EXPLORING THE IMPACT OF IMMERSION ON SITUATIONAL AWARENESS AND TRUST IN REMOTELY MONITORED MARITIME AUTONOMOUS SURFACE SHIPS

ALEXANDER GREGOR

A THESIS SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

GRADUATE PROGRAM IN
ELECTRICAL ENGINEERING & COMPUTER SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO

JANUARY 2023

Abstract

This thesis examines how Situational Awareness (SA) and Trust, along with some exploratory variables, were affected by different immersion levels in maritime remote monitoring. To examine this a simulated Shore Control Centre (SCC) interface for Maritime Autonomous Surface Ships (MASS) was constructed, which had an autonomous container ship traversing the arctic with robotic aids. Three query sets were asked per simulation run, which facilitated tracking how SA, Trust, and Motion Sickness (MS) evolved over time. Three different virtual reality (VR) interfaces were used; Non-Immersive VR (NVR), Semi-immersive VR (SVR), and Immersive VR (IVR). The simulation and query sets were performed on a counterbalanced within-subjects user study with 39 participants. The results illustrated various trade-offs - with NVR showing higher user preference, SVR showing signs of higher SA, and IVR showing moderate Trust but increased MS. Understanding these trade-offs between immersion levels is a requisite step for designing future SCCs.

Acknowledgements

I would like to thank my supervisor Professor Robert Allison, whose guidance and support has been invaluable. Additionally, I would also like to thank Professor Onoise Kio, Faruq Afolabi, as well as the rest of the Virtual Reality & Perception Lab, for their constant support. Finally, I would like to thank the National Research Council of Canada for financially supporting this research.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI . . . . . Artificial Intelligence

ANOVA . . . . . Analysis of Variance

AR . . . . . Augmented Reality

BC . . . . . Box-Cox Transformation

C-ARS . . . . . Camera's Average Rotational Speed

C-NVSC . . . . . Camera's Number of Vantage State Changes

C2 . . . . . Command and Control

CAVE . . . . . Cave Automatic Virtual Environment

COLREGs . . International Regulations for Preventing Collisions at Sea

DrivingYr . . . . . Driving experience, in years

DV . . . . . Dependent Variable

ECDIS . . . . . Electronic Chart Display and Information System

FMS . . . . . Fast Motion Sickness Scale

FOV . . . . . Field-of-View

FullNO . . . . . Full data set, with no outliers

GamesWk . . . . . Video games played, in days per week

GPU . . . . . Graphics Processing Unit

HMD . . . . . Head-mounted display

IAI . . . . . Intelligent Adaptive Interface

IML . . . . . Interpretable Machine Learning

IMO . . . . . International Maritime Organization

IV . . . . . Independent Variable

IVR . . . . . Immersive Virtual Reality

LiDAR . . . . . Light Detection and Ranging

LSD . . . . . Least Significant Difference

MASS . . . . . Maritime Autonomous Surface Ships

MR . . . . . Mixed Reality

MS . . . . . Motion Sickness

MWL . . . . . Mental Workload

NASA . . . . . National Aeronautics and Space Administration

NASA-TLX . . . . . NASA's Task Load Index

NRC . . . . . National Research Council of Canada

NRC-OCRE NRC's Ocean, Coastal and River Engineering Research Centre

NVR . . . . . Non-immersive Virtual Reality

PassedNO . . . . . Passed data subset, with no outliers

Q1 . . . . . Query Set #1

Q2 . . . . . Query Set #2

Q3 . . . . . Query Set #3

RAM . . . . . Random Access Memory

ROC . . . . . Remote Operation Centre

RTLX . . . . . Raw Task Load Index

SA . . . . . Situational Awareness

SA-AT . . . . . Situational Awareness - Activation Time

SA-PC . . . . . Situational Awareness - Percent Correct

SA-RCS . . . . . Situational Awareness - Rate Correct Score

SA-RT . . . . . Situational Awareness - Response Time

SAGAT . . . . . Situation Awareness Global Assessment Technique

SCC . . . . . Shore Control Centre

SCCO . . . . . Shore Control Centre Operator

SE . . . . . Standard Error

SME . . . . . Subject-matter Expert

SPAM . . . . . Situation Present Assessment Method

SSD . . . . . Solid State Drive

SUS . . . . . System Usability Score

SVR . . . . . Semi-immersive Virtual Reality

UAV . . . . . Unmanned Aerial Vehicle

UGV . . . . . Unmanned Ground Vehicle

URPP . . . . . Undergraduate Research Participant Pool

US . . . . . Usability

USV . . . . . Unmanned Surface Vehicle

VR . . . . . Virtual Reality

XAI . . . . . Explainable Artificial Intelligence

YJ . . . . . Yeo-Johnson Transformation

Chapter 1

Introduction

As automated vehicles and robots become increasingly ubiquitous, there has been a push from shipping solutions providers to extend these technologies to maritime shipping. This is known as Maritime Autonomous Surface Ships (MASS), and has been a growing area of research over the past decade with potential benefits of mitigated environmental impact, reduced risk for seafarers, and economic efficiency (Dybvik et al., 2020; Yara, 2021). Various projects have made strides designing MASS systems including the Munin Project, ReVolt, and the Yara Birkeland (Munim, 2019). The latter has made the most progress, with manned pilot voyages being carried out across the Oslofjord in 2021 and 2022, and un-crewed commercial usage expected to begin in late 2023 (Kongsberg, 2022).

