Evaluation of Control Modalities in Highly Automated Vehicles

by

Chongren Sun

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2023

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contribution

The contents are collaborative effort with two supervisors of first author: Chongren Sun. They are:

- Dr. Siby Samuel
- Dr. Oliver Schneider

Abstract

Autonomous driving vehicles are classified by researchers into six levels, ranging from 0 to 5. The level of autonomy increases with the level number, with vehicles at levels 4 or 5 possessing the capability for full self-driving without human intervention. In high-level autonomous driving, user control model can be adapted to meet user demands since drivers are not required to focus on the road. Thus, measuring the metrics and trade-offs of control modalities under this new driving paradigm is crucial. This study proposes an evaluation framework for control modalities in level 4 and 5 autonomous vehicles, particularly in distraction scenarios.

The research comprises two parts. The first part is a user requirement study. A questionnaire, which surveyed 150 participants, investigated potential control modalities and features in self-driving vehicles. Following this, a user study that incorporated both between-participant and within-participant designs was conducted. The between-participant design aimed to compare three control modalities: physical buttons, voice, and hand gesture. Additionally, a within-participant design tested each participant's performance while being distracted. The study collected both objective and subjective data, including user error rates, physiological data, the NASA TLX rating scale, and interview feedback.

The evaluation revealed that the hand gesture control modality yielded the lowest user performance without distractions and was least affected by distractions compared to the other models. Users who engaged with the voice control modality experienced a lower error rate and workload but were also more susceptible to distractions.

Acknowledgements

I would like to thank my superviors, Dr. Siby Samuel and Dr. Oliver Schnieder, for funding this research. You provided valuable guidance of this research and feedback for every revisions of this thesis. You answered my questions perfectly and helped me solve problems in research immediately. It was a smooth, wonderful research experience with you.

I also want to thank Dr. Shi Cao, and Dr. Jian Zhao. Both of them are professional experts in this field and served as my thesis committee.

I would like to thank my parents for supporting me though my years of study and academic research.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACC	Acceleration 45
ADAS	Advanced driver-assistance system 8, 13, 16
ADHD	Attention Deficit Hyperactivity Disorder 72
BPM	Beats per minute 46
BVP	Blood Volume Pulse 35, 45, 46
DALI	The driving activity load Index 13
EDA	Electrodermal Activity 35, 45
GSR	Galvanic skin response sensor 42
HMD	Head-Mounted Display 16
HR	Heart Rate 35, 45, 46
HVI	Human-Vehicle Interaction 13, 17
IVIS	Information and Communication Systems 13
MWL	Mental Workload 12, 13
SCL	Skin Conductance Level 46
SCR	Skin Conductance Response 46
TEMP	Temperature 45
TLX	Task Load Index 36, 37

UER User Error Rate 44

VRDS Virtual Reality Driving Simulation 16

Chapter 1

Introduction

1.1 Motivation

The vehicle industry is growing rapidly. More and more advanced features such as Lane Keeping Assist Systems (LKAS), Head-up display (HUD) are added into people's daily driving vehicles [1]. With these features as early prototype of self-driving features, self-driving started to be a popular topic in both industry and academia. The new capacities of self-driving cars lead to profound impacts of society. It is important to address autonomous driving technology related concerns with more research.

The classification and categorization of different autonomous driving styles is crucial for a better understanding of autonomous vehicles. In 2016, the SAE International committee (SAE J2016) published a new taxonomy and definitions for Driving Automation Systems, which established six levels of vehicle driving systems [2]. These levels range from 0 to 5, with Level 0 indicating the absence of driving automation and Level 5 representing full automation. The automation in the driving systems between levels 1 to 3 is defined as requiring varying degrees of human involvement for self-driving. Conversely, driving automation systems with levels greater than 3 (levels 4 and 5) are deemed autonomous systems that do not necessitate human involvement. In the instance of full autonomy, drivers are not required to participate in any driving functions, although they can still communicate with other features of the vehicle [2].

With driving, the methods of interaction are traditionally limited due to the need for drivers to concentrate on the road and avoid distractions. However, in the context of fully autonomous driving, drivers are relieved of this burden and can interact with the self-driving vehicle in a different manner than in conditional autonomous driving vehicles. Currently, the predominant forms of interaction in vehicles on the market are physical buttons and touch screen controls [3]. Despite some companies exploring alternative control modalities such as mid-air hand gesture control and voice control, these are still restricted to a limited number of features [3]. Despite the current standard of vehicle automation being at autonomy level 2, the advancement towards levels 3, 4, and eventually level 5 continues to progress steadily. Hence, the examination of potential control modalities in fully autonomous driving vehicles is both important and necessary.

In the field of driving research, two commonly utilized research methods include real-road experiment, and Simulation [4]. Due to the limitations of real-world testing, such as safety and cost, simulation has become a preferred choice for researchers to test their work [4]. There are various types of driving research simulation methods, including Hardware-in-the-loop, Software-in-the-loop, and Human-in-the-Loop [5]. Hardware-in-the-Loop simulation involves the use of real-world hardware components, such as sensors and actuators, connected to a computer simulation of the vehicle and environment. This type of simulation is useful for evaluating the real-world performance of control algorithms and sensor systems [5]. Software-in-the-Loop simulation, on the other hand, employs a computer simulation of the vehicle and environment, with both the control algorithms and sensor systems being simulated. It is useful for testing the software components of a driving system without the need for real-world hardware [6]. Human-in-the-Loop simulation, on the other hand, is focusing on Human's performance in the simulation task. Human-in-the-loop simulation represents an integral part of developing and testing driving automation systems, providing a controlled environment in which various driving scenarios can be systematically studied and evaluated. In the realm of driving simulations, the human (typically a driver) interacts with the simulation environment, providing crucial inputs to the system and responding to the system outputs as they would in a real-world driving scenario.

This study conducted an user study for researching fully self-driving control modalities with Virtual Reality simulation. As a method of Human-in-the-Loop design, it utilizes virtual reality technology to create a realistic simulation of the vehicle and environment, and is useful for examining human factors, such as driver behavior and perception, in a simulated driving scenario [7].

In the field of driving research simulation, the scenario customization and data collection capabilities provided by driving simulators could offer researchers significant flexibility for conducting research. Over the past decade, there has been a rapid progression in the field of Virtual Reality, which has proven to be a valuable tool for various academic research fields. The utility of a Virtual Reality driving simulator has been demonstrated in the literature [8]. The immersive experience and multiple interaction methods offered by Virtual Reality provide great advantages for driving research. In this study, the researcher is using Virtual Reality driving simulator to provider better simulation environment. It is further discussed in other sections.

1.2 Research Questions

In this study, the main research questions are:

1. What are the features that users want in a fully self-driving vehicle?

2. Which control modalities are possible to appear in fully autonomous driving vehicles?

3. How is users' performance and preference of each tasks with different interaction methods and why they perform well or bad with certain tasks?

1.3 Contributions

The findings in this paper are applicable to future research in autonomous driving, input control modalities, Virtual Reality driving simulators. This paper could provide an insight of a general understanding of combination of these fields to future researcher. And also, data collected and analyzed could help future application development in industry. The more specified contributions are listed below.

1. This study built a evaluation system of input control modalities in self-driving vehicle.

2. This study have built a Virtual Reality self-driving interaction simulator

3. This study found control models are easily affected by distraction, especially with the same modality.

1.4 Thesis Organization

In this thesis, the content is organized into five chapters in accordance with the following structure.

Chapter 2 presents a comprehensive literature review that encompasses relevant literature in the fields of autonomous driving, control modalities, and driving simulators. The first four sections of this chapter provide an elaboration of terms that were introduced in the introduction, including Driving Automation Systems Levels, categorization of control modalities, driving research simulators, driving tasks, and distractions. The fifth section of the chapter conducts a general review of the literature based on the conjunction of these topics, examining the relationship between them and presenting the research questions posed by previous authors. A detailed discussion of the literature review is provided in both this chapter and Chapter 5: Discussion, with a final conclusion and guidance for the reader to the next chapter.

Chapter 3 outlines the study methodologies and procedures of the experiment. The design of the entire research is explained in detail in this chapter, with a section devoted to the hypothesis of the study, the expected outcome, and a thorough examination of the considerations taken in each step of the experiment. Upon completion of this chapter, the reader should have a thorough understanding of the author's research design. A conclusion and summary of the chapter are also included.

Chapter 4 presents the results of the study. The author reports on multiple datasets collected from in-lab user tests, utilizing methods such as interviews, questionnaires, rating scales, heart rate (HR), blood volume pulse (BVP), and electrodermal activity (EDA). These datasets are analyzed through both subjective and objective methods, and the author reflects on the lessons learned from the data analysis.

Chapter 5 engages in a discussion of the implications of the results for control modalities in level-4 autonomous driving vehicle simulations within a virtual reality driving simulator. The author compares the hypothesis with the insights gained from the data analysis.

Chapter 6 concludes the thesis and the research as a whole, providing a general conclusion and discussing limitations and future work. This chapter serves as the main conclusion of the thesis.

Chapter 2

Background

To provide a general insight for readers of this thesis, this report provided readers with an overview of previous work related to self-driving vehicles, driving simulators, control modalities in driving, self-driving vehicle features, and user performance with different interactions in autonomous driving. All of these topics are essential for the study of this thesis. And also, it is important to review the past work of other researchers. Thus, we could better understand the general context of each topic and identify the need for additional research and justify our research. By reading through these chapters, readers could gain a better understanding of the research questions.

This chapter was divided into five main sections, with multiple subsections. These sections will further discuss each topic. In section 2.1, this report discussed the definition of automation driving systems and also review the current self-driving technologies that appear in both industry and academia. In section 2.2, this report discussed the common control modalities in both people's real life and also in the ergonomics research field. And especially, how they perform and are evaluated in the driving section. In section 2.3, this report discussed tasks in self-driving, which include existing self-driving interaction tasks and potential tasks in the future. In section 2.4, this report discussed the mental workload (or cognitive workload) of different control modalities while doing different types of tasks. In section 2.5, this report discussed the previous study which related to the combination of some of these topics and how other researchers did their study. At the last, in the section. And, A general conclusion of this chapter is provided.

2.1 Levels of Driving Automation

The application of AI (Artificial Intelligence) in the Automotive industry has led to a huge technological evolution [9]. More and more advanced features are adapted to automotive products which helps people a lot in their daily driving. To regularize the trend of automation in the vehicle industry, the major factor that helps people determine the automation level of AV is the percentage of human involvement in the driving task. The International Society of Automotive Engineers has published six levels of vehicle automation (from 0 to 5) [2]. This scale is the most widely accepted scale in both academia and industry. Following the improvement of Advanced driver-assistance system (ADAS), the level of vehicle automation increases from no automation (Level 0) to full automation (level 5). Between level 0 and level 5 five, level 2∼4 is defined as conditional autonomous driving. The human driver is responsible to monitor the driving environment along with ADAS's support in level 2. In level 3, the driving is called conditional driving which the human driver is responsible to take over the vehicle in certain circumstances. And, in level 4, the vehicle is capable to drive fully automatically. But the driving system (gas pedal, steering wheel, etc.) is still built in this level's vehicle. Compared with level 4, which is also defined as "fully self-driving", level 5 autonomous driving is a vehicle that can be fully automated with no human driving system. The chart with more specified details is attached below [2].

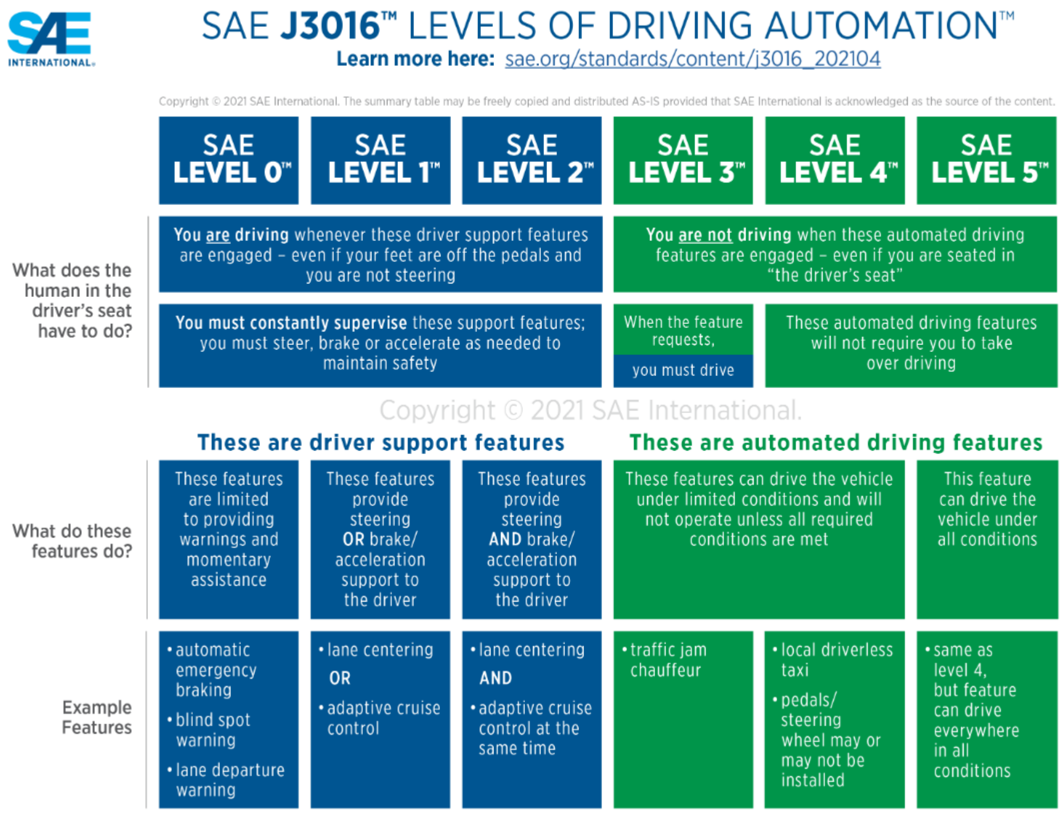


Figure 2.1: SAE Levels of driving automation: The SAE Levels 0 to 5 describe driving automation, with 0 being no automation and 5 being full automation without human intervention. As levels increase, vehicles assume more driving responsibilities.

