.3)$$

$$NG_{st} = \left(\frac{log_2 N}{log_2(N+t_i)}\right)(0.021N + 1.04) + 1 \quad (6.4)$$

$$NG_{un} = \left(\frac{log_2 N}{log_2(N+t_i)}\right)(0.044N + 0.81) + 0.0071N + 1.96 \quad (6.5)$$

$$MGD_{st} = \frac{TGT_{st}}{NG_{st}} \quad (6.6)$$

$$MGD_{un} = \frac{TGT_{un}}{NG_{un}} \quad (6.7)$$

where:

st stands for structured,

un stands for unstructured,

N is the total number of selectable items on the screen,

t_i is the number of exposures to the interface,

D is the distance to target from hand position on steering wheel and

W is the target width.

6.1.8 Other models

Lee et al. (2016) recently proposed a model that predicts the visual demand of a new HMI design based on the visual saliency of the elements present in every screen. The model was published as a web tool where the experimenter can upload screen shots of the interface design evaluate its visual demand. One limitation of this approach is that it considers the saliency dimension of attention allocation only, and as shown by several authors there are many other factors involved in drivers' attention allocation (e.g. Horrey et al., 2006; Wickens and Horrey, 2008; Large et al., 2018). Moreover, the web interface does not provide the user with data output (only visualisation) making it hard to further analyse the generated results.

The extended KLM model (Pettitt and Burnett, 2010) could be another candidate in this context to calculate task completion times and total off-road time. It is based on a combination of the principle of the occlusion method where periods of vision and non-vision occur in sequential order and the Keystroke Level Model (KLM) for task execution (Card et al., 1980). HMI tasks are represented by a set of subtasks, each one requiring a specified amount of time to be completed, based on the performance of expert users. Moreover, there is still some subjective element to how it is constructed and which modules are used to represent parts of the task (Burnett et al., 2011).

Finally, the ACT-simple architecture (Salvucci and Lee, 2003) and CogTool interface (John et al., 2004) can both be used to define HMI tasks that are automatically translated and simulated in ACT-R. Due to modelling limitations, however, they require multiple hours to run and simulate the necessary interactions (Salvucci, 2005)

6.2 Predicting Visual Attention Sharing During HMI Engagement

Based on the review above and revisiting the inclusion criteria defined at the beginning of the previous Section, only two of the described models appear to meet them. Namely, Distract-R and the predictive equations by Large and colleagues were chosen to be evaluated here, since they generate metrics relevant to HMI evaluation, they do not require expert knowledge of the model to be used (a novice user can easily create and simulate a model) and can generate results fast.

Being based on the ACT-R cognitive architecture, Distract-R could be considered as having been thoroughly validated, given the rich body of distraction related modelling work using ACT-R (e.g. Salvucci, 2005, 2006; Brumby et al., 2007). Moreover, Distract-R has recently been used in assessing driver distraction induced by in-vehicle displays (see e.g. Lee, 2014). The predictive model by Large et al. (2018), on the other hand, given its recent publication, has not yet been further validated beyond the data used by the authors.

Given the results from previous validations, the underlying frameworks used to identify them, as well as the type of interface and task used in this study, the models were expected to only partially capture the observed HMI interaction behaviour, since they were created and validated against different HMI tasks and do not explicitly address specific behavioural aspects that were observed during the experiments, such as the visual chunking behaviour discussed in Chapter 5. To investigate that assumption, both of the models were validated here using the data from the UoLDS experiment. In particular, model data were compared against the fixed base, constant speed simulator data, mirroring the validation done by the original authors and, thus, providing a “fair” evaluation. After the results from the two existing models are discussed, a novel model is proposed that accounts for the visual chunking discussed in Chapter 5 which has not been previously considered as a behavioural phenomenon in such HMI interaction models.

Model performance was evaluated through the *Root Mean Square Error* ($RMSE$) for different visual behaviour metrics. In particular, model performance for each behavioural metric was evaluated by a comparison of the mean values of the metric across the three HMI tasks, between the observed and model generated data. This comparison yielded $RMSE_{abs}$, which refers to how well the models predict the exact values of the observed means (i.e. their absolute fidelity). With y_i representing the observed mean value of the metric for task i (where $i = 1$ refers to the easy HMI task, $i = 2$ refers to the medium HMI task and $i = 3$ refers to the hard HMI task), $RMSE_{abs}$ can be calculated by the following equation:

$$RMSE_{abs} = \sqrt{\frac{\sum_{i=1}^{3} (y_i - f_i)^2}{3}}, \tag{6.8}$$

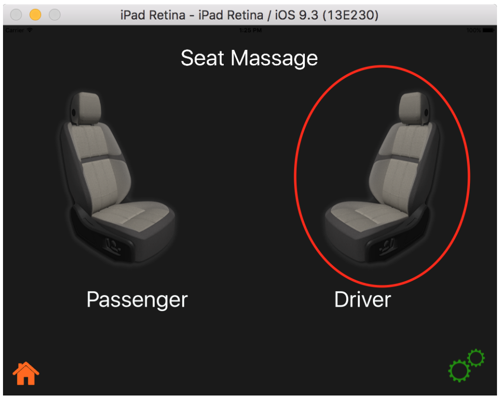
Additionally, $RMSE_{rel}$ was also calculated for the ratios of the metrics between medium and easy, hard and easy and hard and medium tasks, using the same formula. These scores can be used to evaluate how well the models predict the relative differences between HMI tasks for each metric (i.e. their relative fidelity). In this case, however, instead of using the values of the means, the ratios of those values are used, pairwise, to describe the relative difference between the tasks. With y_j^r representing the ratio of the observed mean values of the metric for combination j (where $j = 1$ refers to the ratio of medium over easy HMI task, $j = 2$ refers to the ratio of hard over easy HMI task and $j = 3$ refers to the ratio of hard over medium HMI task), $RMSE_{rel}$ can be calculated by the following equation:

$$RMSE_{rel} = \sqrt{\frac{\sum_{j=1}^{3} (y_j^r - f_j^r)^2}{3}}, \tag{6.9}$$

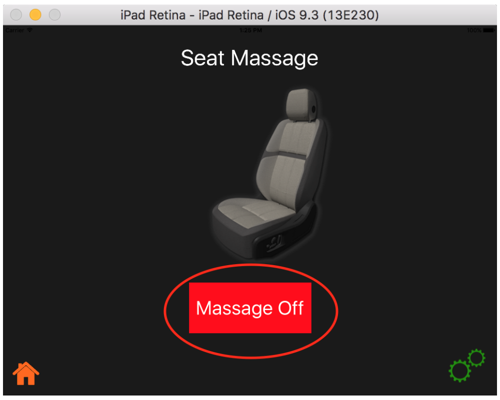
6.2.1 Distract-R

The three HMI tasks described in Chapter 3 were manually recreated in the graphical interface of Distract-R, using the “button” module to represent icons. The scrolling action of the hard task was represented by a rotary control action in Distract-R, using different angles of rotation to simulate a range of different approaches to the scrolling action (e.g. one swift or a slow and gradual scroll). Figures 6.1 through 6.3 illustrates how the three HMI tasks were replicated within the Distract-R environment. Since the interface does not allow for multiple screen design, all buttons were placed in their respective positions in a single Distract-R screen, as they would appear if all HMI screens were drawn onto the same Distract-R screen all at once. If two buttons from different screens overlapped, they were placed next to each other. This approach had no effect on results, as Distract-R uses the same assumptions as ACT-R to define visual search and encoding times, which are defined as fixed values (Salvucci, 2009). Consequently, the additional clutter on the screen does not increase any visual related metrics.

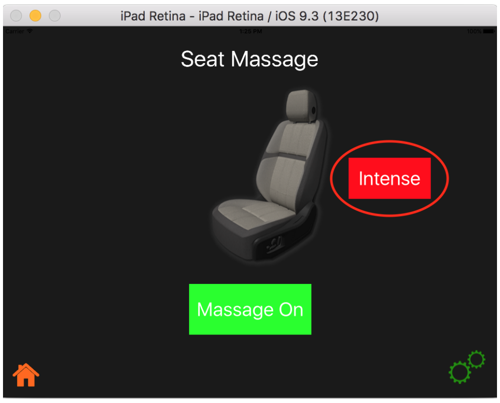
In terms of action mapping, the vast majority of the individual interactions, as described in Chapter 3, are target touch actions (equivalent to button presses), hence the *press* operator was used. As mentioned above, the *rotate* operator was used to simulate scrolling in the hard task. The Distract-R model was initially tested with all parameters set to default values (henceforth referred to as “Distract-R”). In particular, driver age was by default set to “Younger”, representing the age range $20 - 30$ and the “Steering Aggressiveness” and “Stability Factor” were set to represent the following values in the underlying ACT-R driver model; $k_{near} = 3.4$, $k_{far} = 13.6$, $k_I = 2.55$, $\theta_{stable} = .025$ and $\dot{\theta}_{stable} = .0125$. Next, to investigate the range of behaviour the model can exhibit, a grid exploration of parameters was used to vary driver age between "Younger" and "Older", as the interface allows, and steering behaviour between 0.75 and 1.25 for the "Steering Aggressiveness" and "Stability Factor", as well as visual chunking, i.e. how many consecutive actions the driver performs in a single glance towards the interface, using the the +*press* operator (henceforth referred to as "Distract-R + Drivers" and "Distract-R + Chunking", respectively). All possible chunking strategies and value combinations were evaluated. Regarding the steering behaviour factors, values outside of the aforementioned range were found to generate erratic behaviour and unrealistic results (most probably due to the increased controlled errors that led to uncontrollable lane weaving). Finally, a model variant containing both of the variations explained above was also tested to investigate their combined effect (henceforth referred to as "Distract-R + All").



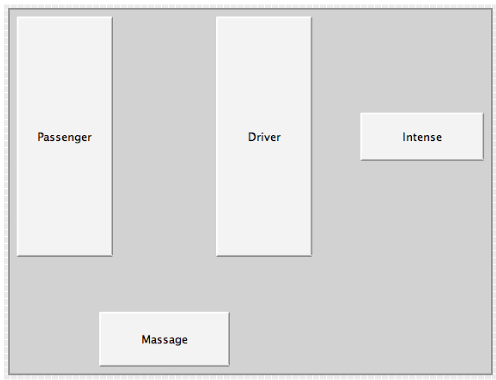
(a) Screen 1 - Press



(b) Screen 2 - Press

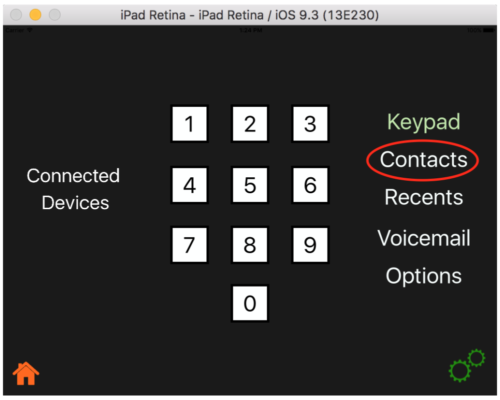


(c) Screen 3 - Press

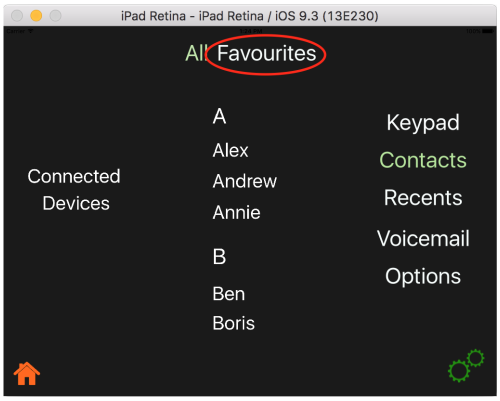


(d) Distract-R representation

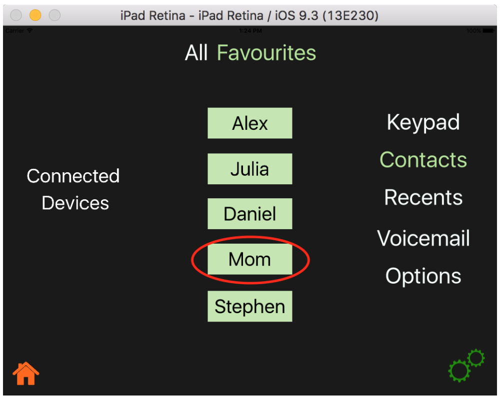
Figure 6.1: Easy HMI task and corresponding Distract-R representation. Red circles denote the areas on the screen / buttons that the user needed to press in order to move to the next screen and complete the task.