Autonomous vessels can have various levels of automation, exemplified by the four Degrees of Shipping Automation defined by the International Maritime Organization (IMO) (IMO, 2021). These range from partial automation of manned ships at Degree 1, to fully autonomous vessels with no human intervention at Degree 4. Degree 3 is the major goal of many research groups,

since this technology is deemed plausible in the near future (Dybvik et al., 2020). Degree 3 vessels would rely on autonomous navigation, but still have a remote human-in-the-loop to make course impacting decisions and take control if necessary. This remote monitoring and control would occur at a Shore Control Centre (SCC), sometimes called a Remote Operation Centre (ROC), which are command centres for autonomous MASS systems (Van Den Broek et al., 2020; Kongsberg, 2022). The humans that will monitor these systems are typically given the title of SCC Operator (SCCO) in the literature (Saha, 2021). Key factors when building interfaces for future MASS systems are the need to optimize Situational Awareness (SA) and manage Trust (Heffner and Rødseth, 2019; Dybvik et al., 2020). This is because they serve to mitigate risk when properly accounted for (Thieme and Utne, 2017).

Endsley defines SA as "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1995; Loft et al., 2013; Nguyen et al., 2019). This relates to the concept of mental models, wherein a user has an internal conceptualization of the environment that can vary from the ground truth. As such, an SCCO would always want a relatively high SA. Trust is a more complex concept, and instead of a higher level, requires calibration when dealing with automation (Lee and See, 2004). Too little Trust and an SCCO might intervene too much, which wastes time and resources and can be dangerous if an SCCO has an inaccurate mental model. Conversely, too much Trust can be just as dangerous, as an over-reliant SCCO may not notice signs that the system is malfunctioning or in danger.

One factor that has the potential to impact these constructs is a user's sense of immersion, which is a concept associated with virtual reality (VR)

and mixed reality (MR) technologies. Immersion involves constraining a user's perceptual experience to a virtual or remote environment (Rizzo and Koenig, 2017). As such, increased visual peripheral occlusion, naturalistic control, haptic feedback, and focused auditory stimuli correlate to a greater sense of immersion. Since immersion depends on varying degrees of engagement with numerous senses, it is not a binary concept, and can have differing levels. For practicality, the immersion levels herein are categorized and referred to as Non-immersive Virtual Reality (NVR), Semi-immersive Virtual Reality (SVR), and Immersive Virtual Reality (IVR), and serve as the main Independent Variables (IVs) in this thesis.

NVR, sometimes called Low-immersion VR, involves interacting with a virtual environment using limited sensory immersion (Rizzo and Koenig, 2017; Martirosov et al., 2021). As such, playing a video game on a TV would constitute NVR. The function of NVR in the present study is to see what a basic teleoperation interface would look like. SVR, sometimes called Medium-immersion VR, is ambiguously defined in the literature, but typically involves using large or concave displays that semi-occlude the user's peripheral vision (Pollard et al., 2020; Martirosov et al., 2021). In SVR, camera movement around the virtual environment can be accomplished through a secondary medium like a joystick. SVR also often uses earphones, or speakers, for controlled audition. Finally, IVR, sometimes called High-immersion VR, uses Head-mounted Displays (HMDs), which occlude peripheral vision and use sensors to map head movement to the rendered vantage point in the virtual environment (Pollard et al., 2020). Other sensory tools are often utilized to increase immersion, such as noise-cancelling headphones for controlled audition, and haptic controllers for proprioceptive feedback.

This thesis sought to examine how SA and Trust are affected by different immersion levels in SCC interfaces, and in turn how these cognitive assessments correlate to each other. For exploratory purposes, a small battery of supplementary cognitive tests and behavioural measures were added as well. This included Motion Sickness (MS), Mental Workload (MWL), Usability (US), Camera's Average Rotational Speed (C-ARS), and Camera's Number of Vantage State Changes (C-NVSC). All of these measures serve as Dependent Variables (DVs) in this thesis, and their specifics are discussed in Chapter 3.

This is a complicated area of research to explore due to the lack of operational SCCs, let alone those employing MR interfaces. Considering this, a VR simulation of a MASS system and SCC interface was constructed and used to mimic how these systems will navigate their environment and how the SCCO will receive navigational and sensory data. As such, in this study VR is used as a proxy for future SCC MR interfaces.

The simulation scenario involves a MASS container ship and its robotic aids traveling through the Canadian arctic. The lack of appropriate telecommunications infrastructure in remote areas of Canada's arctic will pose one of the most extreme bandwidth-limited scenarios a MASS vessel is likely to encounter. Additionally, the MASS vessel was tasked with navigating around floating ice. This simulation assumed an IMO automation level of 3, so the participant who is acting as the SCCO was tasked with remotely monitoring the container ship's progress. Having different immersion levels helped to determine if there is a statistically supported benefit to an operator's SA and Trust in the utilization of immersive MR equipment in SCCs.

Chapter 2

Related Work

I am unaware of any research focusing on VR or MR interfaces for the teleoperation or remote monitoring of MASS systems. As such, comparable areas of research are discussed to piece together the state of the literature.