Despite the introduction of numerous technological advancements in the automotive industry, the development of fully autonomous vehicles remains an elusive goal, facing numerous technological and ethical challenges. These challenges encompass technological difficulties, ethical and moral considerations, political and legal obstacles, and safety hazards [10]. Nonetheless, the significance of these challenges has made the pursuit of level 5 autonomous driving a highly-researched topic, attracting the attention of numerous research labs, scientists, and scholars who are working towards resolving the aforementioned challenges. For instance, researchers are addressing technological challenges through the implementation of faster data transmission technologies, such as 5G, and the training of more sophisticated machine learning models to enhance the decision-making processes [10].

2.2 Human-Vehicle Interaction

With new generation of automotive vehicles being released with more and more complex technologies, the functionalities of a vehicle are also changing. The vehicle has become a place for communications, media consumption, and infotainment center, as driving is already a complicated task that requires a varying level of cognitive and physical load. Correspondingly, the user interaction inside the vehicle has become overcrowded. As a human, how to interact with these features is an important factor to achieving a comfortable, satisfying ride. For the driver to use the complicated features in a vehicle in the most comfortable, intuitive way, HVI should be considered as the most important key in ergonomics study [11]. There are a lot of previous studies conducted to categorize HVI models and its related subjects. For example, brain vehicle interface [12], personalized touch interface [13], etc. When considering the design of Driver-vehicle interfaces, the main goal of engineers is to manage different interactions between the driver and in-vehicle systems in the following disciplines refer to: avoid interference between different information, avoid negative impact from information sources such as information overload, distraction [11]. And also, to manage the functions, these factors should be considered: decision on which type of information should be delivered, driver's adaptation, and personalization for the individual driver.

More particularly, some researchers divided human-Vehicle Interaction into three groups: Driver-centric HVI, User-centric HVI, and Customer-centric HVI [14]. For each kind of focus group, researchers studied different kinds of features. For instance: they mainly studied Mobility, Safety, and control of Driver-centric HVI, since these are the factors that most important from a driver perspective. For the User-centric HVI, the three factors are: control interaction method, feedback, and Customization. And for the Customer-Centric HVI, these are: infotainment, Telematics, and connectivity [14].

In the realm of fully autonomous vehicles (Level 4/5), User-centric Human-Vehicle Interaction (HVI) models are of utmost importance. In such a scenario, drivers are relieved of the responsibility of driving tasks and can instead be considered as users or passengers. Consequently, this study will primarily focus on examining the impact of various individual interaction methods on users. The field of in-vehicle interactions encompasses a diverse range of methods, including physical buttons, gestures, auditory cues, tactile/haptic feedback, and wearable sensors [15]. While these interaction methods have been previously evaluated by researchers, their performance in the context of fully autonomous vehicles remains an under-explored area.

2.3 Distraction and Mental workload in HVI

Task distraction in ergonomics refers to the interruption or deviation from an individual's primary task, which can lead to decreased productivity and increased risk of injury or discomfort. Task distraction can occur due to external factors such as noise, movement, or visual stimuli, or internal factors such as fatigue or boredom [16]. In Driving tasks, distraction refers to any activity or behavior that diverts the driver's attention from the task of driving. This can include activities such as using a mobile phone, eating, drinking, grooming, or adjusting the car's controls. Driving distraction is a serious issue, as it can significantly reduce a driver's ability to react to road conditions and increases the risk of crashes, injury or death [16].

In other words, when designing a user interface in a vehicle, engineers have always to consider that driving likely involves the processing of various external cues including visual and auditory. Besides that, User performance and safety are critically dependent on internal and external environmental factors, such as visibility, wind speed, crowd density, and precipitation [17, 18]. It is no doubly that environmental factors can greatly impact driver in a manual driving vehicle or partial self-driving vehicle. For example, poor visibility in outside environment caused by fog, rainfall may impair driver's hazard avoidance ability, which could lead to more sudden braking and increased accident rate [17].

However, when scenarios transfer to fully self-driving, the explanation is different. Since the driver (or passenger) do not have to deal with driving distraction and interruption, they are more tend to focus on their tasks. Their task distractions, in this case, is the interaction with the self-driving vehicle. Different control modalities in a self-driving vehicle could bring different level of Mental Workload (MWL) to driver/passengers on their own ongoing tasks.

It is common that human has a limitation in the ability of processing information. Usually, information overload could lead to poor performance. The workload is defined as a mental construct that could reflect the interaction of mental demands imposed on users by tasks they are doing. Specifically, MWL refers to the mental effort (or amount of mental resources) needed by the user to perform a set of tasks.

The examination of MWL is very important for Ergonomics, Human factors research.

One way researcher study it is by objectively investigating the task itself, which is also referred as task load. For instance, driving in a automatic transmission demands a lower taskload than driving in a vehicle with manual transmission [19]. Because user have to pay attention to more features (clutch, shifting gears) which increases user's mental workload [19]. On the other hand, another way to study MWL is by investigating workload through subjective experience of users, typically through interviews, surveys, or questionnaires. It is well established that different people may have different capabilities to handle tasks. So subjective measurement of mental workload can effectively calculate workload efficiency of single user as well [20].

In Human-Vehicle Interaction (HVI), Productive interaction between humans and vehicles needs that a user must effectively manage his/her attention of the features that are competing for it. An improper interaction method may reduce a user's performance, and emotional state, and increase mental workload [21]. In the driving tasks, driver's MWL is affected my multiple factors, for example, complex driving tasks, interaction with infotainment systems, surrounding environment, and distractions, etc. Paizié in 2008, has build MWL in driving evaluation framework, The driving activity load Index (DALI), to help researchers to evaluate driver's tasks. It shows that the evolving advanced in-vehicle assistant systems has greatly enhanced in-road safety, mobility, and reduced driver's mental workload, such as Information and Communication Systems (IVIS), ADAS [22]. However, in fully self-driving vehicle, the mental workload is still an important factor when considering design interactions, research user preference. Since user in autonomous driving vehicle, even released from complex driving tasks, still need to spend mental resources on different tasks.

2.4 Wizard-of-Oz Experiment

The Wizard-of-Oz (WoZ) method is a research technique used in the field of Human Factors, Human-Computer Interaction, and other disciplines, that involves simulating an automated system with a human 'wizard' controlling its operations behind the scenes. This method enables researchers to examine how users interact with systems that are perceived to be autonomous, even when the autonomous capabilities of these systems are under development or non-existent [23].

In a typical WoZ study, participants interact with a system they believe to be fully autonomous. However, their interactions are actually facilitated by a hidden human operator, the 'wizard', who responds to user controls in real-time. In the realm of autonomous driving study, Woz often represent a human driver driving an "autonomous car", to mimic a real self-driving scenario [24].

The advantage of the WoZ method is that it allows for the collection of authentic user data and reactions to a proposed system or feature before fully developing it. This can offer invaluable insights for the design and development process.

2.5 Driving Simulators

To study driving ergonomics topics, driving simulator is a very important tool. Driving simulators allow researchers to investigate complex driver/user behaviours in an environment that easy to control and safe for users [25]. In 1930s, driving simulators firstly appeared in research. The early function of a driving simulators are: technology effects, road infrastructure, in-vehicle systems [26].

However, the driving simulators also has its side effect. For example, an issue with a laboratory-based simulator is the reliability and validity when it is comparing with real road driving research. Reliability is the ability of a simulator to deliver consistent data/results. In the driving research, to boost the reliability of a simulator, researchers always ask participants to compete the driving task multiple times, with analyses comparing machine performance across time [27]. On the other hand, validity means the ability of a simulator that representing real world driving. A low validity refers to a certain kind of driving simulator perform poorly in represent real world driving. And also, validity could be in different forms, such as absolute validity and relative validity [28].

In last 40 years, graphics technology, advanced computer processing, and more accurate controlled equipment have boosted the both the validity and reliability of driving simulators greatly [29]. Most simulators are dynamic and the driving environments is more adaptive which lead to better drivers preference. The key point in a driving simulator is the visual quality of its display. A poor visual quality could lead to high rate of Simulator Sickness, higher mental and physical workload, and fatigue. On the other hand, the physical equipment improvement also lead to a more realistic environment for driving. For example, in the past, most basic simulator are on a fixed seating with a limited movement steering wheel and gas pedal. Recently, driving simulator with incorporating with partial vehicle body or flexible movement seating, steering wheel, or gas pedal [30].

Moreover, simulator fidelity is another essential factor when considering driving simulators. Fidelity means the ability of equipment (simulators) to appear real world scene. Technological advances have lead to increase of different types of driving simulator. however, the fidelity performance of each simulators could be greatly different [30]. Based on research by Kaptein [30], a simulator with good performance should has a large field of view, full-feedback interaction, and flexibility in control. A high-level driving simulator should be able to provide matching physical realism.

With the development of Virtual Reality in the past 20 years, researchers has started using Virtual Reality headset as the driving simulator, to provide high fidelity, high immersion driving simulation. Vitual Reality, on the other hand, could be called as 3D multi-sensory highly interactive artificial environment [31]. For example, the evaluation of a new ADAS could be benefited by using Virtual Reality as it providing highly immersive driving simulation. It is very common that most of the automotive manufacturers are using VR in different phases in their product development [32]. VR Driving simulation with a Head-Mounted Display (HMD), for example HTC Vive or Oculus Rift/Quest, could provide another perspective to conventional simulation in the design of automotive products. Moreover, Virtual Reality driving simulator also have another advantage, which is fast prototyping, could help researchers/manufacturers to develop their driving model. Fast prototyping means the low time-consumption on prototyping and adjustment of the driving model, once researchers found that they need new virtual prototype or adjustment on the driving model, or driving scenario, they could make the change or deploy the implementation with the lowest cost in time. To use VR as driving simulator, a control loop with programmed applications need to be integrated in the device. Generally, just like other driving simulators, input hardware should always be integrated as well, such as steering wheel, gas pedal.

In other words, Virtual Reality also has a lot of limitations when conducting driving simulation. Some typical limitations related with Virtual Reality Driving Simulation (VRDS) are: inconsistency of speed in virtual environment and the real world, lack of haptic feedback on other part of user's body, and shorter duration that people can comfortably play it [33]. Just like advantages of VRDS, all of these limitations are also needed to be considered in implementation of driving models.

2.6 Previous Works

There are several previous works did by other researchers. In 2016, Debernard et al. have investigated in HVI Cognitive work analysis, and also cooperation and transparency in Autonomous vehicle [34]. The authors presented a novel methodology for evaluate the transparency of an interface system [34]. However, in this research, only display and visual control models were evaluated. Multi-model evaluations in Autonomous vehicle HVI is still being unexplored by this work. Another work, by Udara, et al, investigated how could Hand-gesture interface help driver in autonomous Vehicle feeling a more realistic interaction with a semi-autonomous vehicle [35]. Udara used several Hand-gesture patterns to control certain features of a self-driving vehicle. They have concluded that hand-gesture could bring higher flexibility in interaction with a semi-autonomous Vehicle [35]. However, there is still a gap since control modalities were not fully investigated in a full autonomous driving condition. And also, comparison of different interaction modilities should also be evaluated. In 2019, Seul et al have studied three different driving support agents in Autonomous driving vehicles [36]. The driving support agents are robotics based with different interaction modalities included. However, the driving support agents are mainly voice based (informative agent, conversational agent) [36]. More study on different control modalities could be investigated based on this work.

Chapter 3

Methodology

This study constructed an experimental design that was based on the research questions. To ensure comprehensive consideration of all aspects, two studies were conducted: User Requirement Study and User performance and Preference Study. In the User Requirement Study, a User Requirement Questionnaire following with data analysis was conducted to set the scope for the second study. For the User performance and preference study, multiple experiment models were employed. The experiment methodologies included: Wizard of Oz design, mixed factorial design with within subject design and between subject design. This chapter provides a thorough explanation of the design of the user experiment, along with the logical steps involved. Furthermore, this report addressed potential problems or limitations of the user experiment and delve into the environment in which the experiment was conducted, which encompasses the laboratory setup and the selection of equipment. In addition, this report discussed participant recruitment, both in terms of the pre-study questionnaire and the in-lab user study.

3.1 User requirement study

Prior to the initiation of the user experiment, a user requirements survey was conducted. The goal of this research was to investigate the real and potential features, along with control models, that may occur in a high-level (4/5) autonomous vehicle, as most people lack insight into full autonomy. The requirement analysis assisted in defining the research questions and limiting the scope of the study. The user requirements survey provided an understanding of the features that individuals most desire in a fully self-driving vehicle. With the analysis of the user requirements, the simulator experiment scenario and tasks related to self-driving were developed. The objectives of collecting user requirement data in this research were twofold: firstly, to determine which control models users find most intuitive or appealing in a self-driving vehicle, and secondly, to identify the tasks or features that individuals expect a self-driving car to possess in the future.

The participants ($N = 150$) for the user requirement survey were recruited from an online platform. The survey consisted of three open-ended questions, namely: "What features do you think will be present in a self-driving vehicle?", "Which type of control method do you think is the most intuitive in a self-driving vehicle?" and "Do you think different control models have different performance for different features in a self-driving vehicle? If so, please rate the performance of the control models in conjunction with each feature." To familiarize the participants with the concept of self-driving, brief introductions of each controls models are provided to participants. The descriptions aimed to provide an understanding of what self-driving is and the various control models and features that are present in a fully self-driving vehicle. The survey took approximately 5 minutes to complete.

3.2 User Performance Study Design

Following along with the research questions, we have to develop the experiment by these rules:

- The experiment should based on user's needs
- The experiment should minimize learning effect
- The experiment should minimize user's fatigue
- Reduce the random noise as much as possible when comparing control groups

In this study, the between-subject approach was selected as the primary framework for the user experiment. Three groups of participants were divided to evaluate three control models, which included physical buttons, voice control, and hand-gesture control. In this experiment, each group of participants was only exposed to one condition. The utilization of a between-subject experiment presents several advantages. Firstly, it minimizes the learning effect among participants, as participants would be able to apply their prior knowledge of one control model to another if they are involved into multiple control models. For instance, a participant who has completed tasks with voice control control may perform better with other control models due to an increased level of knowledge. Secondly, the between-subject design reduces the impact of fatigue. In the user experiment, each participant would interact with each task multiple times, which could cause fatigue and potentially affect the outcome of the experiment.