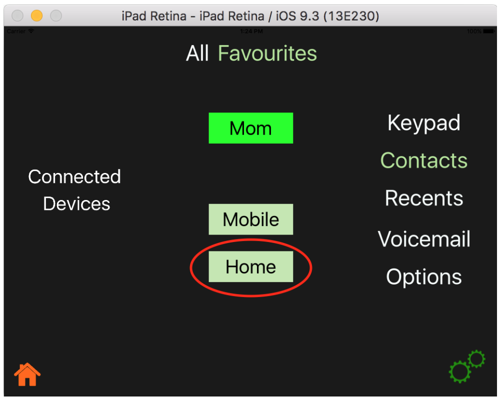
(a) Screen 1 - Press



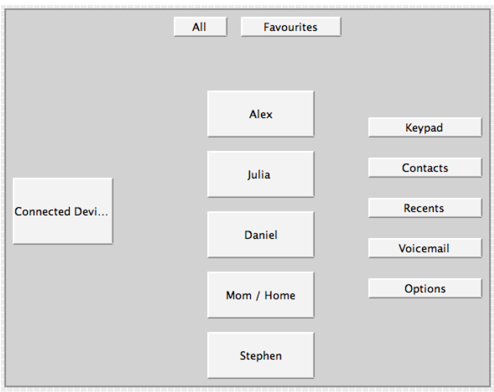
(b) Screen 2 - Press



(c) Screen 3 - Press

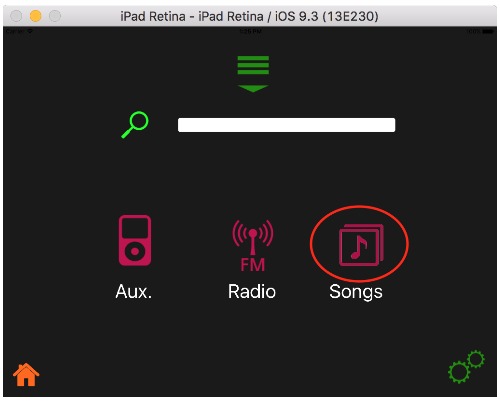


(d) Screen 4 - Press

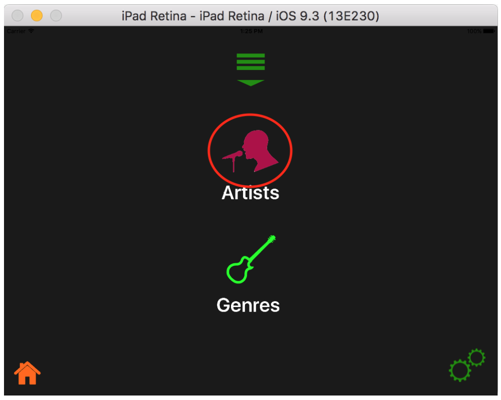


(e) Distract-R representation

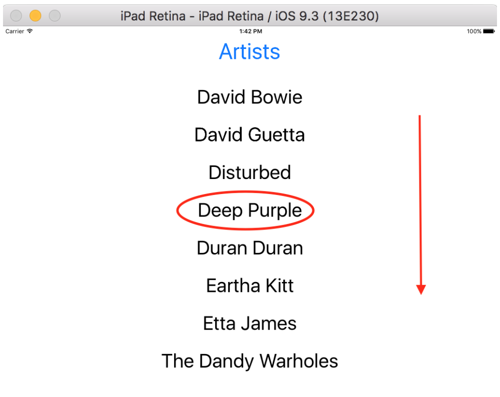
Figure 6.2: Medium HMI task and its corresponding Distract-R representation. Red circles denote the areas on the screen / buttons that the user needed to press in order to move to the next screen and complete the task.



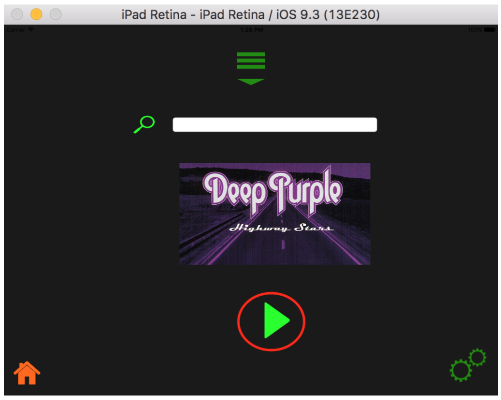
(a) Screen 1 - Press



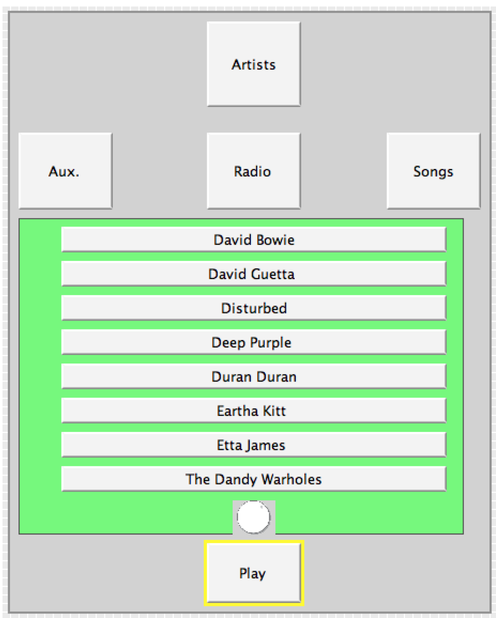
(b) Screen 2 - Press



(c) Screen 3 - Scroll and Press



(d) Screen 4 - Press



(e) Distract-R representation

Figure 6.3: Hard HMI task and its corresponding Distract-R representation. Red circles denote the areas on the screen / buttons that the user needed to press in order to move to the next screen and complete the task. The red arrow denotes that the user needs to scroll down the list to locate the desired target.

After simulation, total off-road glance duration and mean off-road glance duration for each simulated task execution were extracted from Distract-R. Their mean values were compared against the respective observed mean values from the experimental data. Figure 6.4 shows model calculated metrics, while Table 6.1 summarises model performance against the observed data, based on the $RMSE$ metrics discussed above.

Inspecting the $RMSE$ values for total off-road glance duration, it appears that none of the model variants can predict the observed values with accuracy (especially considering that all $RMSE$ values here are larger than the observed means) for either absolute or relative validity. Visually inspecting Figure 6.4, it appears that all variants of the model overestimate total off-road glance duration for the hard task. As described in Chapters 4 and 5, drivers had different approaches to how they conducted the "scroll" operation in the hard task (only sporadically checking for the desired target or devoting no glances to it at all). As a result, the mean values of total off-road glance duration and mean off-road glance duration for the medium task were slightly higher than those of the hard task. This type of behaviour cannot be accounted for with Distract-R, as it assumes that the drivers would look away as per the sub-task demand. The three tasks, as coded in Distract-R provide a clear difference in difficulty level and effort required to be completed (due to the increasing number of interactions required and how the model calculates metrics based on those), hence Distract-R appears to overestimate the visual demand of the hard task.

For mean off-road glance durations, on the other hand, the Distract-R + Chunking outperforms the other model variants both in terms of absolute and relative validity, while adding the different drivers feature does not seem to improve performance. For absolute validity, in particular, the model manages to predict the observed values quite closely, achieving a small error of 0.28, considering the observed means and range of values. As discussed in Chapter 5, devoting fewer glances to the HMI task leads to longer mean off-road glance durations. This justifies why Distract-R + Chunking provides improved predictions, as it simulates the execution of multiple interactions within one single glance. The fact that the first two variants seem to predict the observed behaviour for the hard task better (see Figure 6.4), while the latter two overestimate it, is because, as mentioned earlier, there is a mismatch between how the task is performed by the model and how real drivers performed it. Hence, lower mean off-road glance durations are observed than what the task structure would dictate.

Overall, in terms of relative validity capabilities, all variants seem to rank the medium and hard task correctly against the easy task but not against each other. This is due to the overestimation of total off-road glance durations for the hard task.

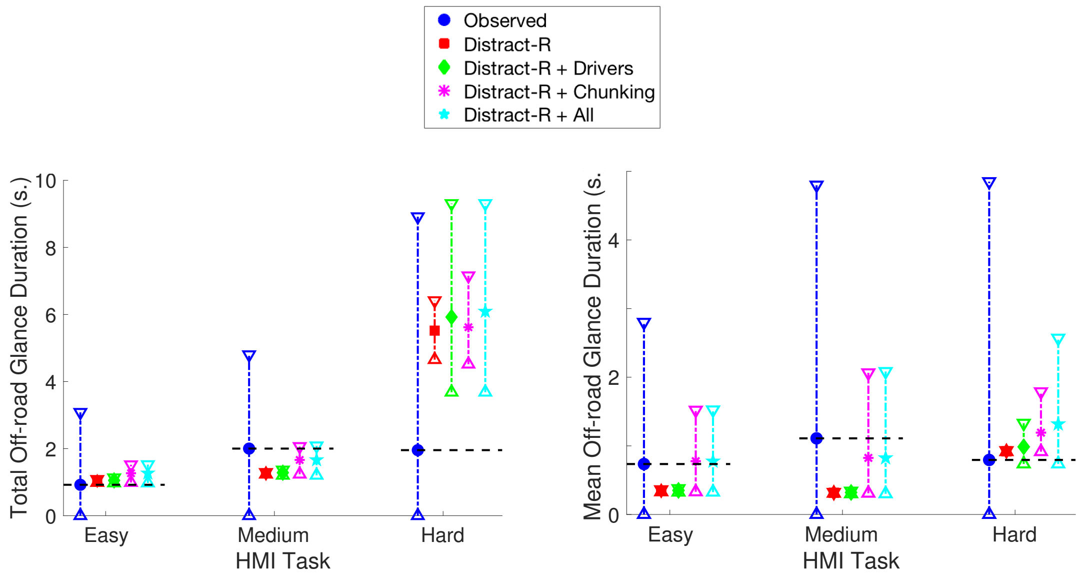


Figure 6.4: Distract-R model performance. Observed mean values for the behavioural metrics are plotted against model generated values. The errorbars illustrate the range of the metric (i.e. the minimum and maximum values in the independent task executions). Distract-R is the model variant with the default Distract-R values, Distract-R + Drivers is the variant where different values for driver age and steering behaviour were tested, Distract-R + Chunking is the variant where visual chunking was tested and Distract-R + All is the variant with both variations.

6.2.2 Large et al. Predictive Equations

The predictive equations as described by Large et al. (2018) were simulated using MATLAB. In consistence with their approach, for each one of the HMI tasks,

Distract-R				
	Default	+ Drivers	+ Chunking	+ All
Absolute				
Total Off-road Glance Duration	2.10	2.33	2.13	2.40
Mean Off-road Glance Duration	1.60	1.72	0.56	0.67
Relative				
Total Off-road Glance Duration	2.75	2.97	1.99	2.25
Mean Orr-road Glance Duration	0.32	0.34	0.39	0.46

Table 6.1: $RMSE$ values for the Distract-R model and variants.

every screen of the task was implemented and simulated independently, i.e. a different equation was used to represent each sub-task. In the lack of a more direct way to implement it, the scrolling sub-task was modelled as a single button press, i.e. in the same way that all other individual interactions were modelled, assuming that drivers would separate it from the subsequent press. Unlike Distract-R, these equations do not lend themselves well to investigating the effect of visual chunking. Since they were not developed to account for single glance durations, no manipulation could be performed to adjust results for multiple interactions compressed in a single glance. The equations were used to calculate total off-road glance duration, mean off-road glance duration and number of off-road glances, for both the structured and unstructured versions of the equations. Figure 6.5 shows model calculated metrics, while Table 6.2 summarises model performance against the observed data, based on the same $RMSE$ metrics as previously.

For total off-road glance durations and number of glances, on the other hand, both models appear to fail to replicate the observed behaviour. Visually inspecting Figures Figure 6.5(a) and 6.5(b) reveals that both models overestimate total off-road glance durations, as well as the number of glances. The overestimation of the number of off-road glances can be attributed to the fact that no visual chunking is supported and the model assumes that a task execution requires at least as many glances as the individual interactions required. The overestimation

Large et. al Predictive Equations

	Unstructured	Structured
Absolute		
Total Off-road Glance Duration	2.67	2.98
Mean Off-road Glance Duration	0.50	0.21
Number of Off-road Glances	8.14	4.30
Relative		
Total Off-road Glance Duration	0.58	0.15
Mean Off-road Glance Duration	0.28	0.29
Number of Off-road Glances	0.03	0.25

Table 6.2: $RMSE$ values for the Large et. al predictive equations.

of the total off-road glance durations, on the other hand, can be attributed to the fact that the model assumes that a visual search is performed with each task execution. However, that is not always the case as drivers usually anticipate where the next task target will appear, especially after they get more familiar with the task (Cockburn et al., 2007).