2.1 Comparing Levels of Immersion

Several research groups have explored the impact of immersion levels. For example, Pollard et al. (2020) used a within-subjects study using 61 participants to examine how learning differed between levels of immersion. The IVs in this instance were the level of immersive technology used, described as Low, Medium, High. Respectively, these were a desktop monitor with speakers, a NVIS HMD with supra-aural headphones, and an Oculus Rift HMD with circum-aural headphones. Their study used various performance metrics related to learning and memory, and found that in general the higher level of immersion provided better learning performance - and suggested worse learning performance for the medium immersion level. This not only illustrates an example of an immersion

level comparison, but also shows that different immersive technologies can uniquely impact various human factors variables. They also argue for the use of within-subjects study designs when doing VR research.

Similarly, Martirosov et al. (2021) performed a user study on 89 participants examining how non-immersive, semi-immersive, and immersive VR affect cyber sickness. Herein, they used a desktop monitor for non-immersive VR, a Cave Automatic Virtual Environment (CAVE) system for semi-immersive VR, an Oculus Rift for immersive VR, and a physical task for the control group. The study showed immersive VR led to very high cyber sickness, with numerous participants unable to complete the full 10 minutes of the study. Here, semi-immersive VR also showed increased cyber sickness, but to a lesser extent than immersive VR. This illustrates how cyber sickness, or MS, is thought to increase as immersion level increases.

These studies show how comparing immersion levels can be useful for understanding immersion's effect on cognitive constructs. However, there appears to be no research of this form when it comes to Trust or SA for teleoperated automation in non-line-of-sight scenarios.

2.2 VR & MR for Autonomous Teleoperation

While it appears no research group has tested the use of MR teleoperation for MASS systems, there are several papers that discuss the utilization of MR teleoperation for robotics. To understand the requirements for remote operators, it is necessary to examine the field of teleoperation. This research examines the area of interface design for teleoperated robots, which can often incorporate MR. For example, Smolyanskiy and Gonzalez-Franco (2017) out of

Microsoft examined how first-person VR perspectives during Unmanned Aerial Vehicle (UAV) flight can aid in navigation. Herein, they built a VR interface for a drone, and then tested it on 7 participants. They then utilized video recordings of these tests to create a VR simulation that they used to separately test the effects of stereo vision and the prevalence of MS in teleoperation in two user studies. This work is useful for two reasons, the first being that it shows that simulations have been used in private industry and academia for testing immersive interfaces for teleoperated robotics. Another reason is that the results showed relatively low levels for MS, which indicates that MS in VR teleoperation is not necessarily a given.

Other groups like Naceri et al. (2021) explored the use of a VR HMD as an interface for a robotic arm to improve SA. This group designed a VR interface for a UR5 robot arm, and then performed a between-subjects user study on 24 participants that used completion time and performance as DVs. The IV conditions in this case were a desktop setup, VR without camera teleporting, and VR with camera teleporting. This concept of teleporting involved implementing multiple cameras as freely navigable viewpoints - a concept that directly influenced this study. Notably, out of 3 IVs VR with teleporting had the best performance, illustrating a potential benefit to this approach. The study also gave an architectural overview of connecting the Unreal game engine with ROS, which could prove useful for future research.

Additionally, groups like Fabris et al. (2021) explored how immersive teleoperation can impact SA. This group created two interfaces, one a traditional computer monitor interface, and the other a computer monitor with an Augmented Reality (AR) headset to augment the user's view - which they refer to as traditional and immersive stations, respectively. They also created a

simulation of a teleoperated Unmanned Ground Vehicle (UGV) that they tested on 11 participants using a within-subjects user study. To assess SA, they used the Situation Awareness Global Assessment Technique (SAGAT) - a freeze probe technique - as well as various performance metrics. Their work showed very little difference for SA between their station interfaces.

These studies demonstrate different ways of designing immersive interfaces for teleoperated robotics, which is useful for narrowing down which features should be explored for a MASS SCC.

2.3 MASS Systems

Several groups have examined the various concerns and design strategies that will have to be considered before MASS systems and SCCs can be developed. Peeters et al. (2020) examined how human monitoring at an SCC might occur for USVs that are large enough to transport cargo. This study is relevant since it highlights the benefits of creating an Inland SCC, and also the requirements to create such systems. Even though they don't explicitly study SA, they also note its importance for remote monitoring stations.

Felski and Zwolak (2020) examined the threats that will be posed for future MASS SCCs. This paper was important for contextualizing risk and threat analysis - which highlighted concepts important for analysis when designing an SCC, such as SA and Trust.

Namgung and Kim (2021) analyzed how MASS systems would need to comply with the International Regulations for Preventing Collisions at Sea (COLREGs) navigational rules laid out by the IMO. This outlined the general system requirements for MASS, and specifically for understanding how ship

avoidance would have to be designed for these autonomous systems so as to mitigate risk.

Yoshida et al. (2020) and Saha (2021) both examined the exact responsibilities of SCCOs, their required background, and the necessary training to prepare them for the role. Yoshida et al. outlined the regulatory needs for MASS operation such as legal gaps, operator training, and operator certification which was important to understand what kind of person would be an SCCO. They suggest that SCCOs, or remote operators as they are called, would be a person of a seafaring background with supplementary education in electronic navigation and autonomous technologies. Secondly, they indicate the importance of SA and understanding how the various levels as defined by Endsley (1995) can be accounted for. Moreover, they indicate the importance of Trust for remote operation. Saha on the other hand focused on examining core competencies for future SCCOs, which include systems understanding, technical communications knowledge, and maritime competence. Both of these illuminate the general requirements for an SCCO.