In addition to the primary process of the experiment, a within-subject design was also incorporated into the research. For each group of participants participating in the between-subject comparisons, they would also be subjected to an additional task aimed at measuring the effect of distraction from other tasks. The utilization of a within-subject design in the main study serves to minimize random noise from participants. This is because individual participants bring their own background knowledge, context, and physical and mental reaction speeds to the experiment, and may also have different emotional or physical states prior to participating.

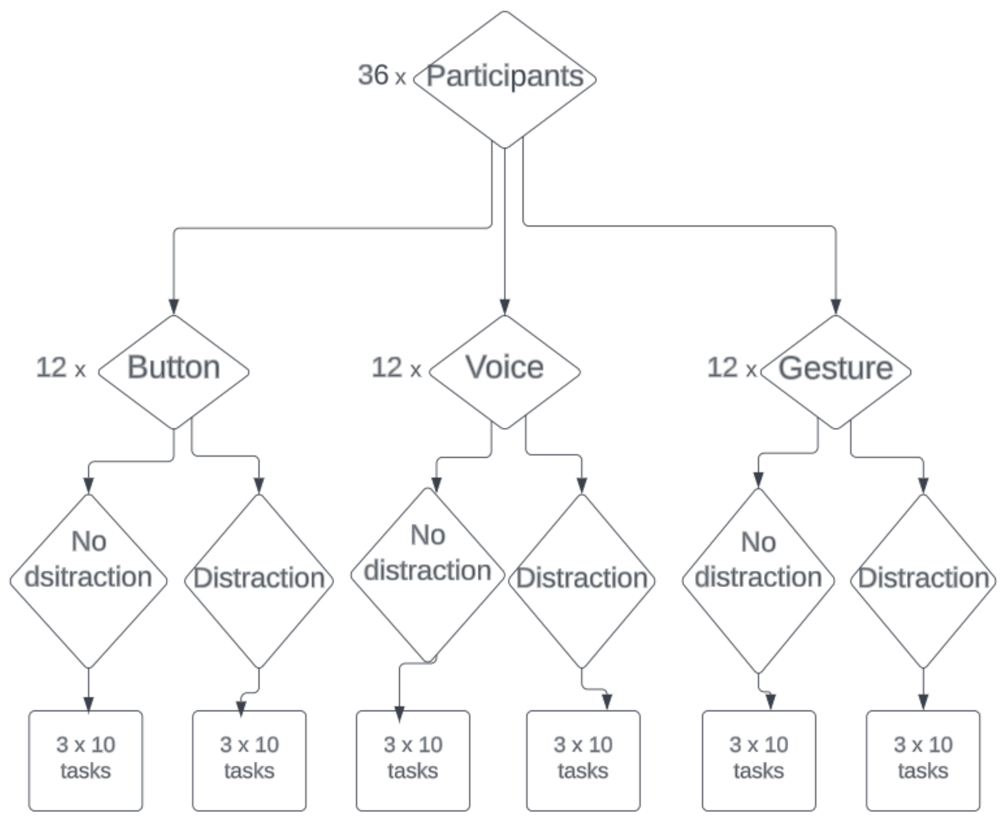


Figure 3.1: Design Pattern of User experiment, three groups of participants (each 12) using three types of control models

3.2.1 Control Models

Based on the results of the user requirement survey, three control models deemed most intuitive for a fully self-driving vehicle by participants were physical buttons (on the steering wheel), voice control, and hand gesture control. The survey also revealed that touch screen and mind control (brain-vehicle interaction) are other potential control methods for a self-driving vehicle in the future, but due to a low rate of selection by the participants, they were not included in this study. This result would be further analyzed in the following chapter.

The driving simulation study was developed using the Unity platform and integrated with Oculus Integration.[37] The initial implementation of the scenario involved a city driving scenario. Subsequently, the results from the user requirements survey were incorporated to implement each of the selected control models.

The first control model employed traditional physical buttons. Utilizing the Logitech G29 steering wheel's connection API, users could interact with the vehicle through these buttons. In the virtual simulation environment, the steering wheel was mirrored to align the user's hand position with that of the physical steering wheel.

The second control model was a voice control interface, which was implemented using the Wizard-of-Oz design method due to the potential lag and uncertainty associated with machine errors in voice recognition. The Wizard-of-Oz design reduced random noise in voice recognition, allowing participants to simply utter commands and operate the interface.

The third control model was hand gesture control, for which we also employed the Wizard-of-Oz design method after testing the built-in hand control API in Oculus Integration. The use of the Wizard-of-Oz method for hand gesture control provided the advantage of obtaining real-time reaction times in conjunction with heart rate, as the built-in function experienced a lag of 0.5 to 1 second.

3.2.2 User's tasks

Following the collection of data from the user requirement survey, we selected the five most frequently requested features for a self-driving vehicle, they are: open/close the window, turn on/off the music, adjust the passenger seat, open/close the map, and turn on/off the ambient light. Other tasks that appeared in the survey responses, such as adjusting vehicle speed, air conditioning, answering phone calls, and activating massage functions, were not included in this research due to their limited popularity and technical difficulty of implementation.

Figure 3.2: Physical button model: each yellow block represents one type of feature

Figure 3.3: Steering wheel in Virtual environment (white blocks for better participants' button recognition). The position is aligned with the real steering wheel

Light	Map	Adjust Seat	Music	Window
"Turn on the light" / "Turn off the light"	"Open the map" / "Close the map"	"Adjust seat"	"Play the music" / "Stop play the music"	"Open the window" / "Close the window"

Figure 3.4: Voice command model: first command as the opening command and second command as the closing command, Adjust seat feature use only one command

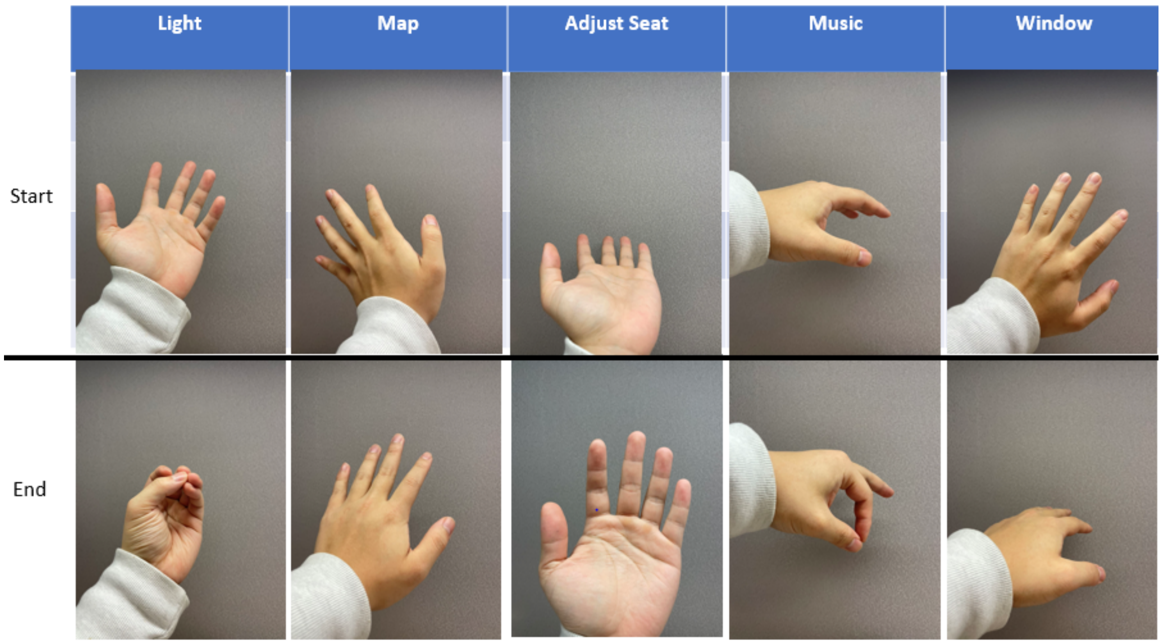


Figure 3.5: Hand-gesture control model: two stages (start and end), users performance one operation by waving/shaping their hand from start stage and end stage. Opening command and closing command using the same set of movement

3.2.3 Task Distraction

The design of distraction from tasks follows the pattern of tasks in a self-driving vehicle. The experiment includes auditory distraction from the task. In a fully self-driving vehicle, distraction from the driving task is not a factor to be considered. However, when the user interacts with in-vehicle interfaces, distraction can also affect the user's reaction and control method. It is well established that most relevant stimuli in driving are visual, and research has shown that visual distractors have a greater impact on driving performance than auditory distractors [38]. Thus, studying auditory distraction remains relevant to our research question.

During the user study, a recording of a series of English characters was played before each control with the simulation interface. The signal was played before the user accepted the visual command. After completing the task, participants were asked to repeat the previous English characters, and researchers recorded any errors made. It was expected that the error rate was higher in tasks with distraction than in tasks without distraction.

3.2.4 Equipment Setup

The following equipment was used in this study.

- **Oculus(Meta) Quest 2**: The Meta Quest 2 is a commonly used virtual reality headset in the market. The headset supports 6 degree pf freedom head and hand tracking. The display is a Fast-switch LCD with a display resolution 1832x1920 per eye. The refresh rate is 72Hz. This headset is integrated with speakers and microphones, and a 6GB RAM. This headset is portable which means it can operate without connection with external computing devices, or connection with both Bluetooth or Lightning 3 cable. The advantage of using Quest 2 in this research is its handtracking feature and its good support of Unity development (Oculus Integration).

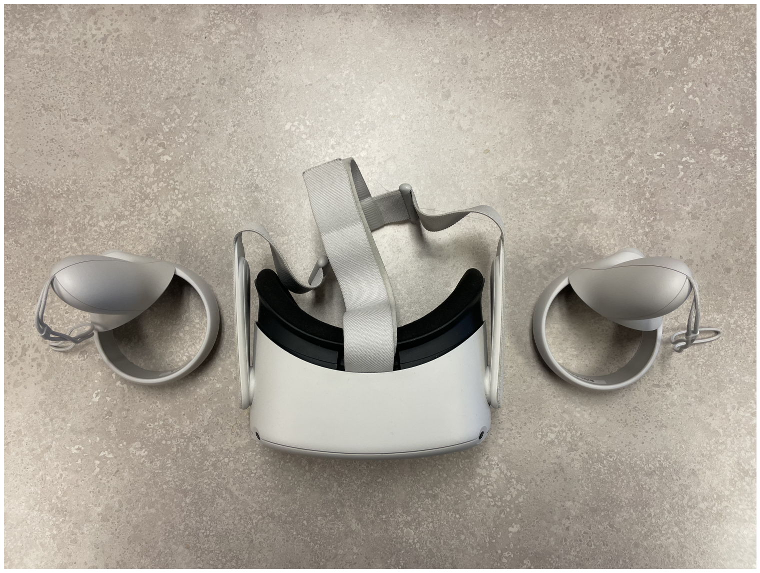


Figure 3.6: Oculus(Meta) Quest 2

- **E4 Empatica wristband**: E4 wristband is an unobtrusive physiological monitoring band. It can collect PPG (photoplethysmography), EDA (Electrodermal activity) data at 4Hz frequency, 3-axis accelerators at 32Hz, and skin temperature at 4Hz. This wristband can connect to most of popular computing devices such as computer, mobile phone, or tablet by using Bluetooth. The data could be transferred to external device in real-time.

Figure 3.7: E4 Empatica wristband

- **Logitech G29 steering wheel**: Logitech G29 is a 10.24 inch racing steering wheel with hand-stitched leather. The maximum rotation degree is 900 degree. The wheel has dual-motor force feedback which can generate real driving experience. Buttons on the wheel can have different functions, directed by its implement. In this research, the buttons on the wheel are acting as different features.

Figure 3.8: Logitech G29 Steering wheel

- **GTR driving simulation seat**: Driving simulator seat to create real driving experience.

3.2.5 Experiment Environment

Although the virtual reality headset already offers an optimal environment for conducting Wizard-of-Oz experiments, an additional secure room setup was constructed. A barrier was erected between the participant and the researcher to prevent the participant from seeing the researcher's actions. The screen in front of the participant was used to provide instruction on the proper use of various features. Prior to the actual experiment (training stage), the participant was able to review the instructions at any time. During the experiment, the monitor was turned off.

Figure 3.9: Wizard-of-Oz experiment room setup, the researcher sits on the left side of the board to operate the features, while the user sits on the right side of the board

3.2.6 Simulation Environment

The simulation application was made by researcher from scratch to met the requirement of this research. The simulation environment is build by Unity with Oculus Integration.

Figure 3.10: Bird view of the city environment, the route is straight along the main road in the mid of his figure

In the simulation environment, the participants are seated in the driver's seat of a self-driving vehicle. The vehicle is driving in a certain route in a big city environment. The speed of the vehicle is 50 km/h to simulate the driving speed in city roads.

3.3 User Performance Study Procedure

In this section, the authors present the procedure of the in-lab user experiment, along with an examination of the considerations and details of each step. The study comprised of 30 participants who were recruited through email (University of Waterloo Engineering Department mailing list) and posters in the main buildings on the University of Waterloo campus. Participants were selected from the age group of 18 to 55 years (young adults and adults) who held a valid full Canadian driver's license (G2). To participate in the study, participants were required to have good or corrected vision with the use of glasses or contact lenses. There are 15 females and 21 males participated in this study, with an average age 26.72 ($SD = 4.05$).

The participants' consent was obtained prior to participating in the main user study, which followed a training session. Participants who had a tendency towards motion sickness or vertigo were excluded from the study. If participants were unable to complete the study, they would be thanked and excused. The main user study was then initiated with a workflow shown in Figure 3.5. Upon completion, participants were compensated with $15 CAD and received an appreciation letter sent to their email. The details of the main user test procedures are discussed below.