For mean off-road glance durations, the Structured model seems to be able to predict the observed data with good accuracy, achieving a low $RMSE$ of 0.21 and indicating high absolute validity. However, this accuracy could be argued as somewhat artificial, since it is essentially driven by the overestimation of both total off-road glance time and number of glances.

When it comes to relative validity, both model variants manage to rank the tasks correctly against each other, something that is can be verified both visually from the plots in Figure 6.5, as well as the low $RMSE$ values achieved when comparing observed and predicted means ratios. Hence, it can be argued that overall, both model variants offer a very good prediction of relative differences in the observed values, but do not manage to approximate the absolute observed values.

6.2.3 Proposed model

Here, a novel model is proposed, that draws inspiration from approaches used both in Distract-R and in Large and colleagues' predictive equations (Large et al.,

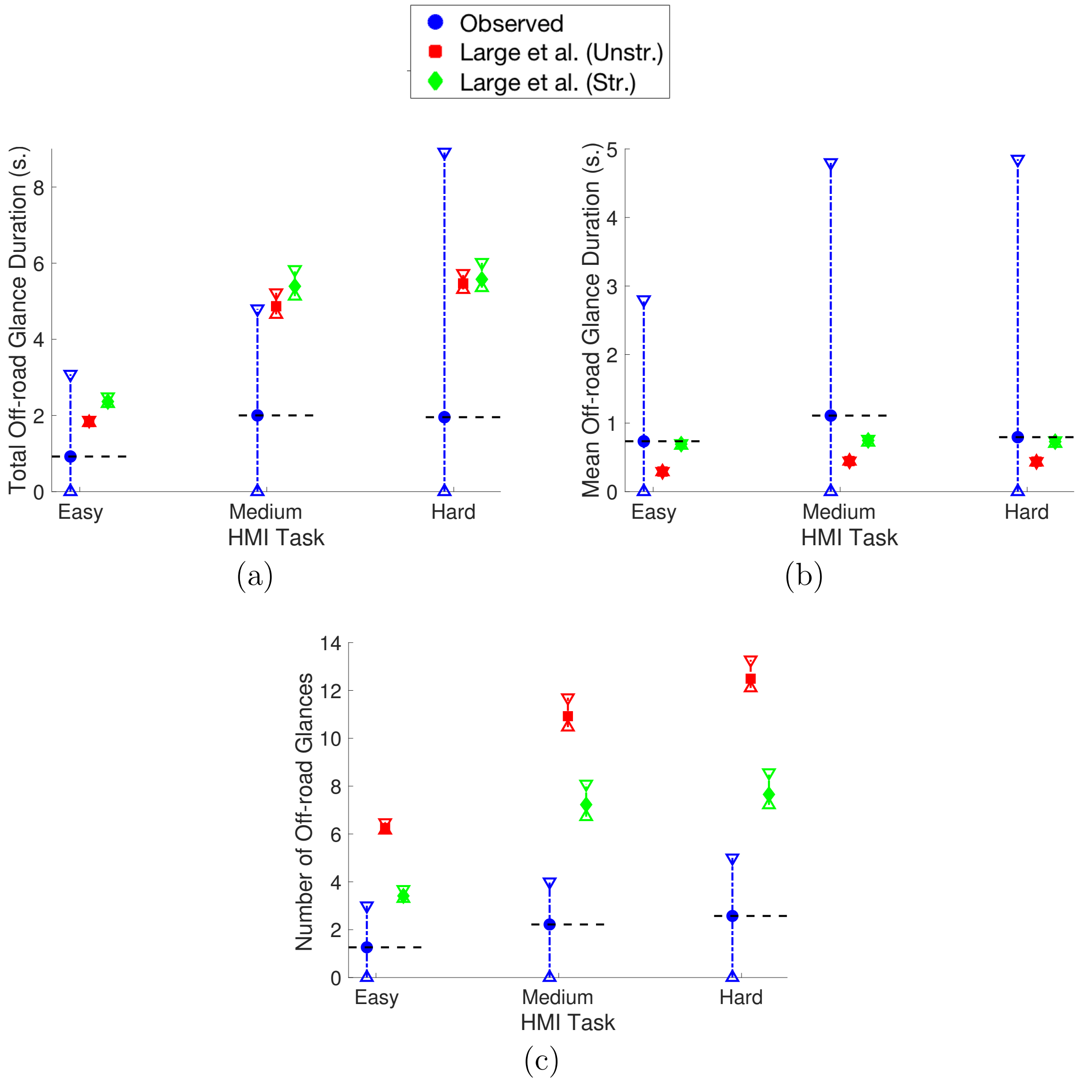


Figure 6.5: Large et al. (2018) predictive equations performance. Observed mean values for the behavioural metrics are plotted against model generated values. The errorbars illustrate the range of the metric (i.e. the minimum and maximum values in the independent task executions).

2018) and is based on Fitts' law (Fitts, 1954) and Hick-Hyman law (Hick, 1952) to calculate visual attention metrics for each interaction. Two variants of the model are discussed; a naive version (henceforth referred to as Naive model) that assumes drivers devote one glance per interaction and a slightly more complex one that accounts for the visual chunking behaviour discussed in Chapter 5 and also implemented in the Distract-R variants (henceforth referred to as Chunking model). The two variants of the proposed model, as well as the features used in each one of them, were decided after reviewing existing models and modelling frameworks but also after examining the behaviour drivers exhibited when performing HMI tasks (through the review of recorded video data, as well as eye-tracking and vehicle control data).

The proposed model was not fitted to data to extract parameter values but was, instead, designed to model the mechanics of the interaction itself, only based on HMI task characteristics. Hence, a few assumptions needed to be made, which were kept to minimal complexity. Similarly to the approach by Large et al. (2018) and in agreement with Cockburn et al. (2007) it was assumed that each glance towards the interface for the execution of a single interaction consisted of a visual search element (referred to as T_{hick} and calculated using Hick's law component - see Equation 6.10) and a manual execution component (referred to as T_{fitts} and calculated using Fitts' law - see Equation 6.11).

$$T_{hick} = 0.1 \cdot log_2(n + 1) \tag{6.10}$$

where n is the number of possible targets on the screen. The 0.1 factor was used so that equation yields a minimum T_{hick} of 0.1 s when only one possible target is available (that can also be used for when the driver knows exactly where to press without the need of an additional visual search).

$$T_{fitts} = 0.1 \cdot log_2(\frac{D}{W} + 0.5) \tag{6.11}$$

where D is the distance of the driver's hand from the target and W is the width of the target. This Fitts' law equation used here is the same used by ACT-R and Distract-R to calculate hand to target pointing time.

Hence, for the Naive model, the duration of a single off-road glance could be calculated as the sum of the two components:

$$SGD = T_{hick} + T_{fitts} \tag{6.12}$$

Given that the naive model assumes one glance per sub-task interaction, the total off-road glance duration for a task execution would be:

$$TORT = \sum_{i=1}^{N} SGD_i \tag{6.13}$$

where i denotes the individual subtask/interaction and N is the total number of interactions. In the case of the Naive model, it holds that $N = NGD$, i.e. the total number of interactions (N) is the same as the number of glances employed during a task execution (NGD), given that the model allocates one glance per interaction.

Finally, the average off-road glance duration for a task execution would be:

$$MGD = \frac{TORT}{N} \tag{6.14}$$

The Naive model, as defined above, although based on a previously validated method of combining Fitts' and Hick's law, has ample room for improvement, especially if one would be looking to create a more realistic and accurate model of the driver/HMI interaction. The findings in Chapter 5 regarding visual chunking (i.e. the act of performing more than one subtasks in a single glance) suggested that this feature should be included in the proposed model. This was enhanced by the fact that it has not been previously coded in a model (the chunking option in Distract-R allows the user to manually select the number of subsequent actions to be collapsed within a glance and does not model the process per se).

In order to model visual chunking, since the aim was to try and replicate the behaviour itself, again some assumptions had to be made. Initially, it was assumed that for a driver to continue looking on the interface after completing a single subtask/interaction, they should know where the subsequent target would appear. Consequently, no visual search would be needed in that case. Not knowing where the subsequent target would be and needing to initiate a novel visual search would prolong the time the driver would need to spend looking away from the road and, thus, they would have chosen to turn their gaze towards the road, to ensure safe driving, before carrying on the remainder of the HMI task. To reflect this, for each subsequent interaction within a single glance, the Hicks search component was set to its minimum value of $T_{hicks} = 0.1$ seconds, as described earlier.

Moreover, and primarily based on observations from the study conducted here, if a driver chose to consecutively execute an additional interaction, they would not remove their hand from the proximity of the interface. This "hovering" distance was arbitrarily set to 2 cm for the model (different distances were tested, too, with no effect in the resulting T_{fitts} time), hence modifying the calculation of the Fitts' pointing component to:

$$T_{fitts} = max\{0.1 \cdot log_2(\frac{2}{W} + 0.5), 0.1\} \tag{6.15}$$

From the above, it can be concluded that each additional action combined in a single glance would require 0.2 seconds of additional visual time, 0.1 from the Hick's locating component and 0.1 from the Fitts' pointing component.

Considering the argument by Wierwille (1993) that drivers would try to get the necessary information from an HMI within a second or less and in no longer than 1.5 s, an"optimal" average single off-road glance duration of 0.5 s can be defined. Hence, it was assumed that the driver would not combine more than five additional interactions (a total of six) in a single glance, since $0.5 + 5 * 0.2 = 1.5$.

In order to define whether a driver would employ visual chunking, a stochastic approach was taken, using a random *chunking* probability p_{ch}. The probability to perform additional interactions in a single glance was assumed to decay based on the number of additional interactions to be performed, i.e. a driver was less likely to combine 3 interactions in a single glance than they were to combine 2. Moreover, to investigate different driving styles and risk taking behaviours, a *risk* probability p_r was also used to scale p_{ch}, where $p_r \in [0, 1]$. A linear decay was used to model this decrease, since it was the simplest method.

$$p_{ch}^n = p_r \cdot (1.2 - 0.2 \cdot n) \tag{6.16}$$

where n is the number of additional interactions combined in a single glance, i.e $n = 1$ denotes that a total of two interactions (one additional) were performed in a single glance. From the above, it stems that a risk adverse driver (i.e. $p_r = 0$ and, consequently $p_{ch} = 0$), would employ one glance per subtask/interaction and the model would default to its naive version. A high risk driver on the other hand ($p_r = 1$) always combines at least one additional interaction in a single glance (since $p^1_{ch} = 1$).

The two variants of the proposed model were used to calculate single off-road glance duration, mean off-road glance duration, total off-road glance duration and number of off-road glances. Figure 6.6 shows model calculated metrics against the corresponding observed values and Table 6.3 summarises model performance against the observed data, based on the same $RMSE$ metrics as previously.

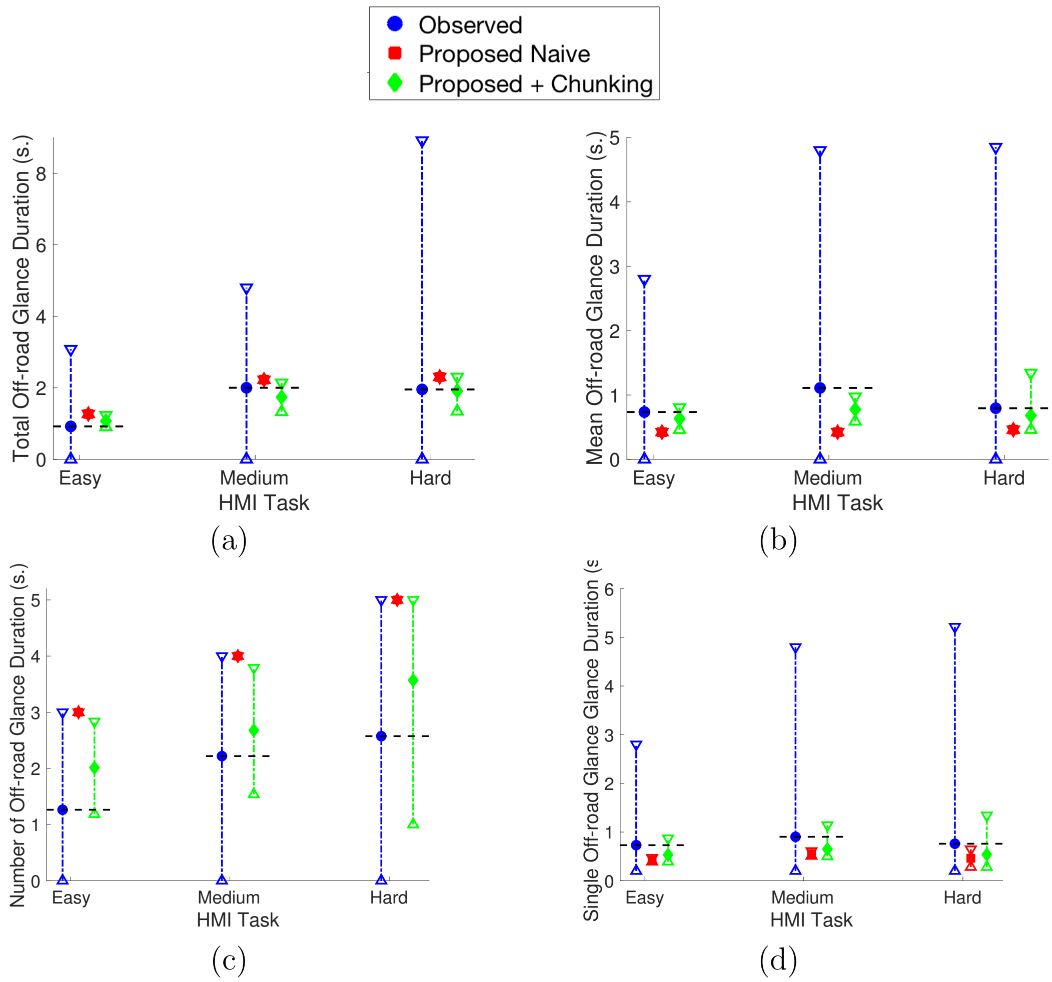


Figure 6.6: Proposed model performance. Observed mean values for the behavioural metrics are plotted against model generated values. The errorbars illustrate the range of the metric (i.e. the minimum and maximum values in the independent task executions).