As a last note, companies such as Rolls-Royce and Kongsberg have been making active strides to construct SCCs (Rolls-Royce, 2016; Kongsberg, 2022). Knowledge of these are sparse due to the proprietary nature of the technology, but highlight how this is an active area of industry research.

2.4 Human Factors for Control Centres

There are multiple factors to consider when designing teleoperated, or remotely monitored, interfaces, such as communications bandwidth and the type of data available. One type of data is a video stream. These are informative, but are

bandwidth-intensive, especially when operating at high latitudes with limited satellite coverage. Importantly, video quality can be reduced without having a large impact on the SA of SCCOs as shown by Yoshida et al. (2021) in their study of how MWL would affect remote operators of MASS systems. In addition to video feedback, some degree of navigational information would be available for control centres.

Marusich et al. (2016) performed multiple Command and Control (C2) simulation user studies to analyze how varying levels of task relevant information can affect SA and Trust in C2 environments. They note how too much information, even if task-relevant, can negatively impede SA and Trust. This emphasizes the need to organize these systems around cognitive processes, and to emphasize simple and intuitive design. Much like the findings of Yoshida et al. (2021), this is beneficial in situations where a control centre has poor bandwidth connection or high latency, since it means design can focus on conveying the core navigational information.

Cunningham et al. (2015) performed a within-subjects user study on 35 pilots to assess SA in an aviation ground control simulation. Their work focused on why Situation Present Assessment Method (SPAM) queries are sometimes left unanswered by participants, but notably they used a multiple query set model. Herein, different SPAM probes were asked over the course of a test in 2-4 minute intervals, with the participant acting the part of various control room roles for the different conditions. This study was valuable for two reasons. First, it lends credence to the idea of using multiple SPAM queries, or probes, in a single test. Second, it shows the use of SPAM queries that mimic remote operators, which is an important concept for creating a believable role-based simulation.

Finally, the group of Loft et al. (2016) performed various within-subjects user studies to assess how uncertainty affects operator SA, MWL, and performance, for remote track management of submarines in combat scenarios. They used SPAM to assess SA, and the National Aeronautics and Space Administration's (NASA) Task Load Index (NASA-TLX) to assess MWL. By manipulating uncertainty in submarine locations, they illustrated how increased uncertainty can negatively impact SA and MWL. Like Cunningham et al. this shows SPAM being used to emulate a role-based scenario, but importantly it demonstrates its utility in a maritime setting in addition to the aviation scenarios it was originally developed for. Additionally, this concept of uncertainty was useful since part of what this thesis analyzed was randomized ice obstacle avoidance. Thus, understanding how uncertainty could impact SCCO SA was important for designing the ice distribution system that will be discussed in the next section.

The research discussed in this chapter illustrates how a lot of pieces required to study human factors in MASS systems are there to build on, but that there are still significant gaps in the literature - the most obvious of which being a lack of focused user study on MR for MASS systems. Additionally, there is a gap in research exploring whether immersive systems can impact SA and Trust in maritime teleoperation scenarios. The goal of this research was to build upon the works discussed to try and bridge parts of these gaps in the literature.

Chapter 3

Methods

This section will cover the methods used in the research. Specifically, this section covers the participants enlisted, the system apparatuses, the simulation, the study procedure, the study's design, and the methods of statistical analysis used.

3.1 Participants

Thirty-nine participants were enlisted through two main sources, the first source being participants from York University's Undergraduate Research Participant Pool (URPP). These students comprised 28 of the total number, and were compensated with course credit. The other 11 of participants were local volunteers.

The majority of participants were undergraduate students, which is perhaps unsurprising given how many people were enlisted via the URPP. These students came from a wide range of majors. Of the volunteers, three were graduate students from the Virtual Reality & Perception Lab, and the rest of the

participants were working professionals. The breakdown of these occupational backgrounds can be seen in Figure 3.1.