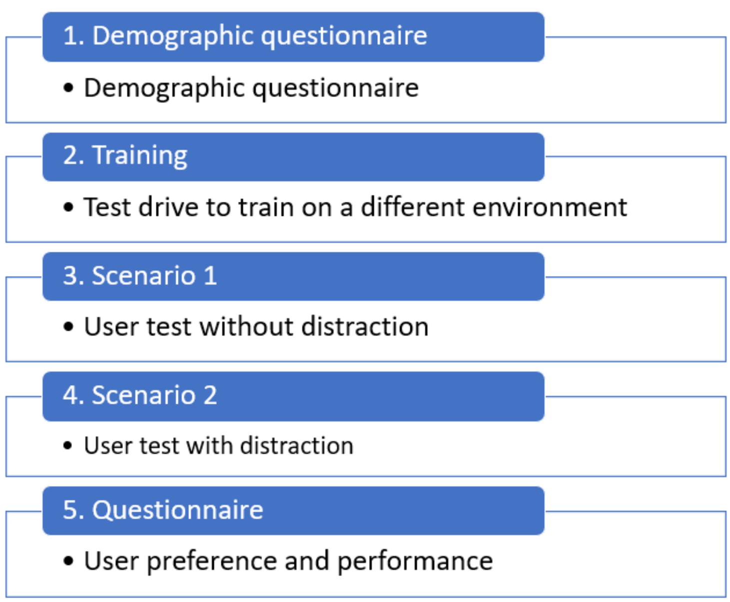


Figure 3.11: Experiment procedure

3.3.1 Preparation

Prior to the initiation of the main user experiment, five sessions of a pilot study were conducted to evaluate the experiment procedure. Based on feedback from the pilot participants, the experiment environment was adjusted and improved to suit for individual participants. After that, participants sat on a driving simulation chair equipped with a steering wheel, while wearing an Oculus Quest 2 virtual reality headset. The researcher, who performed the Wizard-of-Oz methodology, was seated on the left side of the participant and was obscured from view due to the participant's use of the headset. Distraction audio sounds was played by native headset built-in speaker.

3.3.2 User Experiment

Demographic Questionnaire

Before participants started training, they need to do a demographic survey. The demographic information allows researchers know better of participants' background (age, race, driving experience). In the demographic questionnaire, these participants' backgrounds were asked:

- Age
- Gender
- Driving experience
- Driving frequency
- Knowledge of self-driving vehicle

Participant Training

The first step of the user test is the participant training phase. Its aim is to educate participants on the use of the driving simulation system and to screen their suitability. Participants receive instructions from the researcher on the use of the Oculus Quest 2 virtual reality headset and the simulation application. They have the opportunity to practice using the application.

During the screening process, participants are asked to fill out a screening questionnaire. If they are deemed suitable for the study, they are required to sign a consent form. After obtaining the necessary consent, participants undergo a training session on the research equipment.

Upon becoming familiar with the basic functions of the driving simulation system, participants undertake a test experiment. This test uses the same software as the main experiment but takes place in a different virtual environment, specifically a rural setting. Participants are asked to complete five tasks, which are different from the tasks in the main experiment and are prompted by audio commands. A comparison chart with further information is provided in the appendix.

Figure 3.12: Training scenario

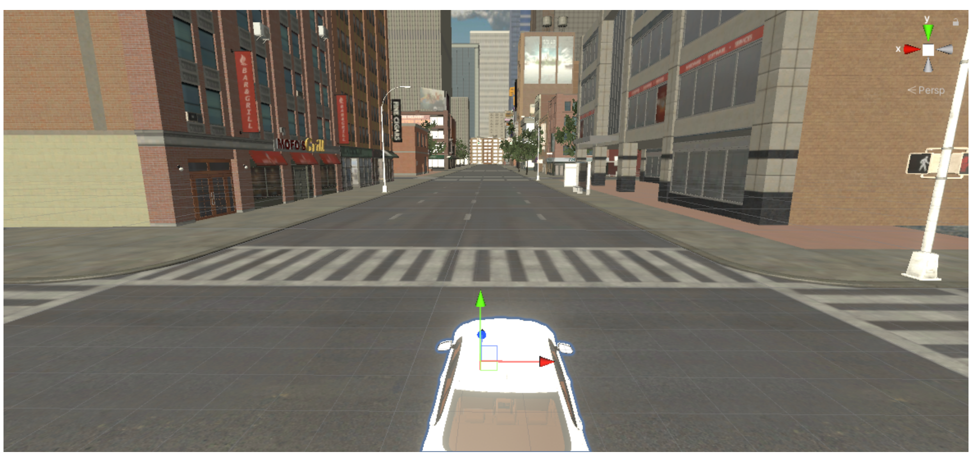


Figure 3.13: Experiment scenario

	Scenario 1 (training)	Scenario 2 (experiment)
Environment	Rural area	Big city
Car velocity	50 km/h	30 km/h
Interaction	Other interaction	Main interaction
Time	5mins	25mins

Figure 3.14: Comparison chart

User Test

In the primary user study, the E4 GSR wristband was worn by participants on their wrist. The wristband was then connected to the researchers' mobile phone via Bluetooth, enabling the recording of the participants' real-time Heart Rate (HR), Electrodermal Activity (EDA), Blood Volume Pulse (BVP) data. Subsequently, visual information in the form of red text was presented in the virtual world to provide instructions to the participants. They were instructed to perform the assigned task. The researcher concurrently recorded the participants' error rate on a computer throughout the experiment.

In order to reduce random noise, the previous mentioned operations were repeated three times. After completion of the three task sets, the experiment transitioned to an examination of countermeasures for distractions, utilizing a within-subject design. During these experimental sets, participants were required to recall a series of numbers (played by headset in English) prior to receiving instructions. Upon completion of the task, they were prompted to recite the numbers. The error rate of the participants was recorded during this period. To minimize random noise, the order of the six task sets are different for different participants.

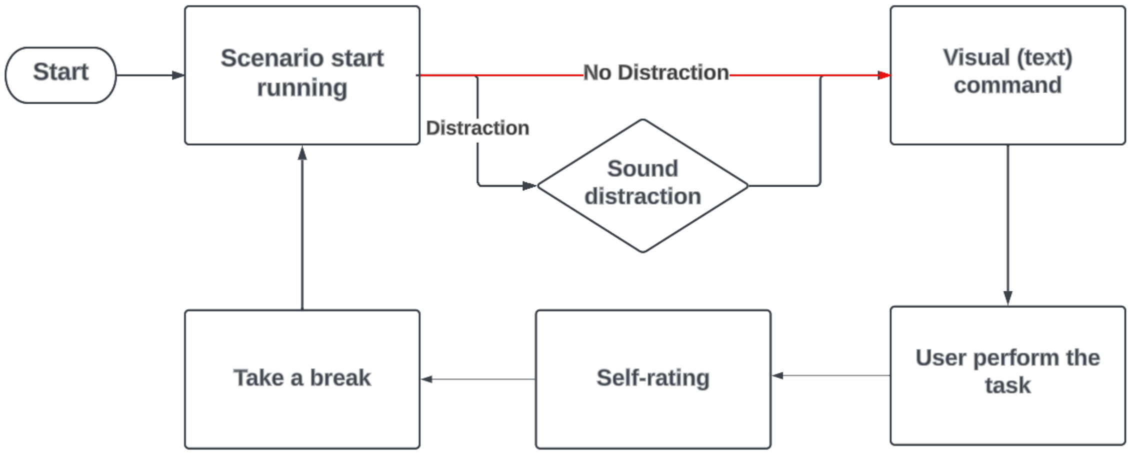


Figure 3.15: Main user test workflow

During each series of tasks, the participants are asked to take a break for half minute. They are asked to take off the VR headset. The purpose of this is to reduce potential fatigue. And, to reduce the learning effect, each series of tasks have a different order, which means participants cannot predict next task. The expected experiment time for one series of tasks is two minutes. And the expected time for the whole user test is 25 minutes, preparation time not included. During the break time, the participants are asked to do a self-rating of three dimensions from NASA Task Load Index (TLX), which are Mental Demand, Temporal Demand, and Effort.

Interview and Questionnaire

After the user study, a user preference questionnaire was conducted. There are two parts of the questionnaire: user experience questionnaire and NASA TLX rating scale. The questions in the user experience questionnaire are:

- Rating scale on intuitiveness of each control models
- Rating scale on complexity of performing each tasks
- General experience of the experiment
- Performance difference on distracted task and non-distracted task
- Fatigue level during experiment
- Nervousness & anxiety during each experiment sets

In the second part of questionnaire, the rating scales in NASA TLX is attached below. NASA TLX is a framework of measuring participants' subjective mental workload. It is a multi-dimensional rating scale with six elements [39]. It is further discussed in Data Collection and Analysis section.

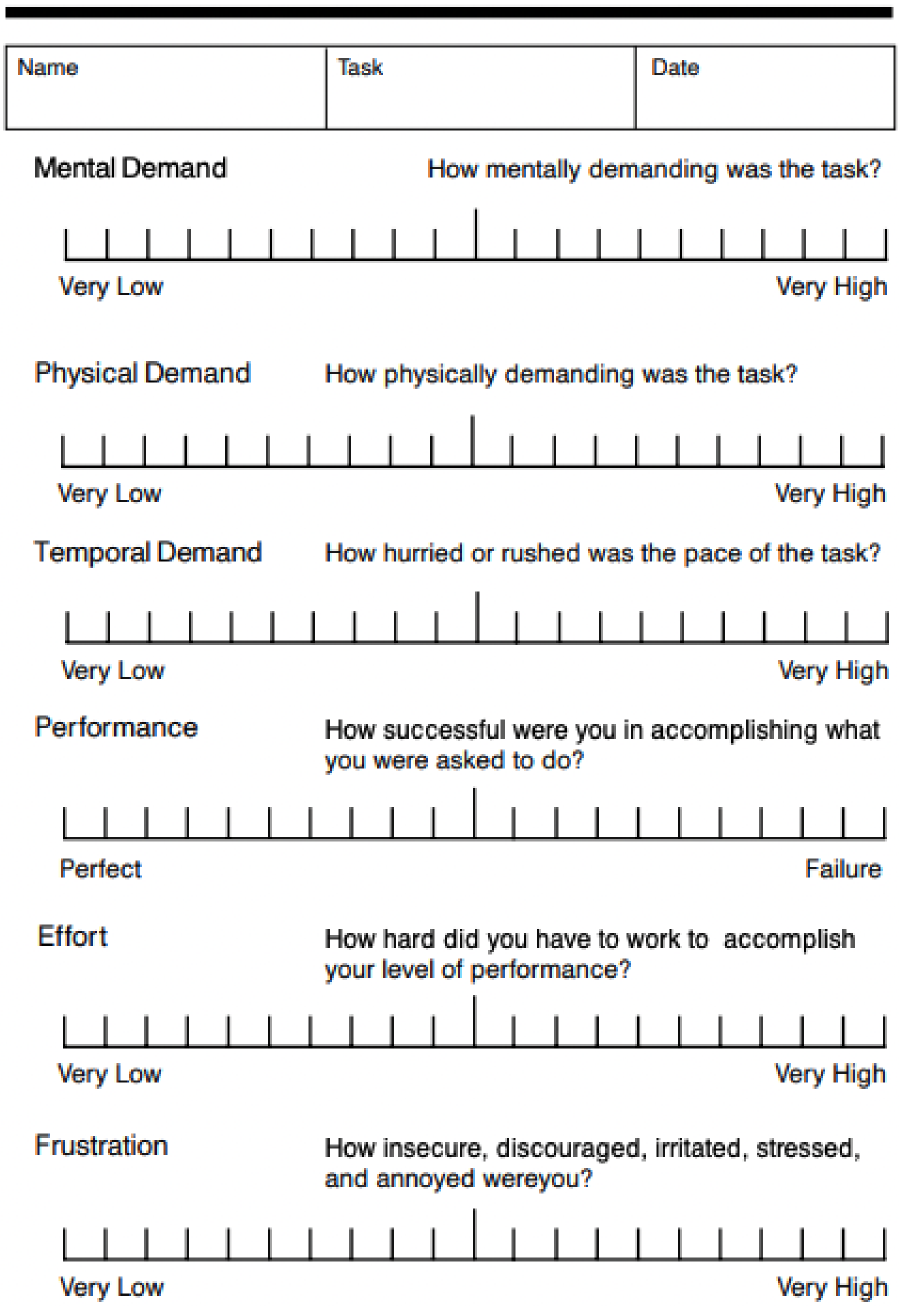


Figure 3.16: NASA Task Load Index chart

After participants finished the questionnaire, an interview was conducted. The researcher was investigated more details in the experiment. The goal of this interview is to obtain subjective data of participants' experience of the whole experiment, and, relate the experiment to the real driving scenario to gain insight of the future self-driving vehicle control models.

3.4 Hypothesis

The prediction of the experiment is made based on the previous study and also the result from the user requirement survey.

Research Question 1: Is there a significant difference of users' mental workload and performance after using different control models?

Null Hypothesis:

H_0: There is no significant difference in the three control models.

Alternative Hypothesis:

H_{A1}: There is a significant difference in the users' mental workload of control models that the physical button control model is more user-friendly and intuitive.
H_{A2}: There is a significant difference in the users' mental workload of control models that the voice interface control model is more user-friendly and intuitive.
H_{A3}: There is a significant difference in the users' mental workload of control models that the mid-air hand gesture control model is more user-friendly and intuitive.

Research Question 2: Is there a significant difference on users' performance with different control models while being distracted?

Null Hypothesis:

H_0: There is no significant difference while being distracted.

Alternative Hypothesis:

H_{A1}: There is a significant difference in the task distraction by control models that the physical button is more easily to be distracted.
H_{A2}: There is a significant difference in the task distraction by control models that the voice command is more easily to be distracted.
H_{A3}: There is a significant difference in the task distraction by control models that the mid-air hand gesture is more easily to be distracted.