Proposed model

	Naive	+ Chunking
Absolute		
Total Off-road Glance Duration	0.31	0.18
Mean Off-road Glance Duration	0.48	0.21
Number of Off-road Glances	2.01	0.77
Single Off-road Glance Duration	0.32	0.23
Relative		
Total Off-road Glance Duration	0.30	0.38
Mean Off-road Glance Duration	0.36	0.19
Number of Off-road Glances	0.33	0.31
Single Off-road Glance Duration	0.05	0.02

Table 6.3: $RMSE$ values for the proposed model variants.

Visual inspection of the plots in Figure 6.6 shows that apart from overestimating the number of off-road glances, both variants approximate the data very well, something that is also evident by the overall low $RMSE$ values for the different metrics. The Chunking model seems to always perform better and achieve more accurate predictions of the observed data. Considering the observed means and ranges of observed data, it can be argued that the Chunking model shows high absolute validity across all metrics, with the exception of number of glances. It is worth noting at this point, that the proposed model also provides a prediction of single glance durations, an important metric in HMI evaluation (NHTSA, 2012) that is not available form the other models here. For number of glances, the Naive model shows poor performance, as expected based on its assumptions, while the Chunking model, although having a medium $RMSE$ value, seems to generate good predictions, especially considering the observed means and range of observed data.

Finally, both models appear to be doing very well in capturing the relative differences in metrics between the different tasks, showing a high level of relative validity. Particularly interesting is the fact that the models could match the relative differences in the observed behaviour between all tasks, across all metrics. This is verified both from the plots in Figure 6.6 and the low $RMSE$ values for both models when comparing observed means ratios.

6.3 Discussion

Despite being evaluated here in previously unseen data, all models managed to generate good predictions for at least some of the metrics investigated, showing their potential for high levels of absolute validity. In particular, Distract-R + Chunking managed to predict mean off-road glance durations with low error and so did the structured version of the Large et al. predictive equations. Neither of those models, however, managed to predict the other metrics accurately. The proposed Chunking model, although simple in its assumptions, managed to outperform all other models, by achieving the lowest error in predicting the observed values across all metrics, consistently. Moreover, the proposed model could account for single glance durations, an important metric in HMI evaluation, that cannot be extracted from the other tested models.

In terms of relative validity, both variants of the Large and colleagues predictive equations model and both variants of the proposed model appeared to be able to rank the tasks against each other correctly with low error. No variant of Distract-R, however, managed to provide high relative validity in this case.

Figure 6.7 provides an overview of the performance of the best model variants, for each metric.

Given the overall good performance of the proposed Chunking model, as well as the Distract-R + Chunking model in predicting mean off-road glance durations, it is evident that including the visual chunking behaviour generates a more realistic model and improves model performance. Consequently, it can be argued that it is a behavioural phenomenon which should be further investigated and more rigorously applied in modelling efforts in the future.

At this point, and in order to validate the proposed model a bit further, both of its variants were evaluated against additional data, namely against the fixed base, constant speed, curved road simulator data, as well as the constant speed (straight road) real world data. The results are presented in Tables 6.4 and 6.5, respectively.

From the $RMSE$ values obtained, it is evident that that the high accuracy

Proposed model		
	Naive	+ Chunking
Absolute		
Total Off-road Glance Duration	0.36	0.28
Mean Off-road Glance Duration	0.22	0.09
Number of Off-road Glances	1.57	0.51
Single Off-road Glance Duration	0.15	0.06
Relative		
Total Off-road Glance Duration	0.86	0.93
Mean Off-road Glance Duration	0.37	0.25
Number of Off-road Glances	0.68	0.65
Single Off-road Glance Duration	0.07	0.06

Table 6.4: $RMSE$ values for the proposed model variants, evaluated in fixed base, constant speed, curved road simulator data.

Proposed model		
	Naive	+ Chunking
Absolute		
Total Off-road Glance Duration	0.67	0.32
Mean Off-road Glance Duration	0.19	0.10
Number of Off-road Glances	1.78	0.55
Single Off-road Glance Duration	0.07	0.07
Relative		
Total Off-road Glance Duration	0.60	0.65
Mean Off-road Glance Duration	0.29	0.12
Number of Off-road Glances	0.45	0.40
Single Off-road Glance Duration	0.16	0.23

Table 6.5: $RMSE$ values for the proposed model variants, evaluated against constant speed, straight road, real word data.

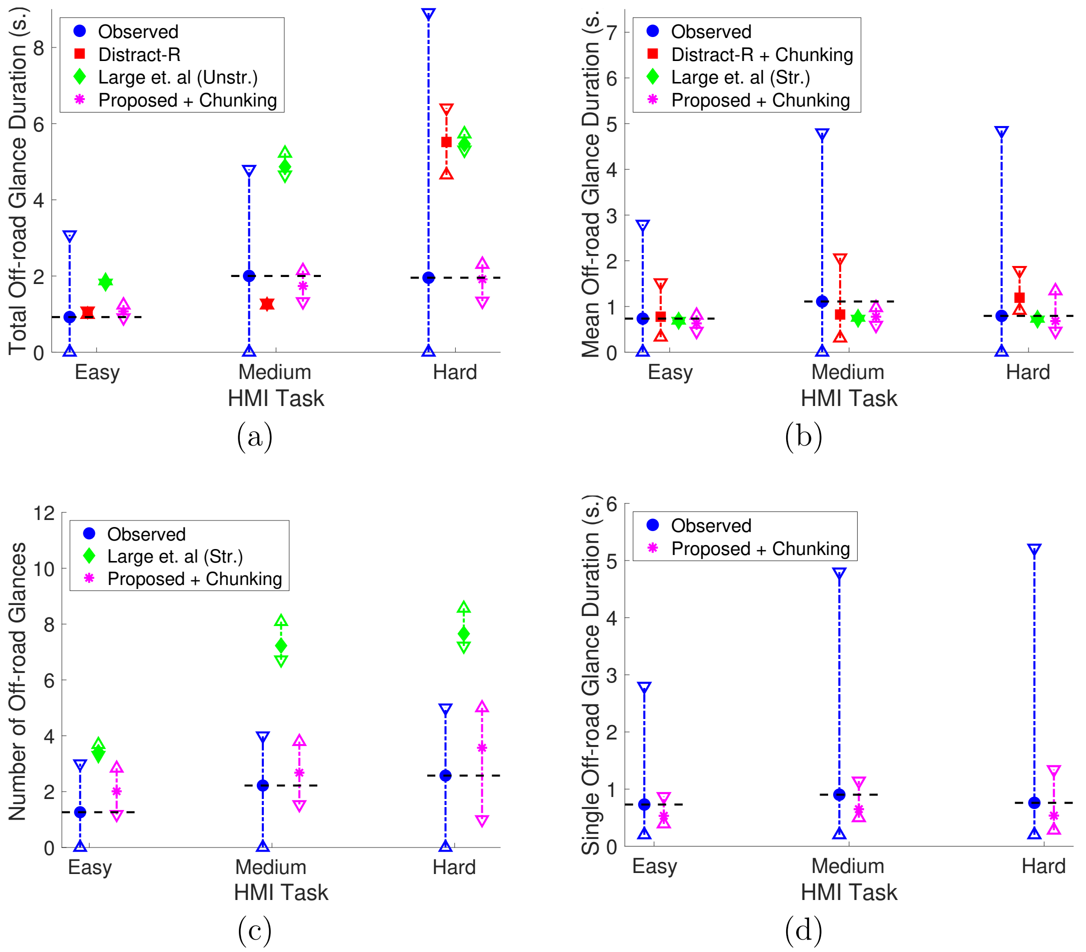


Figure 6.7: Best model variants performance. Observed mean values for the behavioural metrics are plotted against model generated values. The errorbars illustrate the range of the metric (i.e. the minimum and maximum values in the independent task executions).

performance of the proposed Chunking model remains at predicting the observed values. In fact, with the exception of total off-road glance durations, for which the error slightly increased, the model seems to be performing better than in the previous dataset. Moreover, the model seems to maintain a good level of relative validity, too, achieving low $RMSE$ when comparing the observed means ratios for the two new datasets.

This consistency in performance indicates that the Chunking model can provide an accurate account of the driver's interaction with HMI tasks. Finally, the fact that the Chunking model performs better for the majority of the metrics for the real world straight road and simulator curved road data, is an indication that it might be more representative of "conservative" or safe driving, since those are conditions where the primary driving task has an increased difficulty and, as discussed in Chapters 2 and 5, drivers tend to ensure they are driving safely.

Chapter 7

Conclusion

The research presented in this thesis was conducted as part of Theme 3 of the Programme for Simulation Innovation (PSi) project, co-funded by the Engineering and Physical Sciences Research Council (EPSRC) and Jaguar Land Rover (JLR). The overarching aim of the PSi Project was to improve methods and tools towards using computer simulations as part of the automotive development process.

The specific objective of the present work was to apply the above in the context of prototype HMI evaluation, by conducting relevant experiments that would allow for comparing the driver behaviour in real and simulated driving. A further objective was to analyse the collected data and derive insights as to how drivers share their visual attention between the primary and secondary tasks. The design of the driving experiments was primarily driven by the aim to further investigate simulation capabilities in the HMI evaluation context, but also the aforementioned aim to answer questions regarding what dictates the drivers' engagement to HMI tasks. The subsequent data analyses focused on which factors affect attention sharing and in which ways, hence identifying whether any distinguishable patterns arise that could potentially help in better understanding and modelling this behaviour. Finally, two existing computational models were tested to evaluate their accuracy on predicting the observed HMI interaction data. Some additional behavioural factors were considered for the existing models and a novel model was proposed, that accounts for behavioural phenomena that have not been previously considered.

The rest of this chapter will summarise the key findings of this research, as outlined in previous parts of this thesis. It will also reflect upon methodologies and experimental design, discussing issues and limitations that arose. Finally, potential future work will be suggested before drawing a final conclusion.

7.1 Key Findings and Contribution to Knowledge

Three primary Research Questions were introduced in Section 1.2 of the introductory Chapter and, subsequently, investigated in Chapters 4, 5 and 6, respectively. The key findings answering those questions are presented in the same order as they were introduced and analysed.

7.1.1 What type of driving simulator should be used in prototype HMI evaluation related user trials?

In order to evaluate the potential of driving simulators as a tool for HMI evaluation, one needs to identify the degree of behavioural validity that can be achieved, 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 relevant results, a behavioural validity matrix was created that can provide insights on the level of behavioural validity a certain driving simulator type can achieve for different behavioural metrics. The matrix can be used by researchers and human factors specialists to either determine what type of simulator one should use for their testing or to help them interpret the reliability of the results of a driving simulator study.

In general, the level of behavioural validity that a driving simulator can achieve was found to be situation dependent, i.e. for a given behavioural metric, different simulator settings provide a different level of behavioural validity. Moreover, within a given simulator setting, the level of behavioural validity for different behavioural metrics can vary. Hence, it was concluded that there is no single, "one size fits all" solution when it comes to choosing what simulator should be used for prototype HMI evaluation experiments. 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.

The results presented here, addressed the issue of behavioural validity for a hexapod only driving simulator for the first time. Moreover, this has also been the first attempt to collectively present behavioural validity results for all relevant driving simulator types, across the most relevant metrics in the context of prototype HMI evaluation.