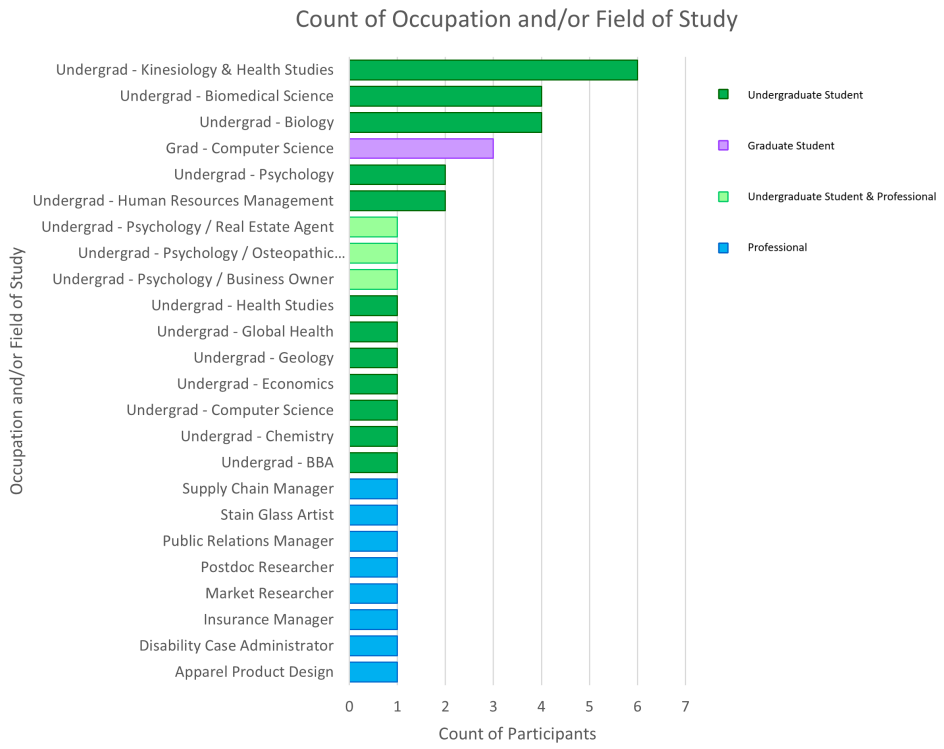


Figure 3.1: Histogram showing the breakdown of participant occupations.

The age of participants ranged from 18 to 61, with the majority of participants being in their twenties. Additionally, for the gender breakdown, 28 participants identified as women, and 11 identified as men. A histogram of participant ages with their respective gender identities can be seen in Figure 3.2.

3.2 Apparatus

A laptop was utilized for running all of the simulations and gathering output data. This was a DELL G15 with Windows 11, 16 *GB* Random Access Memory

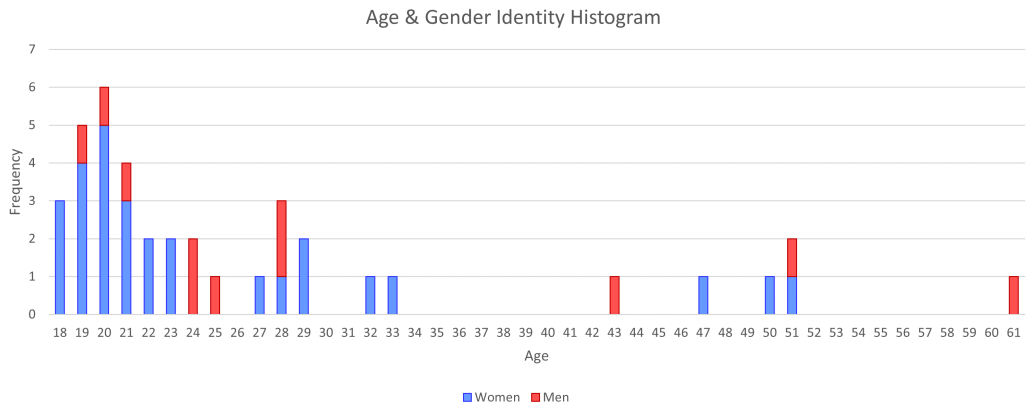


Figure 3.2: Histogram of participant ages with gender identity breakdown.

(RAM), 1 TB Solid State Drive (SSD), an Intel Core i7 Processor, and a Nvidia Geforce RTX 3060 Graphics Processing Unit (GPU).

3.2.1 NVR Setup

The NVR immersion level - an example of which can be seen in Figure 3.3 - used a large TV with built-in speakers, a secondary monitor, a keyboard, and a wireless mouse. The TV showed the various panoramic camera perspectives, which were navigable with the W(steer up), A(steer left), S(steer down), D(steer right) keys on the keyboard. This meant that users could affect the camera display's yaw and pitch - however the roll was locked to be level with the horizon. Horizontal movement included the full 360° range, while the vertical movement was locked to a 150° range, which was to avoid issues with camera flipping. The visual details for the TV were as follows:

- Dimensions (Width x Height): 123 cm x 70 cm
- Field-of-View (FOV): 90°
- User Distance from Focal Point: 70 cm

- Resolution (Constrained): 1280 x 768
- Luminance (Box Measurements; calculated by fixating on the central static focal point of two different types of light contrast test background using a Minolta LS-100 Luminance Meter):
  - White Box (10% of screen) on a Black Background: 62.270 cd/m^2
  - Black Box (10% of screen) on a White Background: 0.728 cd/m^2

The second display showed the console and query set interfaces, which could be interacted with using the mouse. The second display blocked roughly 10% of the participant's camera FOV when looking directly forward. This was an unfortunate necessity to maintain continuity with the IVR setup, which would invariably have some level of obstruction from its game space console. The details for the secondary monitor were as follows:

- Dimensions: 38 cm x 31 cm
- User Distance from Focal Point: 45 cm
- Angle of the Secondary Monitor (Relative to the Table): 80°
- Luminance (Box Measurements):
  - White Box on a Black Background: 78.490 cd/m^2
  - Black Box on a White Background: 4.472 cd/m^2

3.2.2 SVR Setup

The SVR immersion level - an example of which can be seen in Figure 3.4 - used an eLumens concave projector display, a secondary monitor, a joystick,

Figure 3.3: A participant using the NVR immersion level

a wireless mouse, and supra-aural headphones. The eLumens display showed the various panoramic camera perspectives, which were navigable via joystick. Similarly to NVR, the users could affect the camera display's yaw and pitch - however the roll was locked to be level with the horizon. Horizontal movement included the full 360° range, while the vertical movement was locked to a 150° range, which was again to avoid issues with camera flipping. The details for the eLumens display were as follows:

- Dimensions: 165 cm x 160 cm
- FOV: 90°
- User Distance from Focal Point: 100 cm

- Resolution (Constrained): 1280 x 768
- Luminance (Box Measurements):
  - White Box on a Black Background: 8.470 cd/m^2
  - Black Box on a White Background: 1.720 cd/m^2

The second display showed the console and query set interfaces, which could be interacted with using the mouse. The second display blocked roughly 10% of the participant's camera FOV when looking directly forward. This was again to maintain continuity with the IVR setup. The details for the secondary display were as follows:

- Dimensions: 38 cm x 31 cm
- User Distance from Focal Point: 45 cm
- Angle of the Secondary Monitor: 45°
- Luminance (Box Measurements):
  - White Box on a Black Background: 47.700 cd/m^2
  - Black Box on a White Background: 1.028 cd/m^2

3.2.3 IVR Setup

The IVR immersion level - an example of which can be seen in Figure 3.5 - used a Meta Quest 2 HMD[1] in Rift mode, one haptic controller, and circum-aural headphones. It should be noted that there were numerous types of higher

[1]https://www.meta.com/ca/quest/products/quest-2/

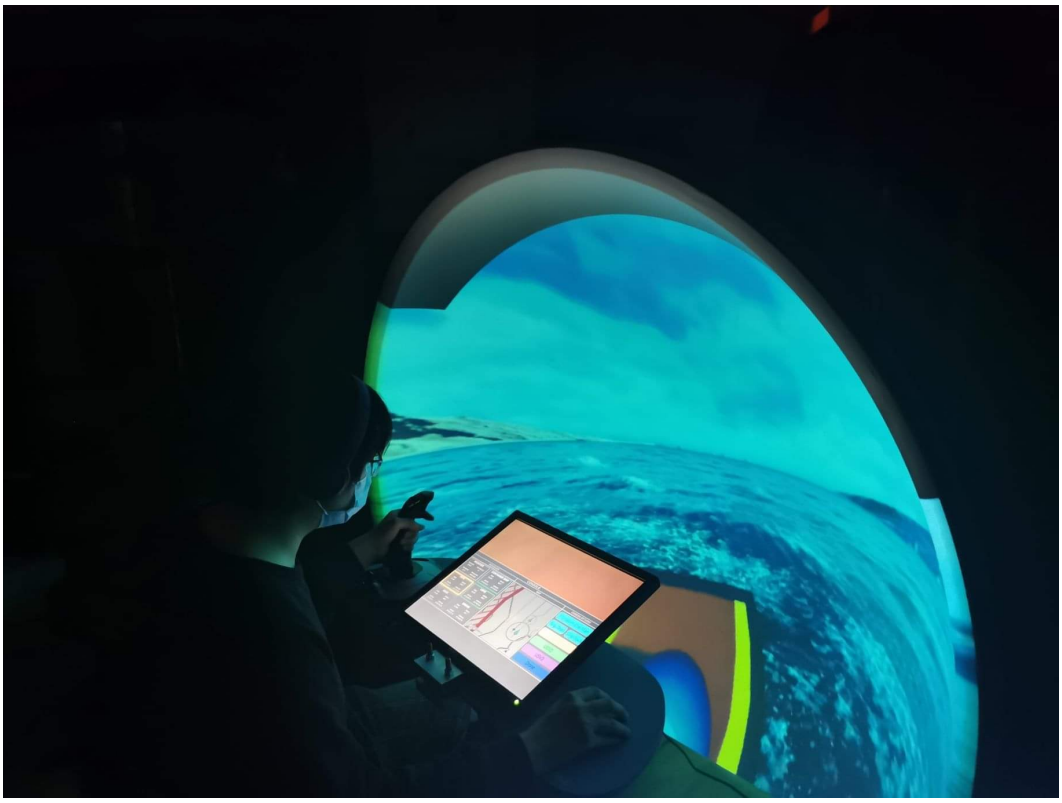


Figure 3.4: A participant using the SVR immersion level

quality IVR systems that could have been used instead - such as the Pro Series HMDs developed by Vive[2]. Additionally, Imagine4D have non-HMD immersive display systems like the Station IX[3], which could perhaps have produced very different results. However, IVR is consistently represented by HMDs in the literature, so I opted not to break with that - and the Quest 2 HMD was chosen since it was the most readily available to myself during the height of the COVID-19 pandemic when the simulation was being developed. The HMD was connected to the laptop via USB-C cord with a 10 *Gbps* bandwidth. Within the HMD the horizontal and vertical movement enabled the participant to see the full spherical range of the video feedback. The details for the HMD display

[2]https://www.vive.com/ca/product/
[3]https://imagine-4d.com/multimmersion/

were as follows:

- FOV: 90°
- Resolution:
  - Panoramic Background (Constrained): 1280 x 768
  - Interfaces and Remote: 1920 x 1832
- Luminance (Box Measurements):
  - White Box on a Black Background:
    - Left Eye: 38.340 cd/m^2
    - Right Eye: 36.170 cd/m^2
  - Black Box on a White Background:
    - Left Eye: 10.730 cd/m^2
    - Right Eye: 10.310 cd/m^2

Connecting to the laptop allowed for a consistent frame rate and the simultaneous ability to display the simulation to the laptop for observer monitoring. Participants could look around the panoramic camera view with their head motion, and could use their haptic controller to interact with the console and query set interfaces - which were part of the rendered world space in this instance. The console blocked roughly 10% of the participant's camera FOV when looking directly forward. Additionally, audio was stereo and directional, coming from the centre of the query set interface. This allowed participants to quickly orient themselves towards the query set interface if they were facing away from the console.