Chapter 4

Data Collection and Analysis

In this section, we present a comprehensive analysis of the data collected during our experiment. We begin by describing the types and features of the data that were collected, including both quantitative and qualitative data. This section provides a detailed overview of the various forms of data involved in the research, as well as the context and background of the data collection methods and processes. Our analysis of this data enables us to draw meaningful conclusions regarding the research questions posed in this study, and to provide valuable insights into the relevant research domain.

In the second section of the discussion, the data is analyzed using various statistical methods to gain deeper insights into the findings of the research. The statistical methods used and descriptive data will be discussed in detail. The comprehensive data analysis presented in this section will provide a solid foundation for the findings and conclusions presented in the next chapter.

4.1 Participants

For the study of user requirements, a sample of 150 participants was recruited. This sample comprised 71 females and 75 males, with 4 participants preferring not to disclose their gender. In terms of age distribution, 76 participants were aged between 18 and 30 years, 54 were between 30 and 45 years, 20 were between 45 and 60 years, and 4 were aged above 60 years.

For the user study, the sample comprised of 17 females (20-37 years; $M = 22.14, SD = 3.96$), and 19 males (20-45 years; $M = 27.7, SD = 6.46$). 29 out of 36 participants possessed a Valid G driver's license and 7 participants had a valid G2 driver's license. 5 participants driving 30 mins per week (2 females and 3 males), 4 participants driving one hours per week (4 males), 14 participants driving two hours per week (9 females and 6 males), 7 participants driving four hours per week (4 females and 3 males), and 4 participants driving eight hours per week (2 females and 2 males).

4.2 Data Collection

In this section, the various types of data are presented in chronological order, starting from the earliest data collected to the latest data obtained. The order is: User Requirement Questionnaire (two months before user study), User Error Rate (during the user study), Physiological data collected by Galvanic skin response sensor (GSR) (during the study), User Survey and Interview (after the user study), and NASA-TLX (after the survey).

4.2.1 User Requirement Questionnaire

Each participant was compensated with 4 CAD for completing the online survey, which took an average of 5 minutes to complete. The survey aimed to investigate participants' preferences for various control methods and presented descriptions of five control methods: physical buttons, touch screen, mid-air hand gesture control, voice control, and a separate controller. According to the survey data, 88 participants considered physical button control to be intuitive and requiring minimal effort, while 86 participants viewed voice control as intuitive and requiring limited mental resources. The popularity of mid-air hand gesture control was comparable to that of touch screen control, with 41 and 36 participants, respectively. And lastly, the separate controller was selected by only 6 participants as an intuitive control method.

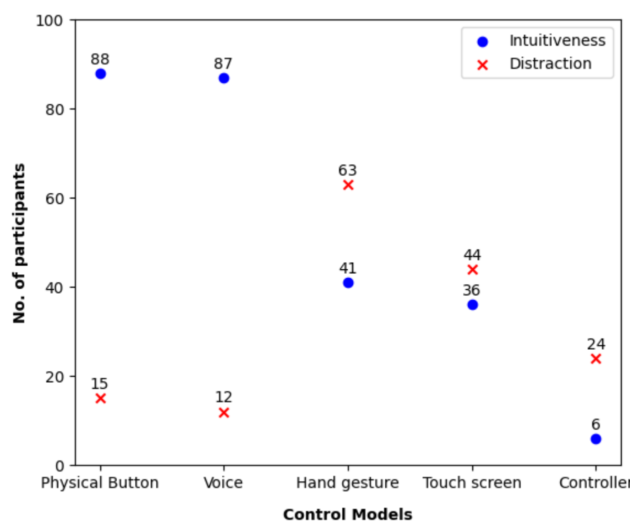


Figure 4.1: Two questions asked in Survey: "which control model is most intuitive and most distracted?" Most of people think Physical button (88) and Voice control (87) are the most two intuitive ones. Most of people think Hand gesture control (63) is the most distracted one.

4.2.2 User Error Rate

User Error Rate (UER), also know as human error rate, is the rate of the times of a participant makes a wrong entry (control interaction in this research) divided by the total interaction times. The UER is usually used to evaluate the usability and utility of a system or product. The lower the user error rate, the system is more intuitive.

In this experiment, the UER is recorded by researcher by direct observation of the user test. There are two types of UER being recorded in this study: UER of participant successfully interacting with the features, and the UER of participants remembering the numbers when it is in the distraction task sets. The reason of recording two-way User Error rate is to understand user's performance when they interacting with self-driving vehicle in distraction, and additionally, understanding each control models' performance, as a distraction from user's current task.

When participants make a wrong entry (control interaction), the researcher will record a score 0 for the current task. If the participants made a correct interaction input of current task, the score is recorded as 1. Each task sets have ten different control tasks. And there are totally six task sets for each one participants. In the six task sets, there are three task sets with distraction and three tasks sets without distraction. Totally, each participants are asked to perform sixty interaction tasks.

In the task sets which distraction is added, wrong repeat of the previous numbers are also recorded by researcher. The default score of all task distractions is one, which assume participants will repeat the number correctly. However, if the repeat is wrong, researcher will record a zero of the current task.

4.2.3 Physiological Data

The physiological data is collected by E4 Empatica Wristband. The data collected by E4 Empatica Wristband sensors is in multiple form, they are: HR, EDA, BVP, Temperature (TEMP), Acceleration (ACC). Since temperature data is limited in this experiment as the participants are not doing high amount of exercise, the temperature data is dropped out from this experiment. And, the acceleration data is irrelavant to this research, it was dropped out as well. In this sub-chapter, I will discuss:

1. Data types and features
2. Data-cleaning pipeline

In this section, we discussed all the pre-processing tasks of physiological data gathered from E4 Empatica Wristband. The processing pipeline is same for all kinds of data since they are in a same format. The format of data is: .csv files with one column of data. The first row is the initial time of the session and the second row is the sample rate in Hz.

Data types and features

EDA: Electrodermal Activity refers to electrical changes of a part of human body. It is usually measure at the surface of skin. The mechanics of EDA measurement is: when the skin receive signals from the brain, then the level of sweating is higher and the EDA activity arise. For the E4 Empatica wristband is using a method that could capture electrical conductance on the skin. The EDA activity score is usually affected by human's emotional activity, physical activity, cognition workload.

In EDA measurement, there are two types of measurement - tonic skin conductance level and phasic skin conductance response. Skin Conductance Response (SCR) is the increases in the conductance of the skin. And Skin Conductance Level (SCL) is the level of conductance of the skin.

BVP: The Blood Volume Pulse is also measured by E4 wristband sensor. The data is collected by a PPG sensor with an algorithm that can combine light signal of both green and red exposure. The sample size of BVP is 64 Hz(64 times per second).

HR: The heart rate recorded by the wristband is the average heart rate values of ten seconds. The frequency of recording the heart rate is one second. Since the value recorded is not the current heart rate value, the data is not in real time. The wristband started recording after ten seconds of initialization. The unit is expressed as Beats per minute (BPM).

Data-cleaning Pipeline

First step: Generating time stamps. The time stamps of each interaction tasks are auto-generated by a Python Script while the participants doing user test. The time stamps are in Unix time stamp to match the E4 data.
Second step: Data synchronization. In this step the data collected from E4 Empatica wristband is synchronized with the time stamps generated before.
Third step: Data cleaning. In this step, each data file (csv file) is separated into six files to represent six task sets of each participants. And also, each time stamps-related data is extracted from each file. After these operations, the cleaned data from each task sets are further combined to better represent the performance of one participant.

4.2.4 Post-study User Interview

The post-study interview was conducted after the user test. These questions were asked:

- "Did you feel tried or dizziness during the experiment?"
- "How do you think about the control method in general?"
- "Do you think the control method you used is intuitive? How intuitive is it."
- "Are you willing to use this control method in your day-to-day driving?"
- "Do you think this control method will be used in the fully self-driving vehicle?"

4.2.5 NASA Task Load Index

NASA Task Load Index (NASA TLX) is a subjective, multidimensional rating scale designed to obtain the overall level of perceived task workload. It was developed by the National Aeronautics and Space Administration (NASA) in the mid-1980s as a tool for evaluating the usability and workload of various systems [39].

NASA-TLX consists of six dimensions of workload: mental demand, physical demand, temporal demand, performance, effort, and frustration [39]. Participants are asked to rate each dimension on a scale form 0 to 100, where a higher score indicates a higher level of perceived workload. The scores for each dimension are then averaged to produce an overall task workload rating. The interpretation of each dimensions are: [39]

- **Mental Demand**: The mental demand needed for accomplishing the task
- **Physical Demand**: The physical demand needed for accomplishing the task

- **Temporal Demand**: The time consuming of accomplishing the task, how hurry is it
- **Performance**: The success rating of accomplishing the task
- **Effort**: The level of hard-working when accomplishing the task
- **Frustration**: The frustration level while doing the task [39]

NASA TLX has been widely used in a variety of fields, including human-computer interaction, ergonomics, and psychology, and has been found to be a reliable and valid measure of perceived workload. The tool is considered particularly useful for evaluating complex, multitasking environments, such as those encountered in aviation, space flight, and other high-stakes industries [40, 39].

NASA TLX rating scales are conducted two times in the user study: one time after participants finished all the tasks without distraction and one time after participants finished tasks with distraction.

4.3 Data Analysis

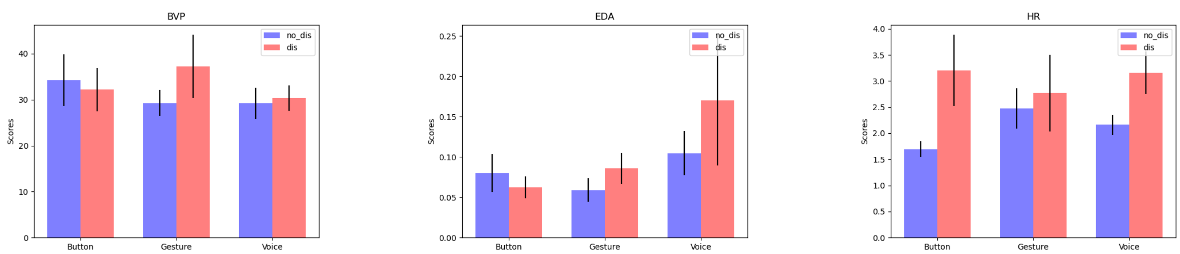
This section of the chapter presents a detailed discussion of our findings, with a focus on objective and subjective data analysis. We begin by examining the objective data collected during our study, specifically the physiological data, and use statistical methods such as t-test and ANOVA to better illustrate our statistical findings. We then move on to analyze subjective data from various sources such as interviews, questionnaires, to validate our statistical findings.

In this study, we are using factorial design to conduct the data analysis and findings. A factorial design is an experimental strategy that allows researchers to simultaneously examine the effects of multiple independent variables (or factors) and their interactions on a dependent variable. The key advantage of a factorial design is its efficiency, as it provides a comprehensive view of multiple factors and their interactions within the same experiment. We aligned factorial design with multiple forms of data to illustrated our findings. In this study, we are using two-way (2X3) ANOVA as factorial design to analyze the data and test the result. There are two category of independent vaiables: control models and distraction. For control models, there are three independent variables: Hand-gesture, button, voice. For distraction factor, there are two independent variables: distraction and non distraction. Two-way ANOVA is a statistical test that allows for the simultaneous evaluation of the influence of two different categorical independent variables on one dependent variable. It is an extension of the one-way ANOVA. One-way ANOVA is a commonly used methodology that used when the experiment has a quantitative outcome and a single categorical explanatory variable with more than two levels. As an extension of one-way ANOVA, two-way ANOVA fit this study better since for every data types, there are two independent variables: control model and distraction.

4.3.1 Physiological Data Exploration

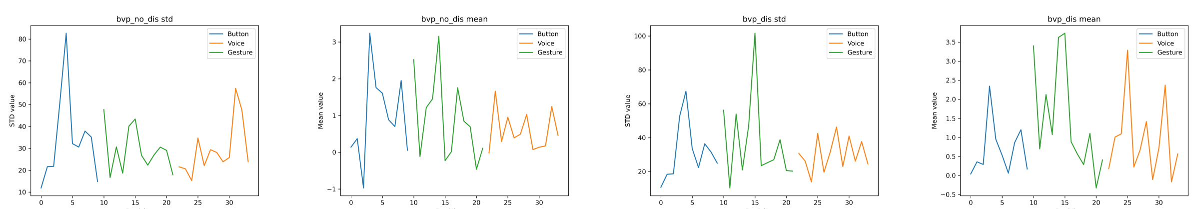
After data processing by the cleaning pipeline mentioned in last chapter, the data was divided to small portion to represent each task sets (10 tasks). For each data sets, it currently contains six different types of data: BVP data with/without distraction, EDA data with/without distraction, and HR data with/without distraction. After outlier filtering by using Z-score (Z=2), the data set is distilled and calculated into mean and standard deviation.

When Evaluating the control model factor, Mid-air hand Gesture has a relatively low performance with distraction based on physiological data, with average heart rate 2.48 (highest) and BVP 37.19 (highest). And for voice and button control model, their physiological data are cross balanced. Further more, when evaluating distraction factor among three control models, hand gesture control model have a higher difference between distraction data and non-distraction data.