7.1.2 How do drivers engage with HMI tasks while driving?

Being able to predict how drivers will interact with a novel secondary HMI demands a deep understanding of the mechanisms involved in dual tasking while driving. Although there has been ample research on the effects of secondary tasks on driving performance, little focus has been placed on what drives the interactions with such tasks. The work conducted here aimed at identifying which factors affect the drivers' decision to engage with an HMI task while driving and in which way, by investigating drivers' visual time-sharing behaviour during HMI task executions.

It was found that there is a big effect of individual differences on both when and for how long drivers choose to look away from the road to perform a secondary HMI task. In particular, drivers were found to employ two distinct strategies to structure their visual time sharing towards the HMI: they would either employ more and shorter glances or fewer and longer ones. Moreover, it was found that the use of fewer glances was also associated to a reduced total time that drivers spent looking away from the road.

A novel hypothesis was proposed as to how drivers might use Time-to-Line-Crossing (TLC) to decide when and for how long to look away. An in-depth analysis revealed that glance durations away from the road are not affected by TLC per se, but revealed a strong relationship between TLC and off-road glance onset. In particular, it was found that drivers ensured they were safe enough before looking away, by doing so only when TLC was above a safe threshold.

Finally, the effect of longitudinal vehicle control on the drivers' visual attention sharing was investigated, through headway distance (HW) from the lead vehicle. In agreement with research by Tijerina et al. (2004), drivers were found to only look away when the visual angle rate of change of the lead vehicle was effectively zero. However, an additional interesting observation was made here; although the drivers tended to initiate off-road glances when visual angle rate of change approached zero, examination of the baseline data also revealed that this is where those values lied the majority of the time.

7.1.3 What types of computational models could predict the observed behaviour?

Although there is a multitude of models and frameworks that could be employed in the early stages of the HMI evaluation cycle, such tools often cannot be used due to their complexity or lack of rigorous validation. Two existing models, that are readily available and easy to use, were evaluated against data collected through the driving study presented here (previously unseen). Namely, Distract-R (Salvucci et al., 2005*a*) and the predictive equations by Large et al. (2018) were used to predict the observed behaviour. A novel model was also proposed, based on some simple equations and assumptions about how drivers engage to HMI tasks, that featured visual chunking (execution of multiple consecutive interactions in a single glance) as a behavioural phenomenon.

Overall, all models were able to predict mean off-road glance durations with low errors, while the proposed Chunking model managed to generate high accuracy predictions for all metrics (with the additional one of single off-road glance durations). Additionally, all variants of the Large et al. predictive equations and the proposed model achieved low error predictions for relative validity, i.e. when comparing the ranking of tasks against each other, for all the investigated metrics. Hence, computational models could prove valuable tools for early and quick prototype HMI evaluation, both in terms of predicting absolute values of behavioural metrics but also as tools to accurately evaluate the relative differences between alternative HMI designs.

In order to provide additional validation for the proposed model, it was also evaluated against data from the fixed base, constant speed, curved road, simulator scenario, as well as from the constant speed real world scenario. The proposed Chunking model performed better, achieving lower error, across all metrics, except for total off-road glance durations, where its performance degraded by a small margin. Interestingly, it generated the best predictions against the curved road data, indicating that it might be more tuned towards safe driving. Based on the consistently good performance of the proposed Chunking model, it was concluded that visual chunking is a behavioural pattern that should be further investigated and included in future computational models of HMI interaction.

7.2 Methodology and Design Issues and Limitations

As every piece of academic work, the present one also comes with its limitations. Since the majority of the work presented here focuses on drivers' visual behaviour, it is important to evaluate the reliability of both the raw data and the calculated metrics involved. Although significant due diligence was done to properly treat the data (please refer to Chapter 3 for more details) some inherent limitations of the equipment used, as expected from all hardware devices to a certain extent, should be reported. The SMI eye-tracker used in the real world setting automatically classifies gaze points into saccades and visual intakes, hence not logging data in a completely raw format for analysis. The Facelab eye-tracker used in the UoLDS setting does not have AOI annotation capabilities, hence mandating the definition of AOIs post-processing. Both eye-trackers also classify gaze points as "good" and "bad" in terms of quality. As is intuitively evident, this can lead to data loss due to the inability of the eye-tracker to record certain data points.

Moving forward, limitations related to the experimental choices made arise, with the first ones being issues regarding the participants; comparing driver behaviour in simulator against reality would ideally require a within-subjects design, i.e. the same participants 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 more appropriate to eliminate as much as possible the effect of individual differences, as was used, for example by Wang et al. (2010) and Klüver et al. (2016). In the experiments presented here, the sample size was relatively limited (eleven participants in the real world and twelve participants in the simulator). This decision, however, was justified by the fact that emphasis was put on obtaining a large number of repetitions per participant, hence having a strong body of data to use in model identification and development.

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 simulator (fixed base with constant speed, fixed base with varying speed, hexapod with constant speed and hexapod with varying speed), a total of 24 (i.e. 4!) participants would be needed for full counterbalancing. In this case however, only half of those were available, thus not allowing for a full counterbalancing of all the possible combinations.

Secondly, issues that could affect driving and HMI engagement performance, based on the experimental design need to be considered. As mentioned in Chapter 3, the majority of participants in the UoLDS experiment and about half of the participants in the Gaydon experiment had previous experience with the driving simulator and the physical test track, respectively. This could potentially affect their behaviour during the experiment either positively (participants being already familiar with the environment and equipment would show more consistent behaviour throughout the experiment as there would be minimal adaptation) or negatively (participants anticipating features they were previously familiar with or behaving as they would based on their past experience but instead having to perform under new conditions). However, in this case, given that their previous experience both in the simulator and the test track was limited and distant in time, it is assumed that the effect on their performance should be minimal.

The fact that not all driving conditions were tested in both the simulator and the real world setting could also potentially have an effect in the generated data. In particular, driving on curves was only tested in the simulator and not in the real world setting, something that could potentially provide a confound for the task engagement behaviour during straight sections, for the UoLDS participants. As a result of being exposed to different driving task complexities, the perception of the HMI task itself, as well as their abilities may have been different for the two sets of participants.

Looking into participants not being exposed to the same driving conditions in the two settings, it is also important to consider the lead vehicle speed profiles. Although the lead vehicle driver in the test track was provided with a detailed schematic of the speed profile they needed to follow in each drive, the actual speed of the lead vehicle was not recorded due to equipment malfunction. Consequently, the lead vehicle speed profiles could not be validated against the target ones and verify whether the two sets of participants were exposed to identical conditions. An additional potential issue, relating to the speed profile of the lead vehicle, is the fact that, in these studies the lead vehicle was not bound to the subject vehicle, i.e. it was moving independently of their distance. As a result, there was the possibility of the participants "losing" the lead vehicle from their field of view. Fortunately, this only partially happened in a few instances in the simulator, where the participants were able to recover the distance. In hindsight, not binding the lead vehicle to the subject vehicle seems to have been the right decision for this type of experiment (as it is more ecologically plausible) but maybe the speed profiles should have been treated differently to ensure that such an issue would not be prominent (e.g. by having a more sinusoidal profile than a random one).

Finally, in terms of experimental design related issues, the absence of counterbalancing for the tasks could create some ordering effects in the collected data. The fact that the baseline drive was always the first, for example, might have caused drivers to exhibit a different driving behaviour than they would have, had it been counterbalanced or just followed a drive of HMI task execution. In that case, this could potentially cause issues with the derived analyses (e.g. the driving behavioural patterns discussed in Chapter 5). However, given that all drivers, in both experiments were exposed to the exact same task ordering, any effects would be consistent throughout the collected data and, at least, alleviate any relative differences that could have arisen due to them.

Apart from the experimental design related issues described above, some limitations in data and analysis quality were also present that should be mentioned here. For the analysis presented in Chapter 4, for example, it has already been discussed how the method of calculating SDLP lacked validity. Using an approximation method such as numerical integration always affects the quality of the resulting data, as it seems to have been the case here (particularly based on the misalignment between the results obtained here and those reported in existing literature). Perhaps, given the issues in calculating it, SDLP should not have used from the experiments here but rather only from reported results in the literature, so as to not affect the quality of the behavioural validity matrix.

An additional point of consideration regarding the behavioural validity matrix, is, as already discussed in Chapter 4, the different levels of behavioural validity that can be elicited by different tasks and driving scenarios. In other words, if the same studies were conducted, with the same equipment but using a different HMI task or primary driving task, it is probable that the resulting levels of behavioural validity for the different simulator setting would be different. Hence, it is important to remember that the behavioural validity matrix comes with its limitations and is only a product of the data used to construct it, in need of timely revisions and updates.

Finally, when considering the model comparison presented in Chapter 7, although every action was taken to ensure that tasks were modelled in the same way, the different techniques used by each model did not always allow for identical replication (e.g. how the scrolling action was modelled). Although this is more of an issue related to the model capabilities rather than a limitation per se, it is still important to keep in mind when comparing model performance.

All in all, although it is important to critically reflect on such issues, in the present case it is also important to note that the experiments, analyses and simulations presented in this thesis still managed to produce a variety of meaningful novel results, as well as to align with existing work in this area of research.

7.3 Future Work Suggestions

The work presented in this thesis, apart from answering the research questions that were initially set out, has also raised some new ones that could be explored in the future. Firstly, regarding driver simulator behavioural validity, a more thorough meta-analysis of already published work would be in order (e.g. also comparing experimental scenarios and types of HMI tasks) as it could help expand the behavioural validity matrix and provide more detailed guidelines as to which simulator should be employed under which conditions. As it stands, the body of related published research is rather heterogenous, since different authors have used different experimental and analysis methodologies. Consequently, there is a need for rigorous assessment of the published results, to determine how accurately they can be used to make inferences about the behavioural validity of the driving simulator in question. A more detailed breakdown of said research, looking into different scenarios and types of secondary tasks, can shed light both on the differences arising from different modalities of each study, as well as on what might be missing from the existing research and needs to be explored further. More importantly, the above can help develop an expanded and improved version of the behavioural validity matrix, that can be used reliably by researchers and specialists trying to design a study or interpret their results. Moreover, additional metrics that have been widely used in distraction studies could potentially be considered in the context of driving simulator behavioural validity evaluation, such as reaction times under dual-tasking conditions. However, tasks like that are not easy to replicate and test in real world scenarios without safety implications. For reaction times, in particular, using a surrogate task like the peripheral detection task might not always be plausible in a real world setting due to environmental interference (e.g. glare). Creating a scenario where the driver would have to react to a near-critical event, on the other hand, such as a lead vehicle breaking, could compromise the safety of the driver. Hence, there is a need for further investigation into how such metrics could be tested and validated in different conditions.

Secondly, as in-vehicle infotainment technology advances, new types of interfaces need to be tested to ensure there is enough driver performance data for virtual methods validation. A series of further studies should be conducted to investigate how drivers interact with modern interfaces, as only a limited amount of work has been published where drivers are using contemporary, real HMIs. Moreover, there is ample room for additional analysis towards understanding which factors dictate drivers' engagement with HMI tasks, as well as in which way. The effects of TLC on off-road glance onset and duration, for example, have not been previously published. Additional follow-up analyses could be performed in that realm to validate the findings presented here and also further investigate the relationship of TLC with other metrics related to drivers' visual attention sharing behaviour.

Finally, regarding the modelling of HMI interactions, it would be important to evaluate the performance of existing models against a variety of data sets to quantify their potency in replicating observed behaviour and verify that previously noted good performance is not just the result of over fitting to training data. Results form the behavioural analyses discussed above could be used to drive the conceptualisation, implementation and improvement of such models. Moving forward, such models should also be able to predict the variability observed in human driver data, so that they can account for different driving styles, risk taking behaviours and be able to provide researchers with a more complete picture of the predicted human behaviour.

Bibliography

15007-1:2014(E), I. (2014), 'Road vehicles-measurement of driver visual behaviour with respect to transport information and control systems-part 1: Definitions and parameters'.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Trans Auto Control* 19(6), 716–723.

Allen, R. W., Klein, R. H. and Ziedman, K. (1979), 'Automobile research simulators: a review and new approaches', *Transportation Research Record* 706, 9–15.

Anderson, J. R. (1993), *Rules of The Mind*, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Anderson, J. R. and Lebiere, C. J. (2014), *The Atomic Components of Thought*, Psychology Press, Mahwah, New Jersey.

Angell, L. S., Auflick, J., Austria, P., Kochhar, D. S., Tijerina, L., Biever, W., Diptiman, T., Hogsett, J. and Kiger, S. (2006), *Driver Workload Metrics Task 2 Final Report, Technical report.*

Aust, M. L., Dombrovskis, S., Kovaceva, J., Svanberg, B. et al. (2013), 'An empirically based suggestion for reformulating the glance duration criteria in NHTSA's visual-manual interaction guidelines', *SAE International Journal of Passenger Cars-Electronic and Electrical Systems* 6, 444–453.