Figure 3.5: A participant using the IVR immersion level
Left: Full view of the IVR setup, Right: The view of the IVR immersion level from the observer's vantage point.

3.3 Simulation

The simulation was built by the Virtual Reality and Perception Lab at York University, and further customized to accommodate this study. All software was designed inside of the Unity game engine, version 2020.3.12f1. Additionally, the High Definition Render Pipeline was used for emulating naturalistic lighting. Several packages were used for creating the simulation including the Oculus

Integration[4], Crest Ocean System[5], and the A* Pathfinding Project[6]. Finally, all response and reaction time data were automatically stored in an external time-stamped CSV file for each immersion level.

The simulation is set in the Davis Strait at approximately 70°0'N 70°0'W. The scenario involved an autonomous shipping vessel making its way from the Eastern entrance of the Northwest Passage out to the Western entrance. In total, there were 5 autonomous vessels in the simulation:

- One Container Ship, ice class 1A (Canada, 2017)
- Three Unmanned Surface Vehicles (USVs)
- One Drone / UAV

Each of these had one panoramic camera affixed to it, except the container ship which had three cameras; one on the main observation deck, one on the ship's bow, and one on the ship's stern. This made for a total of seven freely navigable camera perspectives to select and use.

All immersion levels had a monoscopic panoramic video feedback with 1024 x 768 resolution and a 20 Hz frame rate to emulate a limited bandwidth. The exact bandwidth that will be available in the Canadian Arctic over the next few decades is unclear. The Canadian Government has invested in a satellite project to enhance non-line-of-sight communications in the Arctic, but that project is not slated to be finished until 2031 (Canada, 2020). However, even if this managed to finish on time, it can be assumed non-line-of-sight bandwidth would still remain fairly limited - especially for civilian and commercial uses.

[4]https://developer.oculus.com/downloads/package/unity-integration/
[5]https://crest.readthedocs.io/en/stable/
[6]https://arongranberg.com/astar/

As such, the implementation of a 20 Hz frame rate helped to instill a sense of limited communications bandwidth that is likely to be present in any initial Arctic SCC.

Navigational information for the simulation was available on the console in front of the user, as well as the electronic chart display and information system (ECDIS) map that showed all vessel locations, path plans, paths travelled in the last 2 minutes, vessel domains, known obstacles, and projected location in 2 minutes. These data all had a 0.25 Hz refresh rate, which served the dual purpose of emulating limited communications bandwidth and improving legibility. The console also included weather information, such as wind speed, wind direction, temperature, and weather - however, these were functionally static in every simulation. An example of this display can be seen in Figure 3.6.

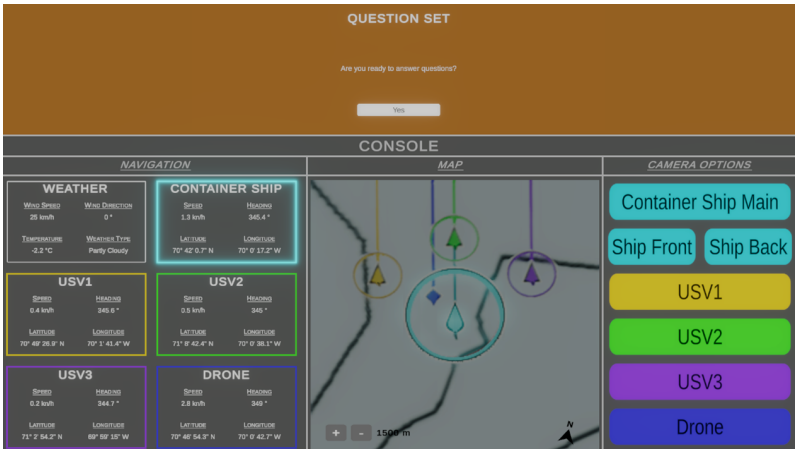


Figure 3.6: An example of the console used in the user study.

Pathfinding was implemented using an A* grid system with 40 m^2 squares. The grid was 900 x 120 squares to mimic the scaled dimensions of the National Research Council of Canada's (NRC) Ocean, Coastal and River Engineering Research Centre (NRC-OCRE) ice tank discussed by Gash et al. (2020), as

it may be used for testing in future research (Canada, 2022). All vessels operated independently and collectively shared sensor data to make path adjustments. All vessels had their vertices modeled as a rigid body for ocean interaction, but accurate hydrodynamics were not used for any of the vessels in the simulation. This is a feature that the research group intends to implement in future iterations of the simulation but was not required for this stage of human factors analysis.

The vessels emulated the detection of hazardous ice through a spheroidal contact flags with a radius of 250 m for the USVs, and 500 m for the container ship, intended to imitate the range of Light Detection and Ranging (LiDAR) sensors – a roughly 100 m radial berth was given to all detected ice. A Reciprocal Velocity Obstacle (RVO) modifier was used on each of the waterborne vessels to avoid inter-vessel collision. Like the LiDAR sensors, these RVO modifiers used radial domains of 250 m for the USVs, and 500 m for the container ship - upon interaction between these domains the respective A* path plans would adjust to move out each others way, with precedence being given to the container ship.