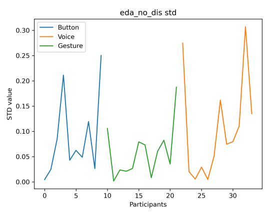
(a) BVP (b) EDA (c) HR

Figure 4.2: Average std values

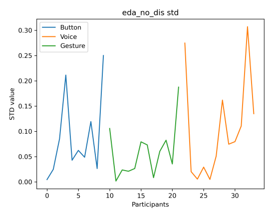


(a) no dis STD (b) no dis Mean (c) dis STD (d) dis Mean

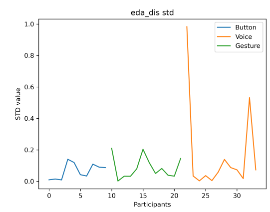
Figure 4.3: BVP data



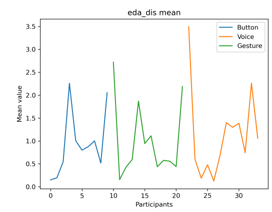
(a) no dis STD



(b) no dis Mean

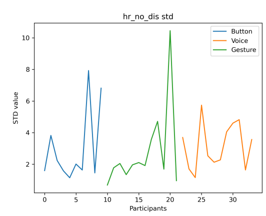


(c) dis STD

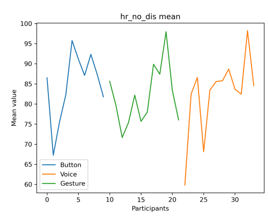


(d) dis Mean

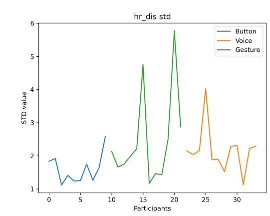
Figure 4.4: EDA data



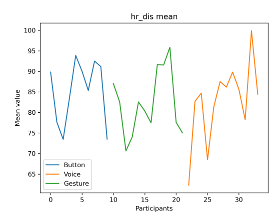
(a) no dis STD



(b) no dis mean



(c) dis STD



(d) dis Mean

Figure 4.5: HR data

Three two-way ANOVA tests are applied to measure the Physiological data. From the observation of the data, it is concluded that data groups acquired from GSR wristband are not significantly different, except the comparison of distraction factor in Heart Rate data.

Source	DF	Sum of Square (SS)	Mean Square (MS)	F Statistic (df_1,df_2)	P-value
Factor A - rows (A)	1	106.0391	106.0391	0.3942 (1,64)	0.5323
Factor B - columns (B)	2	187.8885	93.9443	0.3492 (2,64)	0.7066
Interaction AB	2	301.5969	150.7984	0.5606 (2,64)	0.5736
Error	64	17215.7879	268.9967		
Total	**69**	**17811.3124**	**258.135**		

Figure 4.6: BVP data ANOVA test. Factor A: Distraction, Factor B: Control models

Hypothesis	P-value	Result
H0_1: there will be no significant difference on BVP data of Distraction Ha_1: there is significant difference	0.5323	No significant difference of three groups of data on distraction Fail to reject H0_1

Hypothesis	P-value	Result
H0_2: there will be no significant difference on BVP data of Interaction models Ha_2: there is significant difference	0.7066	No significant difference on data of different groups of interaction Fail to reject H0_2

Hypothesis	P-value	Result
H0_3: there will be no significant difference Interaction of two factors: Interaction and Distraction Ha_3: there is significant difference	0.5736	No significant difference of the interaction of two factors Fail to reject H0_3

Figure 4.7: Hypothesis test result of BVP data

For Factor A (distraction), the null hypothesis could not be rejected as the p-value was greater than the chosen alpha level (p = .5323, F(1, df) = .3942, $\eta^2 = .0061$). This suggests that the difference between group averages is not statistically significant. The small effect size ($\eta^2 = .0061$) indicates that the magnitude of the difference between averages is minimal.

For Factor B (control model), the null hypothesis could not be rejected, with a p-value greater than the alpha level (p = .7066, F(1, df) = .3492, $\eta^2 = .011$). The difference between group averages was not statistically significant. The effect size was small ($\eta^2 = .011$), indicating a minor difference between group averages.

The interaction between Factors A and B was also not significant (p = .5736, F(1, df) = .5606, $\eta^2 = .017$), supporting the null hypothesis that the averages of all groups are equal. The observed effect size was small ($\eta^2 = .017$), suggesting a minor interaction effect on the dependent variable.

Therefore, from BVP data ANOVA test, either main effect and interaction effect are not significant. No major conclusion could be conducted by just observing BVP data.

Source	DF	Sum of Square (SS)	Mean Square (MS)	F Statistic (df_1,df_2)	P-value
Factor A - rows (A)	1	0.01185	0.01185	0.6166 (1,64)	0.4352
Factor B - columns (B)	2	0.06797	0.03398	1.7688 (2,64)	0.1788
Interaction AB	2	0.02014	0.01007	0.5242 (2,64)	0.5945
Error	64	1.2296	0.01921		
Total	**69**	**1.3295**	**0.01927**		

Figure 4.8: EDA data ANOVA test. Factor A: Distraction, Factor B: Control model

Hypothesis	P-value	Result
H0_1: there will be no significant difference on EDA data of Distraction Ha_1: there is significant difference	0.4352	No significant difference of three groups of data on distraction Fail to reject H0_1

Hypothesis	P-value	Result
H0_2: there will be no significant difference on EDA data of Distraction Ha_2: there is significant difference	0.1788	No significant difference of three groups of data on distraction Fail to reject H0_2

Hypothesis	P-value	Result
H0_3: there will be no significant difference on EDA data of Interaction of two factors Ha_3: there is significant difference	0.5945	No significant difference of the interaction of two factors Fail to reject H0_3

Figure 4.9: Hypothesis test result of EDA data

For Factor A, the null hypothesis could not be rejected as the p-value was greater than the chosen alpha level (p = .4352, F(1, df) = .6166, η^2 = .0095). This suggests that the difference between group averages is not statistically significant. The small effect size (η^2 = .0095) indicates that the magnitude of the difference between averages is minimal.

For Factor B, the null hypothesis could not be rejected, with a p-value greater than the alpha level (p = .1788, F(1, df) = 1.7688, η^2 = .052). The difference between group averages was not statistically significant. The effect size was small (η^2 = .052), indicating a minor difference between group averages.

The interaction between Factors A and B was also not significant (p = .5945, F(1, df) = .5242, η^2 = .016), supporting the null hypothesis that the averages of all groups are equal. The observed effect size was small ($\eta^2 = .016$), suggesting a minor interaction effect on the dependent variable.

Same as BVP data ANOVA test, from EDA data ANOVA test, either main effect and interaction effect are not significant. No major conclusion could be conducted by just observing EDA data.

Source	DF	Sum of Square (SS)	Mean Square (MS)	F Statistic (df_1,df_2)	P-value
Factor A - rows (A)	1	14.9336	14.9336	5.0603 (1,63)	0.02798
Factor B - columns (B)	2	0.6028	0.3014	0.1021 (2,63)	0.9031
Interaction AB	2	4.9687	2.4843	0.8418 (2,63)	0.4357
Error	63	185.9226	2.9512		
Total	**68**	**206.4277**	**3.0357**		

Figure 4.10: HR data ANOVA test. Factor A: Distraction, Factor B: Control model

Hypothesis	P-value	Result
H0_1: there will be no significant difference on HR data of Distraction Ha_1: there is significant difference	0.02798	Significant difference of three groups of data on distraction exists Reject H0_1

Hypothesis	P-value	Result
H0_2: there will be no significant difference on HR data of Distraction Ha_2: there is significant difference	0.1021	No significant difference of three groups of data on Interaction Fail to reject H0_2

Hypothesis	P-value	Result
H0_3: there will be no significant difference on HR data of Distraction Ha_3: there is significant difference	0.8418	No significant difference of data the interaction of two factors Fail to reject H0_3

Figure 4.11: Hypothesis test result of HR data

In this test, The p-value equals 0.02822, ($P(x \leq 5.041) = 0.9718$). It means that the chance of type I error (rejecting a correct H0) is small: 0.02822 (2.82%). For distraction factor (p = .4352, F(1, df) = .6166, η^2 = .0095), pairwise comparisons are made. The results of a post hoc Tukey HSD test revealed no significant differences between Button group under study. The mean value for no-dis scenario group (M = 1.69, SD = 0.518) was not significantly higher than dis scenario (M = 3.20, SD = 2.37), with a mean difference of 1.51, and $p = 0.05223$. The P value is very close to 0.05 which may indicate there is a weak relationship between two datasets.

For Factor B (control model), the null hypothesis could not be rejected, with a p-value greater than the alpha level (p = .1788, F(1, df) = 1.7688, η^2 = .052). The difference between group averages was not statistically significant. The effect size was small (η^2 = .052), indicating a minor difference between group averages.

The interaction between Factors A and B was also not significant (p = .5945, F(1, df) = .5242, η^2 = .016), supporting the null hypothesis that the averages of all groups are equal. The observed effect size was small (η^2 = .016), suggesting a minor interaction effect on the dependent variable.

Aso, the results of a post hoc Tukey HSD test revealed no significant differences between Voice group under study. The mean value for no-dis scenario group (M = 2.47, SD = 1.4) was not significantly higher than dis scenario (M = 2.76, SD = 2.65), with a mean difference of 0.291, and $p = 0.074$.

However, significant difference was found of Heart rate value of Gesture group under study. The mean value for no-dis scenario group (M = 2.16, SD = 0.68) was not significantly higher than dis scenario (M = 3.15, SD = 1.56), with a mean difference of 0.998, and $p = 0.0433$.

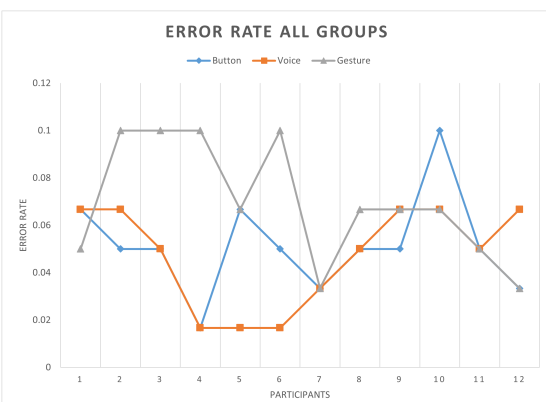
4.3.2 User Error Rate

It is found that participants with Voice control model has the lowest average error rate, and the participants with Hand Gesture control model has the highest average error rate. When investigating distraction factors, all three control models shows that the error rate with distraction are higher than that without distraction. Physical Button group's participants has the lowest average error rate with distraction and the Hand gesture group's participants has the highest average error rate.

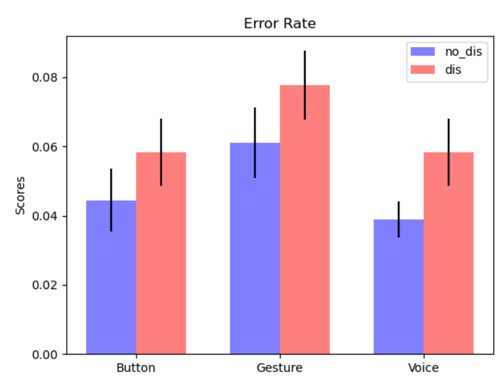
-	no dis	dis
Button	0.0444	0.0583
Voice	0.0389	0.05836
Gesture	0.0611	0.07778

Table 4.1: Error rate Mean values of each group

Error rate under distraction



(a) Error distribution



(b) Average Error rate

Source	DF	Sum of Square (SS)	Mean Square (MS)	F Statistic (df_1,df_2)	P-value
Factor A - rows (A)	1	0.005	0.005	4.5575 (1,66)	0.03649
Factor B - columns (B)	2	0.006142	0.003071	2.7992 (2,66)	0.0681
Interaction AB	2	0.00009259	0.0000463	0.0422 (2,66)	0.9587
Error	66	0.07241	0.001097		
Total	**71**	**0.08364**	**0.001178**		

Figure 4.13: Interaction ANOVA test. Factor A: distraction, Factor B: Control models

Hypothesis	P-value	Result
H0_1: there will be no significant difference on Error rate of Distraction Ha_1: there is significant difference	0.0365	Significant difference of three groups of data on distraction exists Reject H0_1

Hypothesis	P-value	Result
H0_2: there will be no significant difference on Error rate data of Interaction Ha_2: there is significant difference	0.0681	No Significant difference of three groups of data on Interaction Fail to Reject H0_2

Hypothesis	P-value	Result
H0_3: there will be no significant difference on Error rate data Interaction of two factors Ha_3: there is significant difference	0.9587	No Significant difference of data on two factors Fail to reject H0_3

Figure 4.14: Error Rate hypothesis test result

For Factor A (distraction), the null hypothesis was rejected, indicating a p-value smaller than the chosen alpha level (p = .03649, F(1, df) = 4.5575, η^2 = .065). This suggests that the differences between some group averages are statistically significant. The medium effect size (η^2 = .065) further indicates that the magnitude of difference between the averages is moderate.

Factor B (control models), on the other hand, did not reject the null hypothesis, as the p-value was greater than the alpha level (p = .0681, F(1, df) = 2.7992, η^2 = .078). This implies that the difference between group averages was not statistically significant. Despite this, the effect size was medium ($\eta^2 = .078$), indicating a moderate difference between the group averages.

As for the interaction between Factors A and B, it was not significant (p = .9587, F(1, df) = .0422, $\eta^2 = .0013$), which supports the null hypothesis that the averages across all groups are equal. The very small effect size ($\eta^2 = .0013$) suggests a minimal interaction effect on the dependent variable.

Two main effects both have significant difference but the interaction effect does not have significant difference. And the interaction effect of two factors is also not significant. We further performed a Tukey HSD test to evaluate pairwise difference of three groups in distraction factor.

There is no significant difference found of Button group's error rate with distraction factor. The mean value for no-dis scenario group (M = 0.0444, SD = 0.0328) was not significantly higher than dis scenario (M = 0.0583, SD = 0.03516), with a mean difference of 0.014, and $p = 0.328$.

Also, there is no significant difference found of Voice group's error rate with distraction factor. The mean value for no-dis scenario group (M = 0.0389, SD = 0.0192) was not significantly higher than dis scenario (M = 0.0583, SD = 0.03516), with a mean difference of 0.019, and $p = 0.107$.

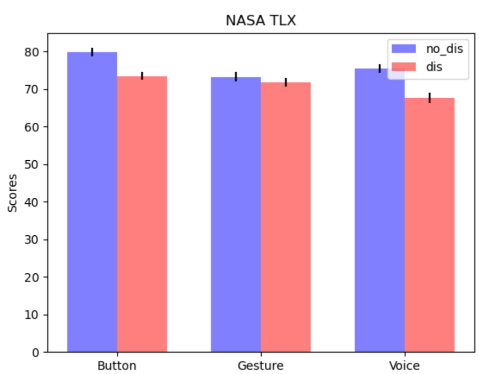
There is no significant difference found of Gesture group's error rate with distraction factor. The mean value for no-dis scenario group (M = 0.0611, SD = 0.0371) was not significantly higher than dis scenario (M = 00778, SD = 0.03516), with a mean difference of 0.017, and $p = 0.275$.