Baber, C. and Mellor, B. (2001), 'Using critical path analysis to model multimodal human–computer interaction', *International Journal of Human-Computer Studies* 54(4), 613–636.

Bach, K. M., Jæger, M. G., Skov, M. B. and Thomassen, N. G. (2008), '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, 193–201.

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

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

Birrell, S. A. and Fowkes, M. (2014), 'Glance behaviours when using an in-vehicle smart driving aid: A real-world, on-road driving study', *Transportation Research Part F: Traffic Psychology and Behaviour* 22, 113–125.

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

Blana, E. (1996), *Driving Simulator Validation Studies: A Literature Review.*

Boer, E. R. (2000), 'Behavioral entropy as an index of workload', in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 44, SAGE Publications, 125–128.

Boer, E. R. and Spyridakos, P. D. (2016), 'Control-theoretic attention-switching driver model'. Presented at the 6^{th} *International Conference on Traffic and Transport Psychology (ICTTP)*, Brisbane.

Broadbent, D. E. (2013), *Perception and communication*, Elsevier.

Brookhuis, K. A., de Vries, G. and De Waard, D. (1991), 'The effects of mobile telephoning on driving performance', *Accident Analysis & Prevention* 23(4), 309–316.

Broström, R., Ljung Aust, M., Wahlberg, L. and Källgren, L. (2013), 'What drives off-road glance durations during multitasking: capacity, practice or strategy?', in *3rd International conference on driver distraction and inattention*.

Brumby, D. P., Howes, A. and Salvucci, D. D. (2007), 'A cognitive constraint model of dual-task trade-offs in a highly dynamic driving task', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 233–242.

Burnett, G. E., Sharma, N., Pettitt, M. A. and Stevens, A. (2011), 'Modelling and predicting the visual demand of in-vehicleinformation systems', in *2nd International conference on driver distraction and inattention*.

Burnett, G., Irune, A. and Mowforth, A. (2007), 'Driving simulator sickness and validity: how important is it to use real car cabins?', *Advances in Transportation Studies* Spec Iss, 33–42.

Caird, J. K., Willness, C. R., Steel, P. and Scialfa, C. (2008), 'A meta-analysis of the effects of cell phones on driver performance', *Accident Analysis & Prevention* 40(4), 1282–1293.

Campbell, J. L., Carney, C. and Kantowitz, B. H. (1997), 'Design guidelines for advanced traveler information systems (atis): The user requirements analysis', in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 41, SAGE Publications Sage CA: Los Angeles, CA, 954–958.

Cao, S. and Liu, Y., (2013), 'Queueing network-adaptive control of thought rational (QN-ACTR): An integrated cognitive architecture for modelling complex cognitive and multi-task performance'. *International Journal of Human Factors Modelling and Simulation* 55, 4(1), 63-86.

Card, S. K., Moran, T. P. and Newell, A. (1980), 'The keystroke-level model for user performance time with interactive systems', *Communications of the ACM* 23(7), 396–410.

Card, S. K., Newell, A. and Moran, T. P. (1983), *The Psychology of Human-Computer Interaction*, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Charlton, S. G. (2009), 'Driving while conversing: Cell phones that distract and passengers who react', *Accident Analysis & Prevention* 41(1), 160–173.

Chiang, D. P., Brooks, A. M. and Weir, D. H. (2001), *An Experimental Study of Destination Entry With an Example Automobile Navigation System*, Technical report, SAE Technical Paper.

Classen, S., Bewernitz, M. and Shechtman, O. (2011), 'Driving simulator sickness: an evidence-based review of the literature', *American Journal of Occupational Therapy* 65(2), 179–188.

Cockburn, A., Gutwin, C. and Greenberg, S. (2007), 'A predictive model of menu performance', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 627–636.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, 2 edn, Lawrence Erlbaum Associates, New York.

Cooper, J. M., Medeiros-Ward, N. and Strayer, D. L. (2013), 'The impact of eye movements and cognitive workload on lateral position variability in driving', *Human Factors* 55(5), 1001–1014.

Cooper, J. M., Vladisavljevic, I., Medeiros-Ward, N., Martin, P. T. and Strayer, D. L. (2009), 'An investigation of driver distraction near the tipping point of traffic flow stability', *Human Factors* 51(2), 261–268.

De Winter, J., Van Leeuwen, P. and Happee, R. (2012), 'Advantages and disadvantages of driving simulators: A discussion', in *Proceedings of measuring behavior*, Vol. 2012, Citeseer, 8.

Department for Transport (2015), *Reported Road Casualties Great Britain, Annual Report: 2015.*

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M. and Hankey, J. (2016), 'Driver crash risk factors and prevalence evaluation using naturalistic driving data', in *Proceedings of the National Academy of Sciences* 113(10), 2636–2641.

Dingus, T. A., Klauer, S. G., Neale, V. L., Petersen, A., Lee, S. E., Sudweeks, J., Perez, M. A., Hankey, J., Ramsey, D., Gupta, S. et al. (2006), *The 100-car Naturalistic Driving Study, Phase II-Results of The 100-car Field Experiment*, Technical report.

Donmez, B., Boyle, L. N. and Lee, J. D. (2009), 'Differences in off-road glances: effects on young drivers performance', *Journal of Transportation Engineering* 136(5), 403–409.

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

Engström, J., Markkula, G., Victor, T. and Merat, N. (2017), 'Effects of cognitive load on driving performance: The cognitive control hypothesis', *Human Factors* 59(5), 734–764.

Engström, J., Victor, T. and Markkula, G. (2013), 'Attention selection and multitasking in everyday driving: A conceptual model' in *Driver Distraction and Inattention*, CRC Press, 27–54.

Eren, A. L., Burnett, G., Thompson, S., Harvey, C. and Skrypchuk, L. (2015), 'Identifying a set of gestures for in-car touch screens', in *Contemporary Ergonomics and Human Factors 2015: Proceedings of the International Conference on Ergonomics & Human Factors 2015*, Daventry, Northamptonshire, UK, 13-16 April 2015', CRC Press, p. 454.

Feyen, R. G. (2002), *Modeling Human Performance Using the Queuing Network-model Human Processor (QN-MHP)*, PhD thesis.

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

Fitch, G. M., Soccolich, S. A., Guo, F., McClafferty, J., Fang, Y., Olson, R. L., Perez, M. A., Hanowski, R. J., Hankey, J. M. and Dingus, T. A. (2013), *The Impact of Hand-held and Hands-free Cell Phone Use on Driving Performance and Safety-critical Event Risk*, Technical report.

Fitts, P. M. (1954), 'The information capacity of the human motor system in controlling the amplitude of movement.', *Journal of Experimental Psychology* 47(6), 381.

Funkhouser, D. and Sayer, J. (2012), 'Naturalistic census of cell phone use', in *Transportation Research Record* 2321(1), 1–6.

Galer, M. (1995), 'Driver vehicle interaction and human factors', in *J. Fagerberg, D. C. Mowery and R. R. Nelson* (eds), *Smart Vehicles*, Swets and Zeitlinger, Oxford.

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

Godley, S. T., Triggs, T. J. and Fildes, B. N. (2002), 'Driving simulator validation for speed research', *Accident Analysis & Prevention* 34(5), 589–600.

Godthelp, H. and Konings, H. (1981), 'Levels of steering control; some notes on the time-to-line crossing concept as related to driving strategy', in *Proceedings of the First European Annual Conference on Human Decision Making and Manual Control*, Delft, May 25-27, 1981.'.

Godthelp, H., Milgram, P. and Blaauw, G. J. (1984), 'The development of a time-related measure to describe driving strategy', *Human Factors: The Journal of the Human Factors and Ergonomics Society* 26(3), 257–268.

Green, P. (1999), 'The 15-second rule for driver information systems', in *Proceedings of the ITS America Ninth Annual Meeting*, Intelligent Transportation Society of America Washington, DC.

Greenberg, J., Tijerina, L., Curry, R., Artz, B., Cathey, L., Kochhar, D., Kozak, K., Blommer, M. and Grant, P. (2003), 'Driver distraction: Evaluation with event detection paradigm', *Transportation Research Record: Journal of the Transportation Research Board* (1843), 1–9.

Groeger, J. A. (2013), *Understanding Driving: Applying Cognitive Psychology to a Complex Everyday Task*, Routledge, London.

Group, D. F.-T. W. et al. (2006), 'Statement of principles, criteria and verification procedures on driver interactions with advanced in-vehicle information and communication systems', *Washington, DC: Alliance of Automobile Manufactures* .

Gu Ji, Y. and Jin, B. S. (2009), 'Development of the conceptual prototype for haptic interface on the telematics system', *International Journal of Human-Computer Interaction* 26(1), 22–52.

Haigney, D., Taylor, R. and Westerman, S. (2000), 'Concurrent mobile (cellular) phone use and driving performance: task demand characteristics and compensatory processes', *Transportation Research Part F: Traffic Psychology and Behaviour* 3(3), 113–121.

Hankey, J. M., Dingus, T. A., Hanowski, R. J., Wierwille, W. W. and Andrews, C. (2000), 'In-vehicle information systems behavioral model and design support: Final report', *US Departement of Transportation, Federal Highway administration, Final Report FHWA-RD-00-135* .

Hankey, J. M., Dingus, T. A., Hanowski, R. J., Wierwille, W. W., Monk, C. A. and Moyer, M. J. (2000), 'The development of a design evaluation tool and model of attention demand', *Paper posted on the NHTSA website (http://www-nrd. nhtsa. dot. gov/departments/nrd-13/driver-distraction/Welcome. htm)* .

Harvey, C. (2011), *Modelling and Evaluating Drivers Interactions With In-vehicle Information Systems (IVIS)*, PhD thesis, University of Southampton.

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

Hibberd, D. L., Jamson, S. L. and Carsten, O. M. (2013), 'Mitigating the effects of in-vehicle distractions through use of the psychological refractory period paradigm', *Accident Analysis & Prevention* 50, 1096–1103.

Hick, W. E. (1952), 'On the rate of gain of information', *Quarterly Journal of Experimental Psychology* 4(1), 11–26.

Hogema, J., Wentink, M. and Bertollini, G. (2012), 'Effects of yaw motion on driving behaviour, comfort and realism', in *Proceeding of the Driving Simulation Conference*, Paris, France', 149–158.

Horberry, T., Anderson, J., Regan, M. A., Triggs, T. J. and Brown, J. (2006), 'Driver distraction: The effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance', *Accident Analysis & Prevention* 38(1), 185–191.

Horrey, W. J., Wickens, C. D. and Consalus, K. P. (2006), 'Modeling drivers' visual attention allocation while interacting with in-vehicle technologies.', *Journal of Experimental Psychology: Applied* 12(2), 67.

Horrey, W. and Wickens, C. (2004), 'The impact of cell phone conversations on driving: A meta-analytic approach (technical report ahfd-04-2/gm-04-1)', *Illinois: Aviation Human Factors Division Institute of Aviation* .

Horrey, W. and Wickens, C. (2007), 'In-vehicle glance duration: distributions, tails, and model of crash risk', *Transportation Research Record: Journal of the Transportation Research Board* (2018), 22–28.

JAMA (2004), 'Guideline for in-vehicle display systems, version 3.0', *Japan Automobile Manufactures Association* .

Jamson, A. H. and Merat, N. (2005), 'Surrogate in-vehicle information systems and driver behaviour: Effects of visual and cognitive load in simulated rural driving', *Transportation Research Part F: Traffic Psychology and Behaviour* 8(2), 79–96.

Jamson, A. H., Westerman, S. J., Hockey, G. R. J. and Carsten, O. M. (2004), 'Speech-based e-mail and driver behavior: Effects of an in-vehicle message system interface', *Human Factors* 46(4), 625–639.

Jamson, H. (2000), 'Driving simulation validity: issues of field of view and resolution', in *Proceedings of the driving simulation conference*, 57–64.

Jamson, H. (2001), 'Image characteristics and their effect on driving simulator validity' in *Proceedings of the First International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 190-195.

Jamson, S. L. and Jamson, A. H. (2010), 'The validity of a low-cost simulator for the assessment of the effects of in-vehicle information systems', *Safety Science* 48(10), 1477–1483.

John, B. E. and Kieras, D. E. (1996), 'The goms family of user interface analysis techniques: Comparison and contrast', *ACM Transactions on Computer-Human Interaction (TOCHI)* 3(4), 320–351.

John, B. E., Salvucci, D. D., Centgraf, P. and Prevas, K. C. (2004), 'Integrating models and tools in the context of driving and in-vehicle devices', in *ICCM*, Citeseer, 130–135.

Kahneman, D. (1973), *Attention and effort*, Vol. 1063, Prentice-Hall Englewood Cliffs, NJ.

Kappé, B., van Erp, J. and Korteling, J. (1999), 'Effects of head-slaved and peripheral displays on lane-keeping performance and spatial orientation', *Human Factors* **41**(3), 453–466.

Kern, D. and Schmidt, A. (2009), 'Design space for driver-based automotive user interfaces', in *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, ACM, 3–10.

Kircher, K. and Ahlstrom, C. (2017), 'Minimum required attention: a human-centered approach to driver inattention', *Human Factors* 59(3), 471–484.

Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweeks, J. D., Ramsey, D. J. et al. (2006), *The Impact of Driver Inattention on Near-crash/crash Risk: An Analysis Using the 100-car Naturalistic Driving Study Data*, Report.

Klauer, S. G., Guo, F., Simons-Morton, B. G., Ouimet, M. C., Lee, S. E. and Dingus, T. A. (2014), 'Distracted driving and risk of road crashes among novice and experienced drivers', *New England Journal of Medicine* 370(1), 54–59.

Klüver, M., Herrigel, C., Heinrich, C., Schöner, H.-P. and Hecht, H. (2016), 'The behavioral validity of dual-task driving performance in fixed and moving base driving simulators', *Transportation Research Part F: Traffic Psychology and Behaviour* 37, 78–96.

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

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

Kountouriotis, G. K., Spyridakos, P., Carsten, O. M. and Merat, N. (2016), 'Identifying cognitive distraction using steering wheel reversal rates', *Accident Analysis & Prevention* 96, 39–45.

Kruskal, W. H. and Wallis, W. A. (1952), 'Use of ranks in one-criterion variance analysis', *Journal of the American Statistical Association* 47(260), 583–621.

Lam, L. T. (2002), 'Distractions and the risk of car crash injury: the effect of drivers' age', *Journal of Safety Research* 33(3), 411–419.

Lamble, D., Rajalin, S. and Summala, H. (2002), 'Mobile phone use while driving: public opinions on restrictions', *Transportation* 29(3), 223–236.

Lansdown, T. (2000), 'Driver visual allocation and the introduction of intelligent transport systems', *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering* 214(6), 645–652.

Large, D. R., Burnett, G., Crundall, E., van Loon, E., Eren, A. L. and Skrypchuk, L. (2018), 'Developing predictive equations to model the visual demand of in-vehicle touchscreen HMIs', *International Journal of Human–Computer Interaction* 34(1), 1–14.

Large, D. R., van Loon, E., Burnett, G. and Pournami, S. (2015), 'Applying NHTSA task acceptance criteria to different simulated driving scenarios', in *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, ACM, 117–124.

Lee, J. (2014), *Integrating the Saliency Map with Distract-R to Assess Driver Distraction of Vehicle Displays*, PhD thesis, The University of Wisconsin-Madison.

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

Lee, J. Y., Lee, J. and Lee, J. D. (2016), 'A visual search model for in-vehicle interface design', in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 60, SAGE Publications Sage CA: Los Angeles, CA, 1874–1878.

Li, P., Merat, N., Zheng, Z., Markkula, G., Li, Y. and Wang, Y. (2018), 'Does cognitive distraction improve or degrade lane keeping performance? analysis of time-to-line crossing safety margins', *Transportation Research Part F: Traffic Psychology and Behaviour* 57, 48–58.

Liang, Y. and Lee, J. D. (2010), 'Combining cognitive and visual distraction: Less than the sum of its parts', *Accident Analysis & Prevention* 42(3), 881–890.

Liang, Y., Lee, J. D. and Yekhshatyan, L. (2012), 'How dangerous is looking away from the road? algorithms predict crash risk from glance patterns in naturalistic driving', *Human Factors* 54(6), 1104–1116.

Lim, J. H. and Liu, Y. (2004), 'A queuing network model for visual search and menu selection', in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 48, SAGE Publications Sage CA: Los Angeles, CA, 1846–1850.

Liu, Y., Feyen, R. and Tsimhoni, O. (2006), 'Queueing network-model human processor (qn-mhp): A computational architecture for multitask performance in human-machine systems', *ACM Transactions on Computer-Human Interaction (TOCHI)* 13(1), 37–70.

Lockyer, K. G. and Gordon, J. (1991), *Critical Path Analysis and Other Project Network Techniques*, Beekman Books Incorporated.

Louw, T., Madigan, R., Carsten, O. and Merat, N. (2017), 'Were they in the loop during automated driving? links between visual attention and crash potential', *Injury Prevention* 23(4), 281–286.

Macdonald, W. A. and Hoffmann, E. R. (1980), 'Review of relationships between steering wheel reversal rate and driving task demand', *Human Factors: The Journal of the Human Factors and Ergonomics Society* 22(6), 733–739.

Maddox, M. E. and Kiefer, A. (2012), 'Looming threshold limits and their use in forensic practice', in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 56, Sage Publications Sage CA: Los Angeles, CA, 700–704.

Mammar, S., Glaser, S. and Netto, M. (2006), 'Time to line crossing for lane departure avoidance: A theoretical study and an experimental setting', *IEEE Transactions on Intelligent Transportation Systems* 7(2), 226–241.

Markkula, G. (2014), 'Modeling driver control behavior in both routine and near-accident driving', in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 58, SAGE Publications Sage CA: Los Angeles, CA, 879–883.

Markkula, G. and Engström, J. (2006), 'A steering wheel reversal rate metric for assessing effects of visual and cognitive secondary task load', in *Proceedings of the* 13^{th} *ITS World Congress*, London, 8-12'.

Medeiros-Ward, N., Cooper, J. M. and Strayer, D. L. (2014), 'Hierarchical control and driving.', *Journal of Experimental Psychology: General* 143(3), 953.

Mehler, B., Reimer, B. and Coughlin, J. F. (2012), 'Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: an on-road study across three age groups', *Human Factors* 54(3), 396–412.

Meixner, G., Häcker, C., Decker, B., Gerlach, S., Hess, A., Holl, K., Klaus, A., Lüddecke, D., Mauser, D., Orfgen, M. et al. (2017), 'Retrospective and future automotive infotainment systems – 100 years of user interface evolution', in Meixner, Gerrit, Mller, Christian (eds.), *Automotive User Interfaces*, Springer, 3–53.

Michon, J. A. (1985), 'A critical view of driver behavior models: what do we know, what should we do?', in Evans (ed.), *Human Behavior and Traffic Safety*, Springer, 485–524.

Morando, A., Victor, T. and Dozza, M. (n.d.), 'A bayesian reference model for visual time-sharing behaviour in manual and automated naturalistic driving', *IEEE Transactions on Intelligent Transportation Systems* (TBC).

Mullen, N., Charlton, J., Devlin, A. and Bedard, M. (2011), *Simulator Validity: Behaviors Observed on the Simulator and on the Road*.

NHTSA (2012), 'Visual-manual nhtsa driver distraction guidelines for in-vehicle electronic devices', *Washington, DC: National Highway Traffic Safety Administration (NHTSA), Department of Transportation (DOT)* .

of Automotive Engineers, S. (2015), 'Operational definitions of driving performance measures and statistics'.

Organization, W. H. et al. (2011), 'Mobile phone use: a growing problem of driver distraction'.

Oviedo-Trespalacios, O., Haque, M. M., King, M. and Washington, S. (2017), 'Effects of road infrastructure and traffic complexity in speed adaptation behaviour of distracted drivers', *Accident Analysis & Prevention* 101, 67–77.

Parzen, E. (1962), 'On estimation of a probability density function and mode', *The Annals of Mathematical Statistics* 33(3), 1065–1076.

Peter D, H. (1985), 'Kernel estimation of a distribution function', *Communications in Statistics-Theory and Methods* 14(3), 605–620.

Pettitt, M. A., Burnett, G. E., Bayer, S. and Stevens, A. (2006), '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, 259–266.

Pettitt, M. and Burnett, G. (2010), 'Visual demand evaluation methods for in-vehicle interfaces', *International Journal of Mobile Human Computer Interaction (IJMHCI)* 2(4), 45–57.

Rakauskas, M. E., Gugerty, L. J. and Ward, N. J. (2004), 'Effects of naturalistic cell phone conversations on driving performance', *Journal of Safety Research* 35(4), 453–464.

Ranney, T. A., Garrott, W. R. and Goodman, M. J. (2001), *NHTSA Driver Distraction Research: Past, present, and future*, Technical report, SAE Technical Paper.

Rayner, K. (1998), 'Eye movements in reading and information processing: 20 years of research.', *Psychological Bulletin* 124(3), 372.

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

Regan, M. A., Hallett, C. and Gordon, C. P. (2011), 'Driver distraction and driver inattention: Definition, relationship and taxonomy', *Accident Analysis & Prevention* 43(5), 1771–1781.

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

Reymond, G., Kemeny, A., Droulez, J. and Berthoz, A. (2001), 'Role of lateral acceleration in curve driving: Driver model and experiments on a real vehicle and a driving simulator', *Human Factors* 43(3), 483–495.

Rosey, F. and Auberlet, J.-M. (2014), 'Driving simulator configuration impacts drivers behavior and control performance: An example with studies of a rural intersection', *Transportation Research Part F: Traffic Psychology and Behaviour* 27, 99–111.

Rousseeuw, P. J. and Leroy, A. M. (2005), *Robust Regression and Outlier Detection*, Vol. 589, John Wiley & Sons.

Salvucci, D. D. (2001), 'Predicting the effects of in-car interface use on driver performance: An integrated model approach', *International Journal of Human-Computer Studies* 55(1), 85–107.

Salvucci, D. D. (2005), 'Modeling tools for predicting driver distraction', in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 49, SAGE Publications Sage CA: Los Angeles, CA, 1149–1152.

Salvucci, D. D. (2006), 'Modeling driver behavior in a cognitive architecture', *Human Factors: The Journal of the Human Factors and Ergonomics Society* 48(2), 362–380.

Salvucci, D. D. (2009), 'Rapid prototyping and evaluation of in-vehicle interfaces', *ACM Transactions on Computer-Human Interaction (TOCHI)* 16(2), 9.

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, 71–78.

Salvucci, D. D. and Gray, R. (2004), 'A two-point visual control model of steering', *Perception* 33(10), 1233–1248.

Salvucci, D. D. and Lee, F. J. (2003), 'Simple cognitive modeling in a complex cognitive architecture', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 265–272.

Salvucci, D. D. and Taatgen, N. A. (2008), 'Threaded cognition: an integrated theory of concurrent multitasking.', *Psychological Review* 115(1), 101.

Salvucci, D. D., Zuber, M., Beregovaia, E. and Markley, D. (2005*a*), 'Distract-r: Rapid prototyping and evaluation of in-vehicle interfaces', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 581–589.

Salvucci, D. D., Zuber, M., Beregovaia, E. and Markley, D. (2005*b*), 'Distract-r: Rapid prototyping and evaluation of in-vehicle interfaces', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 581–589.

Santos, J., Merat, N., Mouta, S., Brookhuis, K. and De Waard, D. (2005), '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* 8(2), 135–146.

Schindhelm, R., Gelau, C., Keinath, A., Bengler, K., Kussmann, H., Kompfner, P. and Martinetto, M. (2004), *Report on the Review of the Available Guidelines and Standards*, Technical report, Technical report, IST-1-507674-IP.

Senders, J., Kristofferson, A., Levision, W., Dietrich, C. and Ward, J. (1966), 'An investigation of driver information processing (no. 1335)', *US Department of Commerce* .

Shannon, C. E. and Weaver, W. (1949), *The mathematical theory of information*, University of Illinois Press, Illinois.

Sheridan, T. B. (2004), 'Driver distraction from a control theory perspective', *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46(4), 587–599.

Siegler, I., Reymond, G., Kemeny, A. and Berthoz, A. (2001),' Sensorimotor integration in a driving simulator: contributions of motion cueing in elementary driving tasks', in *Proceedings of driving simulation conference*, pp. 21–32.

Stanton, N. A. and Baber, C. (2008), 'Modelling of human alarm handling response times: a case study of the ladbroke grove rail accident in the UK', *Ergonomics* 51(4), 423–440.

Strayer, D. L. and Drews, F. A. (2003), 'Effects of cell phone conversations on younger and older drivers', in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 47, SAGE Publications Sage CA: Los Angeles, CA, 1860–1864.

Strayer, D. L., Drews, F. A. and Johnston, W. A. (2003), 'Cell phone-induced failures of visual attention during simulated driving.', *Journal of Experimental Psychology: Applied* 9(1), 23.

Tijerina, L., Barickman, F. and Mazzae, E. (2004), 'Driver eye glance behavior during car following', *US DOT and NTSHA, Report Number: DOT HS* 809, 723.