4.3.3 Participants' Workload from NASA TLX

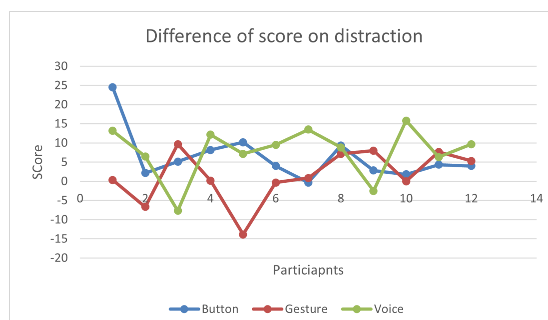
It is observed that most participants, except five, give a lower TLX score in tasks with distraction. This indicated that the average task workload is higher with the distraction, to compare that of the data groups without distraction. For non-distraction scenario, hand Gesture control has the lowest score and the physical button have the highest score. For distraction scenario, however, the voice control model has the lowest score and the physical button model still have the highest score.

-	no dis	dis
Button	79.81943	73.47224
Voice	75.40278	67.69445
Gesture	73.2361	71.70834

Table 4.2: Mean values of each group



(a) Average Score



(b) Difference of distraction factor

The results for Factor A(distraction) indicated a statistically significant difference among group means, F(1, df) = 26.0358, $p < 0.00001$, partial $\eta^2 = .28$. This suggests a large effect size.

Similarly, the results for Factor B(control model) also suggested a statistically significant difference among group means, F(1, df) = 9.4893, $p < 0.0003$, partial $\eta^2 = .22$, indicating a large effect size.

Lastly, the interaction effect between Factors A and B also reached statistical significance, F(1, df) = 3.3922, $p < 0.05$, partial $\eta^2 = .093$, indicating a medium effect size.

Source	DF	Sum of Square (SS)	Mean Square (MS)	F Statistic (df_1,df_2)	P-value
Factor A - rows (A)	1	485.6769	485.6769	26.0358 (1,66)	0.000003049
Factor B - columns (B)	2	354.0306	177.0153	9.4893 (2,66)	0.0002386
Interaction AB	2	126.5581	63.279	3.3922 (2,66)	0.0396
Error	66	1231.1764	18.6542		
Total	**71**	**2197.442**	**30.9499**		

Figure 4.16: NASA TLX ANOVA test. Factor A: Distraction, Factor B: control models

Hypothesis	P-value	Result
H0_1: there will be no significant difference on NASA-TLX Data of Distraction Ha_1: there is significant difference	0.0000003	Significant difference of three groups of data on distraction exists Reject H0_1

Hypothesis	P-value	Result
H0_2: there will be no significant difference on NASA-TLX data of Distraction Ha_2: there is significant difference	0.000002	Significant difference of three groups of data on Interaction exists Reject H0_2

Hypothesis	P-value	Result
H0_3: there will be no significant difference on NASA-TLX data interaction of two factors Ha_3: there is significant difference	0.0396	Significant difference of data on two factors exists Reject H0_3

Figure 4.17: NASA TLX Hypothesis test result

It is observed that both main effects and interaction effects have significant difference. The following tukey HSD post hoc analysis results is below.

In the Tukey HSD test of Button group, significant difference was found with distraction factor. The mean value for no-dis scenario group (M = 79.82, SD = 4.08) was not significantly higher than dis scenario (M = 73.47, SD = 3.81), with a mean difference of 6.347, and $p = 0.0007$.

Significant difference was also found with distraction factor for Voice group. The mean value for no-dis scenario group (M = 75.4, SD = 4.17) was not significantly higher than dis scenario (M = 67.7, SD = 4.77), with a mean difference of 7.708, and $p = 0.00036$.

However, significant difference was not found with distraction factor for Gesture group. The mean value for no-dis scenario group (M = 73.24, SD = 4.74) was not significantly higher than dis scenario (M = 71.7, SD = 4.24), with a mean difference of 1.52, and $p = 0.415$. This suggest the gesture control model is not easy to affected user's work load score, comparing to other two factors.

Post hoc analysis of control-model-wise comparison was also conducted (Figure 2.24, 2.25). It is found that when comparing physical button and Voice control model, they have different effect on user's workload for both with/without distraction scenario. And comparing Button and Gesture control model, only scenario without distraction model has a significant difference. And when comparing button and voice control model, both two scenarios are not significant. This suggested Hand gesture control has a general higher workload for users (M=73.24 without distraction and M=71.7 with distraction).

Pair	Difference	SE	Q	Lower CI	Upper CI	Critical Mean	p-value
x1-x2	4.4167	1.2536	3.5232	0.0665	8.7668	4.3502	0.046
x1-x3	6.5833	1.2536	5.2516	2.2332	10.9335	4.3502	0.002119
x2-x3	2.1667	1.2536	1.7284	-2.1835	6.5168	4.3502	0.4488

Figure 4.18: Tukey HSD test for control model factor without distraction. X1, X2, X3 are Button, Voice, and Gesture

Pair	Difference	SE	Q	Lower CI	Upper CI	Critical Mean	p-value
x1-x2	5.7778	1.24	4.6596	1.4748	10.0808	4.303	0.006495
x1-x3	1.7639	1.24	1.4225	-2.5391	6.0669	4.303	0.5784
x2-x3	4.0139	1.24	3.237	-0.2891	8.3169	4.303	0.07145

Figure 4.19: Tukey HSD test for control model factor with distraction. X1, X2, X3 are Button, Voice, and Gesture

Chapter 5

Discussion

5.1 User Workload

The findings of this study showed that different control models may have different effect of driver's workload (both mentally and physically). Basically all the participants had higher mental and physical workload while being distracted. From the post-study survey, most of people stated that the tasks with distraction caused more workload compared to tasks without distraction. And, it is found, from NASA-TLX rating scale, hand-gesture has a lower workload score comparing that of Voice control and Physical button control. It is explained by Udara, in 2016, that hand-gesture can provide certain level of flexibility of interactions with vehicle, however, it have higher wordload for a long time driving task [35].

From NASA-TLX scores, we found that different control models have significant difference on workload score. And, null hypothesis of distraction have no effect on workload score was also rejected. It is well explained the participants' statement in post-study interview that, users of Voice and physical button groups didn't obviously reflect their workload increasing. When observing the score difference caused by distraction, participants from hand-gesture group have the lowest average difference score (figure 4.29). The possible reason of this might be the body movement have a lower impact by distraction from other perceptions [41].

However, from the Hypothesis test of physiological data, two out of three data types' (BVP and EDA) results are not significant as that of NASA-TLX score. This situation was not expected by researchers. The potential reason of this is the interaction tasks is not intensive. So, it is not obvious to show high fluctuation and significant difference on data acquired by GSR wristband. Since there is a time gap between each single task (4 seconds). And, a break (30 secs) between each task sets are applied, which could also be a reason that the tasks are not intensive. In the future research, a more detailed methodology design and more intensive time and workload setup is necessary. On the other hand, from the heart rate hypothesis test, we found that distraction do have significant impact on heart rate score. It is found that the physical button have a greater impact on heart rate, comparing to voice control and hand-gesture control. On the other hand, there is not significant difference found of control models' effect on heart rate. However, the hand gesture control model has a higher average heart rate than physical button group and voice group. And, from the observation of data, the voice control model do have the lowest average value of heart rate, and significant lower standard deviation data point compare that to hand-gesture interaction, which can support the conclusion from NASA-TLX observation and post-study survey observation.

5.2 User Performance

User performance was mainly evaluated from the Error rate and post-study survey. From the hypothesis test, we can find that the different control model have significant effect on the error rate. Draper's study in 2002 also have revealed the similar result [42]. With the voice control interaction have the lowest average error rate. [42] And, the hand-gesture control have the most highest error rate. It is also a reflection from the post-study survey that, five out of twelve participants think it is easy to make a mistake while using hand-gesture control. To compare with that, only one participant think it is easy to make mistake while using physical button, and zero participant think it is easy to make mistake while using voice control interaction. And, from the hypothesis test, it is proved that distraction have a significant effect of user's error rate ($p = 0.0.0365$). It is expected that each control models are affected by distraction. The distraction affect ratio are 1.313, 1.5, and 1.273 for Button, Voice, and Gesture groups. The equation for calculating the difference ratio is:

$$r = \frac{e(distraction)}{e(non - distraction)} \tag{5.1}$$

It is observed that, even the voice group doesn't have the highest average error rate, it have the highest error rate difference ratio. It is caused by the effect of same modality. In 1993, a study by D.N Card found that control methods are easier to have higher error rate when encounter distractions with same modality [43]. The voice control model and the distraction modality are both in audio modality. And the hand-gesture interaction have the lowest difference ratio, which was proved by Ma in 2017, that gesture based control model as a secondary task is good at avoiding distraction while doing primary task [44].

5.3 User Preference

Usability of Vehicle Interfaces. People generally think that the control model tasks is easy to accomplish, when the task doesn't contain audio distraction. In all three control model participant groups, participants has stated that they feel the task is not very challenge. For example, one participant in hand-gesture interface group said: *"I feel easy to interact with the features with such a novel control method (hand-gesture interface), even though I never used it before."* And one participant in Voice command interface stated that: *"it is easy to use, I expect this feature could be added in my car."* However, there are two participants in Hand-gesture control group, and voice control group, stated that they are not used to interact with vehicle by using hand gestures. One participant said: *"When I was given a task to do (interact with a feature in vehicle), I always tend to do it in the original way (the interface in the existing vehicle)."* And, one participant from Voice control group stated that: *"When I was speaking to the car, I was a little bit shy. I don't know the reason behind it but I think I am not used to speak to a machine."*

Distraction by audio tasks. While distraction is added into the task sets, a part of participant feel stressful to perform. The effect is most obvious with the Hand-gesture control group. And, the effect is not significant for Voice-control group. For example, one participant stated that: *"When I was told that I have to do the task while remembering the number, I thought it will be hard since both control method and distraction are sounds. But it's actually easier than I think."*

Distraction by control models. All three groups of participants are affected by the distraction caused by control models. Most of participants mentioned their short-term memorization was affected by interacting with the vehicle. One participant in Voice control group stated: *"Memorizing the numbers is easy by itself, but repeating it after I do these* *tasks is not too easy."* And a participant in Hand-gesture control group said: *"I have to stay focus on the number in order to not forgetting it."* However, by our observation, the effect by control models in Physical button group is not as great as other two groups. One participant stated that: *"It could effect my memorization, but not too much. There is no big difference."*

User's mental workload, fatigue. After explaining the definition of mental workload, and fatigue, participants illustrated their feeling. Most of people in all three groups didn't have strong feeling of fatigue. With a few people (3 in voice command, 3 in hand-gesture, 3 in physical button) illustrated that they have low level of fatigue. For example: one participant stated that: *"After these task sets, I am little tired."* And most of people didn't express their tiredness or fatigue in the interview. In study by Detjen, authors also found in-vehicle interaction have different effect on user's workload and tiredness [45].

Certain Task findings. From Participant interview and questionnaire, there is no strong evidence from participants believe a certain interaction task (in-vehicle feature) is easier to perform comparing to other tasks. However, three participants, from Hand-gesture interaction group, do stated that the the 'Open the map' task operated by hand gesture is *'intuitive'* and *'easy'*.

Chapter 6

Conclusion

To answer the two research questions raised in this study, we have build an VR simulator to Evaluate three Control Modalities with distraction factor. We have found that distraction is a major factor when evaluating Control Modalities. However, for different modalites, there is no significant difference in Physical aspect, even though our data and analysis show significant difference on Error Rate and Workload.

Based on this study, we found that mid-air Hand-gesture based control method has the lowest performance score and highest workload in autonomous driving comparing to that of physical button and voice control interaction group. However, from the study, it is also found that Hand-gesture control perform well one certain task ("Open the map"). It is a possible conclusion because in some scenarios, hand-gesture has been proved to be well-performed. Micah Alpern, in 2003 has proved that hand-gesture control is good with some secondary tasks in driving and made even fewer mistakes, compared to physical buttons (radio button in the study) [46]. And, for tasks that requires a quick glance, gesture interaction have also shown better performance [46]. However, the study scenario in this study is not in a fully self-driving vehicle, which can be explained the difference of the study result of this study and Micah's work result. Furthermore, It is found that the Hand gesture interaction have a higher error rate in this study, which can lead to a future research of the different on Hand-gesture interaction in both manual driving vehicle and self-driving vehicle.

One the other hand, the study result has show very close performance and error rate of Voice control interface and Physical button interface. The error rate of voice control interface is slightly lower that that of physical button interface, but no significant difference was found in ANOVA test. When evaluation workload, voice control interface and physical button interface also have very close result, with the physical button interface lower workload. This also proved a study Murali in 2021 [15]

In the present investigation, the assessment of error rates in the distraction task has yielded inconclusive results. The obtained data does not exhibit a significant deviation, and the mean values remain notably similar. Possible contributing factors to these findings include a limited sample size (with the distraction task error rate data comprising only half of the total error rate data) and an insufficient range of evaluation parameters (such as interaction accuracy and human processing time, which were not recorded in this study). [47]

6.1 Limitations

The limitation of this study could be divided into two parts: Experimental limitation and System limitation.

6.1.1 Experimental Limitation

Firstly, Due to limitations of in lab user testing and time constraints, the control methods have not been tested with more complex features such as text input, adjusting AC, or switching to next music. The results may vary if the complexity of the features increases. And, the sample size for the in-lab user tests is likely not large enough to produce meaningful findings. Furthermore, the in-lab experiment duration could also be a limitation as it could be longer to better collect the mental demand, or physical demand data. And in this study, distraction was not counterbalanced as the participants are following the same procedure. This may also cause small bias of the test result. And for the physical button control method, seat manual control is usually not placed near the seat as in real cars. In this study, all buttons are on the steering wheel due to experiment lab setup. This may cause difference with actual daily scenario.