Tivesten, E. and Dozza, M. (2015), 'Driving context influences drivers' decision to engage in visual–manual phone tasks: Evidence from a naturalistic driving study', *Journal of Safety Research* 53, 87–96.

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

Tukey, J. W. (1977), *Exploratory Data Analysis*, Vol. 2, Reading, Mass.

Victor, T., Dozza, M., Bärgman, J., Boda, C.-N., Engström, J., Flannagan, C., Lee, J. D. and Markkula, G. (2015), *Analysis of Naturalistic Driving Study Data: Safer glances, driver inattention, and crash risk*, Technical report.

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

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

Wickens, C. D. (2002), 'Multiple resources and performance prediction', *Theoretical Issues in Ergonomics Science* 3(2), 159–177.

Wickens, C. D. and Gopher, D. (1977), 'Control theory measures of tracking as indices of attention allocation strategies', *Human Factors* 19(4), 349–365.

Wickens, C. D. and Horrey, W. J. (2008), 'Models of attention, distraction, and highway hazard avoidance', in Lee, Regan (eds.), *Driver distraction. Theory, Effects and Mitigation*, CRC Press, 249–279.

Wierwille, W. W. (1993), 'Visual and manual demands of in-car controls and displays', *Automotive Ergonomics*, 299-320.

Wilk, M. B. and Gnanadesikan, R. (1968), 'Probability plotting methods for the analysis for the analysis of data', *Biometrika* 55(1), 1–17.

Wu, C. (2007), *Queueing Network Modeling of Human Performance and Mental Workload in Perceptual-Motor Tasks*, Thesis.

Wu, C. and Liu, Y. (2007), 'Queuing network modeling of driver workload and performance', *IEEE Transactions on Intelligent Transportation Systems* 8(3), 528–537.

Wu, C., Tsimhoni, O. and Liu, Y. (2008), 'Development of an adaptive workload management system using the queueing network-model human processor (qn-mhp)', *IEEE Transactions on Intelligent Transportation Systems* 9(3), 463–475.

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

Young, K., Regan, M. and Hammer, M. (2007), 'Driver distraction: A review of the literature', *Distracted Driving* 379–405.

Zhang, H. and Smith, M. (2004), *Safety Vehicles Using Adaptive Interface Technology (task 7) a Literature Review of Visual Distraction Research*, Technical report.

A - Participant Briefing Sheets

Participant Briefing

Thank you for your response to the advertised driving simulator study. Before you take part, please read the following information and sign a consent form. Please ask the experimenter any questions that you have.

This experiment forms part of a Jaguar Land Rover and EPSRC co-funded project called **PROGRAMME FOR SIMULATION INNOVATION (PSi)**. The ultimate aim of this project is to develop the capability of driving simulators as tools for fast prototyping early systems evaluation. This study will consider a particular type of visual/manual distraction. This study will involve four simulator runs administered in one 2-hour session. **Upon completion you will be paid £15 to express our gratitude for your time.**

The Simulator

The University of Leeds Driving Simulator is a controlled and safe environment that facilitates numerous studies on driver behaviour. From outside, the motion system and the large, white projection dome are the two major components that are visible. Inside the dome, is a Jaguar S-type vehicle cab. Entry to the simulator dome is via a boarding platform and you will be accompanied into the simulator by the researcher.

Description of the experiment

In this experiment we are examining the effects of performing visual/manual tasks, using an in-vehicle interface, on driving performance.

You will be presented with 3 different tasks that will require you to interact with a touch screen. You will be thoroughly trained in performing the tasks statically and in the simulator while driving, before moving on to data collection.

The experiment will begin with a practice drive to familiarise you with the driving simulator, the road environment (a closed off, test track circuit) and the tasks that we will be asking you to complete during the study.

You will drive in four separate sessions, accompanied by the experimenter. You will complete two sessions and then you will be given a short break, and then asked to complete the second two sessions.

During each session you will periodically be asked to complete each task on several occasions. The experimenter will direct you on which task to perform and when; the experimenter will direct you to initiate the task by saying "*Engage now*" and you will have to indicate successful completion of the task by saying "*Done*".

Throughout the experiment we would like you to ensure that you are driving safely at all times. You will be following another vehicle. Please maintain a safe following distance behind this vehicle throughout the drive and please do not attempt to overtake this vehicle. Each driving session will end at the same place where it started from, with the lead vehicle displaying break lights. At this point, the experimenter will inform you that the session is completed.

Ethics, Safety and Confidentiality

It is important that you understand that we are **not** looking at your individual driving style or judging your ability as a driver. We are solely interested in the behaviour of a group of drivers to draw conclusions about drivers in general.

As with all our research, this study is subject to the strict ethical guidelines of the British Psychological Society and the requirements of the Data Protection Act. Please note that:

- At no time now, nor in the future, will any information you provide be published that allows you as an individual to be identified.
- You are free to withdraw from the study at any time without having to give any reason for your decision. Withdrawing, however will make you non-eligible for the aforementioned £15 compensation.

Finally, we would like to thank you very much for expressing an interest in this work and we hope that you will enjoy the time spent at the simulator. Your contribution towards the science of road safety is much appreciated.

Participant Briefing

Thank you for your response to the advertised driving study. Before you take part, please read the following information and sign a consent form. Note that you do not need to memorise any details; the researcher leading the experiment will repeat the instructions to you before data collection. Please ask the experimenter any questions that you may have, at any point.

This experiment forms part of a Jaguar Land Rover and EPSRC co-funded project called **PROGRAMME FOR SIMULATION INNOVATION (PSi)**. The ultimate aim of this project is to develop the capability of driving simulators as tools for fast prototyping and early systems evaluation. The collected data will be compared to data collected in an identical simulator study in order to quantify the reliability of the simulator in certain experimental conditions. This study will consider three different visual/manual distraction tasks as well as a non-visual/cognitive distraction task. The study will involve three driving sessions, with short breaks in-between, taking place in the Emissions Circuit test track in Gaydon.

Description of the experiment

In this experiment we are examining the effects of non-driving tasks, using an in-vehicle interface and a mental arithmetic task, on driving performance. You will have to drive with and without engaging in a secondary task, while following another vehicle at all times.

First, you will be presented with 3 different visual/manual tasks that will require you to interact with a touch screen. You will be thoroughly trained in performing the tasks statically and while driving, before moving on to data collection.

You will begin with a practice drive to familiarise yourself with the test track environment and the tasks that you will be asked to complete during the sessions.

You will drive in two separate sessions, accompanied by the experimenter. You will complete one session and then you will be given a short break. You will then be asked to complete the second session. After each session, you will be asked to fill in a short questionnaire.

At all times, the experimenter will direct you on which task to perform and when; the experimenter will direct you to initiate the task by saying “Engage now” and you will have to indicate successful completion of the task by saying “Done”.

For the final session, you will be presented with a non-visual, cognitive task. During this task you will hear a series of numbers through the vehicle speakers. After each number you will have to repeat the one that you heard before that (1-back). For example:

Number heard	0	5	2
Number to repeat	N/A	0	5

You will, again, begin with a practice drive before moving on to data collection.

Throughout the experiment we would like you to ensure that you are driving safely at all times. Please maintain a safe following distance from the lead vehicle, stay within your designated lane and do not attempt to overtake the lead vehicle vehicle at any point. Your final driving session will end by exiting the Emissions Circuit, when the experimenter indicates so.

Ethics, Safety and Confidentiality

It is important that you understand that we are not looking at your individual driving style or judging your ability as a driver. We are solely interested in the behaviour of a group of drivers to draw conclusions about drivers in general.

As with all our research, this study is subject to the strict ethical guidelines of the British Psychological Society and the requirements of the Data Protection Act. Please note that:

- At no time now, nor in the future, will any information you provide be published that allows you as an individual to be identified.
- You are free to withdraw from the study at any time without having to give any reason for your decision.

Finally, we would like to thank you very much for expressing an interest in this work and we hope that you will enjoy the time spent at the simulator. Your contribution towards the science of road safety is much appreciated.

B - Participant Consent Forms

UNIVERSITY OF LEEDS
DRIVING SIMULATOR

PSi HMI

Participant Consent Form

Thank you very much for agreeing to take part in this research. The purpose of this form is to make sure that you are happy to take part and that you know what is involved. Signing this form does not commit you to anything you do not wish to do.

If you suffer from any of the following medical conditions, unfortunately we will not be able to use you as a participant. Therefore, please let the experimenter know now if you suffer from:

- Fear of heights
- Epilepsy
- Serious mobility problems affecting the back, knees or hips
- Claustrophobia
- Feelings of disorientation

Please sign here if you suffer from none of the above ______________________

Have you read the participant briefing sheet?	YES	NO	
Have you had the opportunity to ask questions and discuss the study?	YES	NO	
If you have asked questions, have you had satisfactory answers?	YES	NO	N/A
Do you understand that you are free to withdraw from the study at any time and without having to give a reason for withdrawing?	YES	NO	
Do you agree to take part in the study?	YES	NO	

Name in block letters ____

Signature______________________________________ **Date** ______________________

UNIVERSITY OF LEEDS

Participant Consent Form

Thank you very much for agreeing to take part in this research. The purpose of this form is to make sure that you are happy to take part and that you know what is involved. Signing this form does not commit you to anything you do not wish to do.

If you suffer from any of the following medical conditions, unfortunately we will not be able to use you as a participant. Therefore, please let the experimenter know now if you suffer from:

- Epilepsy or other similar nervous system disorders
- Serious mobility problems affecting the back, knees or hips
- Claustrophobia
- Feelings of disorientation

Please sign here if you suffer from none of the above ______________________

Have you read the participant briefing sheet?	YES	NO	
Have you had the opportunity to ask questions and discuss the study?	YES	NO	
If you have asked questions, have you had satisfactory answers?	YES	NO	N/A
Do you understand that you are free to withdraw from the study at any time and without having to give a reason for withdrawing?	YES	NO	
Do you agree to take part in the study?	YES	NO	

Name in block letters ____

Signature______________________________________ **Date** ______________________

C - Subjective Questionnaires

NAME:

DATE:

How easy was the task to complete?

<u>Task 1</u>

Very Easy Very Difficult

<u>Task 2</u>

Very Easy Very Difficult

<u>Task 3</u>

Very Easy Very Difficult

How easy was the task to perform while driving?

<u>No Task</u>

Very Easy Very Difficult

<u>Task 1</u>

Very Easy Very Difficult

<u>Task 2</u>

Very Easy Very Difficult

<u>Task 3</u>

Very Easy Very Difficult

How acceptable was it to perform the task while driving?

Task 1

Not acceptable Very acceptable

Task 2

Not acceptable Very acceptable

Task 3

Not acceptable Very acceptable

How did you find the frequency of task execution?

Task 1

Not frequent Very frequent

Task 2

Not frequent Very frequent

Task 3

Not frequent Very frequent

How realistic was the overall driving experience (e.g. simulator graphics, simulator motion, etc.)?

Not realistic Very realistic

How realistic were the vehicle controls (steering wheel, pedals, etc.)?

Not realistic Very realistic

Did you feel uneasy at all during the drive (e.g. dizzy, disoriented, nauseated, etc.)?

Not uneasy Very uneasy

D - Counterbalancing

UoLDS Experiment

Table 7.1 illustrates the condition counterbalancing in the UoLDS experiment. The conditions, labelled with numbers $1-4$, represent a combination of simulator motion setting and driving scenario, as noted below:

- $1 \rightarrow$ Fixed base and constant speed.
- $2 \rightarrow$ Fixed base and varying speed.
- $3 \rightarrow$ Hexapod and constant speed.
- $4 \rightarrow$ Hexapod and varying speed.

Table 7.1: Counterbalancing in the UoLDS experiment.

Participant	**Drive 1**	**Drive 2**	**Drive 3**	**Drive 4**
1	4	2	3	1
2	4	3	2	1
3	4	1	3	2
4	2	4	3	1
5	2	3	4	1
6	2	1	3	4
7	3	2	4	1
8	3	4	2	1
9	3	1	4	2
10	1	2	3	4
11	1	3	2	4
12	1	4	3	2
13	4	1	3	2

In the above, Participant 3 was the one to drop out due to simulator sickness, being replaced by Participant 13.

Gaydon Experiment

Table 7.2 illustrates the condition counterbalancing in the Gaydon experiment. The conditions, labelled with numbers $1 - 2$, represent the driving scenario, as noted below:

- $1 \rightarrow$ Constant speed.
- $2 \rightarrow$ Varying speed.

Table 7.2: Counterbalancing

Participant	Drive 1	Drive 2
1	1	2
2	1	2
3	2	1
4	2	1
5	1	2
6	1	2
7	2	1
8	2	1
9	1	2
10	2	1
11	1	2