Secondly, the use of Virtual Reality simulations may also affect participants, as some may experience motion sickness. In this experiment, most participants did not report any symptoms of motion sickness in interviews and surveys, but one participant expressed feeling slight dizziness. However, compared to real-world driving experiments, simulation-based studies still have limitations and the physiological data may be impacted. Additionally, wearing Virtual Reality headset put weight on participants head, which may lead to discomfort which can affect participants' mental workload.

Finally, the simulation software was built from scratch using the Unity Engine by researchers. It is possible that the vehicle dynamics were not fully integrated into the simulator. For instance, the four wheels of the simulator vehicle were not separated but represented as a single game object with the body of the vehicle, which may hinder the vehicle's ability to drive under realistic dynamics constraints. Additionally, during turning, the rotation of the simulator vehicle may not be as accurate as in the real world due to limitations in wheel representation.

6.1.2 System Limitation

Apart from the experimental limitation, the experiment system design could also contains limitations.

Firstly, machine and computing system latency and error is not considered in this research. For mid-air hand gesture control and voice control, both control methods are heavily relied on sensor accuracy and computing system respond time. Since computer vision and Deep learning algorithms and models are highly involved in to the application, model performance may also affect input respond time and machine latency. All of those factors could affect user's performance and preference.

Secondly, in this study, the distraction was made by a pre-recorded audio, this distraction methods might be limited. Visual, haptic distraction could have different affect with participants' interaction, and may further affect the result of the research question.

6.2 Future Work

This research focuses on examining the different control methods in fully autonomous vehicles, leaving room for a multitude of future works to be conducted from various engineering perspectives. One area of future research could be to examine the impact of machine errors in the Human-Computer Interaction system. By combining the findings of this study with an analysis of machine performance, a more comprehensive understanding of the overall research question can be achieved. To more accurately simulate real-world use cases, future research could utilize more advanced technologies such as computer vision gesture recognition and natural language processing frameworks, rather than relying on the Wizard-of-Oz methodology employed in this study. Additionally, to address the limitations of considering only single-form distractions, future research could incorporate other forms of distractions such as visual, cognitive, and fatigue into the evaluation. Especially, when considering voice control interface, studying a multi-model distraction pattern is necessary. Furthermore, the focus of this research is on the evaluation of in-vehicle interface distractions and their impact on the user's current task. Trust is a crucial factor that must be taken into account in future studies. While trust in self-driving scenarios has been widely researched in the real world, this study does not consider it, as the participants were informed that the vehicle was driving in a safe condition. However, it is important to note that if participants were involved in a complex, uncertain situation, their mental and physical workload could be different from the findings of this study. Future studies should therefore consider incorporating a trust factor in their evaluations to more accurately reflect real-world scenarios and to enhance the reliability of the findings.

Furthermore, despite the possible future works mentioned above, more human factors details should be added to future work. For driver with mental disorder symptoms, such as Attention Deficit Hyperactivity Disorder (ADHD), their driving behavior is different from that of neurotypical person [48]. ADHD patient is characterized as person who has symptoms of hyperactivity, inattention, or impulsivity [48]. All of these characters could affect driver's behavior. In self-driving scenario, driving risks are significantly reduced for driver with ADHD. However, their prefered control method could be different due to the symptom mentioned above.

References

[1] Muhammad Iqbal, Jia Cheng Han, Zhi Quan Zhou, Dave Towey, and Tsong Yueh Chen. Metamorphic testing of advanced driver-assistance system (adas) simulation platforms: Lane keeping assist system (lkas) case studies. *Information and Software Technology*, 155:107104, 2023.

[2] Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles.

[3] Dario D. Salvucci. Rapid prototyping and evaluation of in-vehicle interfaces. *ACM Trans. Comput.-Hum. Interact.*, 16(2), jun 2009.

[4] Alessandro Oronzo Caffò, Luigi Tinella, Antonella Lopez, Giuseppina Spano, Ylenia Massaro, Andrea Lisi, Fabrizio Stasolla, Roberto Catanesi, Francesco Nardulli, Ignazio Grattagliano, and et al. The drives for driving simulation: A scientometric analysis and a selective review of reviews on simulated driving research. *Frontiers in Psychology*, 11, 2020.

[5] Brian Paden, Michal Cap, Sze Zheng Yong, Dmitry Yershov, and Emilio Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on Intelligent Vehicles*, 1(1):33–55, 2016.

[6] Lars Creutz, Sam Kopp, Jens Schneider, Matthias Dziubany, Yannick Becker, and Guido Dartmann. Simulation platforms for autonomous driving and smart mobility: Simulation platforms, concepts, software, apis. *Smart Transportation*, page 151–182, 2021.

[7] Andras Kemeny. From driving simulation to virtual reality. *Proceedings of the 2014 Virtual Reality International Conference*, 2014.

[8] Maria T. Schultheis, Jose Rebimbas, Ronald Mourant, and Scott R. Millis. Examining the usability of a virtual reality driving simulator. *Assistive Technology*, 19(1):1–10, 2007.

[9] E S Soegoto, R D Utami, and Y A Hermawan. Influence of artificial intelligence in automotive industry. *Journal of Physics: Conference Series*, 1402(6):066081, 2019.

[10] Daeuk Kim, Ryan Robin Mendoza, Karen Faye Chua, Mark Andrew Chavez, Ronnie S. Concepcion, and Ryan Rhay Vicerra. A systematic analysis on the trends and challenges in autonomous vehicles and the proposed solutions for level 5 automation. *2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, 2021.

[11] Nikolaos Gkikas. *Automotive ergonomics: Driver-vehicle interaction.* CRC Press, 2012.

[12] Tenghuan He, Luzheng Bi, Jinling Lian, and Huafei Sun. A brain signals-based interface between drivers and in-vehicle devices. *2016 IEEE Intelligent Vehicles Symposium (IV)*, 2016.

[13] Carl Jörgen Normark. Design and evaluation of a touch-based personalizable in-vehicle user interface. *International Journal of Human-Computer Interaction*, 31(11):731–745, 2015.

[14] Jung Min Lee and Da Young Ju. Classification of human-vehicle interaction: User perspectives on design. *Social Behavior and Personality: an international journal*, 46(7):1057–1070, 2018.

[15] Prajval Kumar Murali, Mohsen Kaboli, and Ravinder Dahiya. Intelligent in-vehicle interaction technologies. *Advanced Intelligent Systems*, 4(2):2100122, 2021.

[16] Driver distraction: A review of the current state-of-knowledge. *PsycEXTRA Dataset*, 2008.

[17] Wei Lien Liang, Michael Kyte, Fred Kitchener, and Patrick Shannon. Effect of environmental factors on driver speed: A case study. *Transportation Research Record: Journal of the Transportation Research Board*, 1635(1):155–161, 1998.

[18] Chongren Sun and Siby Samuel. Micro scenarios – a theory-based intervention to alleviate simulator sickness for older drivers in driving simulators. *Human Factors in Transportation*, 2022.

[19] Mark S. Young, Karel A. Brookhuis, Christopher D. Wickens, and Peter A. Hancock. State of science: Mental workload in ergonomics. *Ergonomics*, 58(1):1–17, 2014.

[20] Bin Xie and Gavriel Salvendy. Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments. *Work amp; Stress*, 14(1):74–99, 2000.

[21] Brian P. Bailey, Joseph A. Konstan, and John V. Carlis. The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In *IFIP TC13 International Conference on Human-Computer Interaction*, 2001.

[22] Pauzie A. A method to assess the driver mental workload: The driving activity load index (dali). *IET Intelligent Transport Systems*, 2(4):315, 2008.

[23] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. Wizard of oz studies — why and how. *Knowledge-Based Systems*, 6(4):258–266, 1993.

[24] Henrik Detjen, Bastian Pfleging, and Stefan Schneegass. A wizard of oz field study to understand non-driving-related activities, trust, and acceptance of automated vehicles. *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2020.

[25] V.D. Calhoun and G.D. Pearlson. A selective review of simulated driving studies: Combining naturalistic and hybrid paradigms, analysis approaches, and future directions. *NeuroImage*, 59(1):25–35, 2012.

[26] S.L. Comte and A.H. Jamson. Traditional and innovative speed-reducing measures for curves: An investigation of driver behaviour using a driving simulator. *Safety Science*, 36(3):137–150, 2000.

[27] D. Davenne, R. Lericollais, P. Sagaspe, J. Taillard, A. Gauthier, S. Espié, and P. Philip. Reliability of simulator driving tool for evaluation of sleepiness, fatigue and driving performance. *Accident Analysis amp; Prevention*, 45:677–682, 2012.

[28] Gerard J. Blaauw. Driving experience and task demands in simulator and instrumented car: A validation study. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 24(4):473–486, 1982.

[29] Rachael A. Wynne, Vanessa Beanland, and Paul M. Salmon. Systematic review of driving simulator validation studies. *Safety Science*, 117:138–151, 2019.

[30] Nico A. Kaptein, Jan Theeuwes, and Richard van der Horst. Driving simulator validity: Some considerations. *Transportation Research Record: Journal of the Transportation Research Board*, 1550(1):30–36, 1996.

[31] Tao Ni, G.S. Schmidt, O.G. Staadt, M.A. Livingston, R. Ball, and R. May. A survey of large high-resolution display technologies, techniques, and applications. *IEEE Virtual Reality Conference (VR 2006).*

[32] *Automotive user interfaces: Creating interactive experiences in the car.* SPRINGER INTERNATIONAL PU, 2018.

[33] Shi Cao, Krithika Nandakumar, Raiju Babu, and Benjamin Thompson. Game play in virtual reality driving simulation involving head-mounted display and comparison to desktop display. *Virtual Reality*, 24(3):503–513, 2019.

[34] S. Debernard, C. Chauvin, R. Pokam, and S. Langlois. Designing human-machine interface for autonomous vehicles. *IFAC-PapersOnLine*, 49(19):609–614, 2016.

[35] Udara E. Manawadu, Mitsuhiro Kamezaki, Masaaki Ishikawa, Takahiro Kawano, and Shigeki Sugano. A hand gesture based driver-vehicle interface to control lateral and longitudinal motions of an autonomous vehicle. *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016.

[36] Seul Chan Lee, Harsh Sanghavi, Sangjin Ko, and Myounghoon Jeon. Autonomous driving with an agent. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*, 2019.

[37] Oculus developer center: Downloads.

[38] Christopher D. Wickens. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2):159–177, 2002.

[39] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in Psychology*, page 139–183, 1988.

[40] Susana Rubio, Eva Diaz, Jesus Martin, and Jose M. Puente. Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology*, 53(1):61–86, 2004.

[41] Malik Haris and Adam Glowacz. Navigating an automated driving vehicle via the early fusion of multi-modality. *Sensors*, 22(4):1425, 2022.

[42] Mark Draper, Gloria Calhoun, Heath Ruff, David Williamson, and Timothy Barry. Manual versus speech input for unmanned aerial vehicle control station operations. *PsycEXTRA Dataset*, 2003.

[43] D.N. Card. Defect-causal analysis drives down error rates. *IEEE Software*, 10(4):98–99, 1993.

[44] Jun Ma and Yuchun Du. Study on the evaluation method of in-vehicle gesture control. *2017 3rd IEEE International Conference on Control Science and Systems Engineering (ICCSSE)*, 2017.

[45] Henrik Detjen, Sarah Faltaous, Bastian Pfleging, Stefan Geisler, and Stefan Schneegass. How to increase automated vehicles' acceptance through in-vehicle inter-

action design: A review. *International Journal of Human–Computer Interaction*, 37(4):308–330, 2021.

[46] Micah Alpern and Katie Minardo. Developing a car gesture interface for use as a secondary task. *CHI '03 extended abstracts on Human factors in computing systems - CHI '03*, 2003.

[47] Nicolas Saunier, Tarek Sayed, and Karim Ismail. Large-scale automated analysis of vehicle interactions and collisions. *Transportation Research Record: Journal of the Transportation Research Board*, 2147(1):42–50, 2010.

[48] Laurence Jerome, Liat Habinski, and Alvin Segal. Attention-deficit/hyperactivity disorder (adhd) and driving risk: A review of the literature and a methodological critique. *Current Psychiatry Reports*, 8(5):416–426, 2006.

Appendix

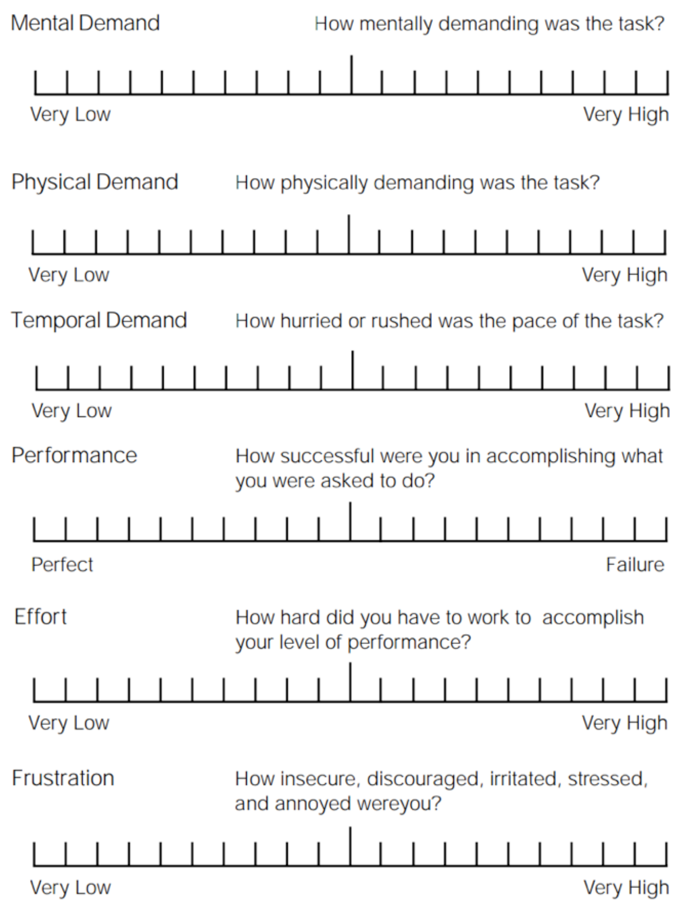


Figure 1: NASA-TLX Questionnaire