The simulation used a mixture of sea ice and land ice that are typically present in the Northwest Passage during the summer melt, so a "Bergy Water" distribution of $< 1/10^{th}$ ice coverage was utilized (Canada, 2005). For placement, the A* grid was broken down into sub-quadrants of 500 m^2. Prefabricated ice objects were placed in randomized sub-quadrants. This avoided objects overlapping at run-time. An example of this distribution can be seen in Figure 3.7. A caveat to the system is that icebergs in neighbouring blocks cannot be placed in squares directly next to each other. This was to avoid unnatural-looking linear arrangements.

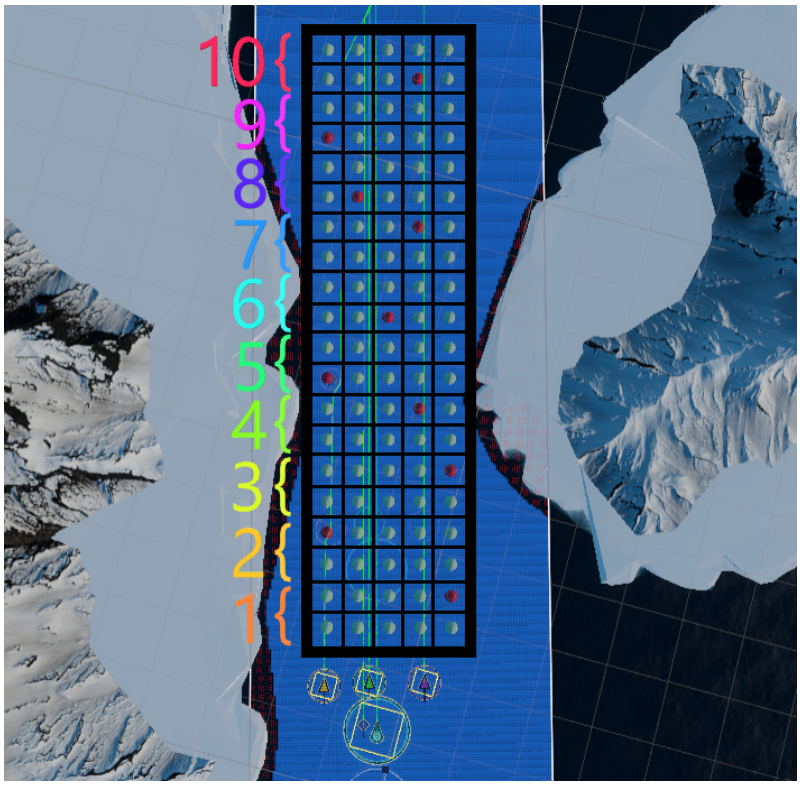


Figure 3.7: An example of the ice distribution system.
Each red dot represents an iceberg that was randomly placed in one of the ten 2 x 5 sub-quadrants at runtime.

Iceberg distribution was further randomized by assigning each iceberg one of 5 possible prefabricated iceberg shapes based on real-world shape distributions. It should be noted that the metrics used were from Romanov et al.'s observed iceberg shape distributions in the Antarctic ocean over the course of 52 years (Romanov et al., 2011). This is not an exact match to the iceberg shape distribution one could expect in the Canadian Arctic, but serves as a rough correlate in this instance. The iceberg shapes and their relative distributions can be seen in Figure 3.8. Notably, this attempt at creating a high degree of realism did not extend to relative sizes. Given the focus of this thesis being on how human factors variables are impacted by immersion level, a high-fidelity

model of iceberg size distribution was not developed. Instead, all of the iceberg models had varying degrees of size built into their prefabs, which allowed for sense of size change for the user.

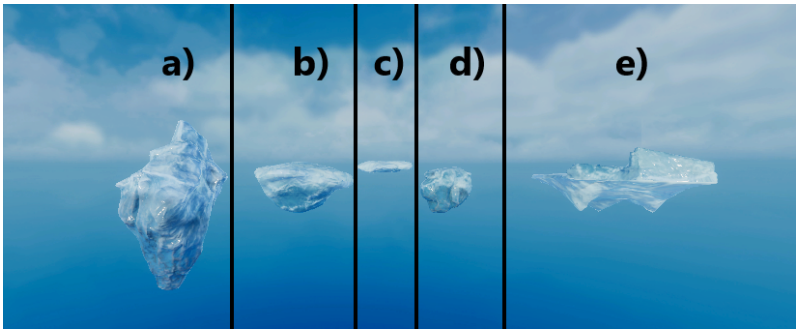


Figure 3.8: A picture of the iceberg prefabs that were used.
a) Pinnacle Iceberg, Dist: 18%; b) Tabular Iceberg, Dist: 24%; c) Weathered Iceberg, Dist: 24%; d) Dome Iceberg, Dist: 11%; e) Sloping Iceberg, Dist: 23%.

The fast ice added a further degree of complexity as it does not always pose a threat to the hull of ships with a sufficient ice class. As such, not all visible ice was a direct danger to the ship, which meant the participant needed to rely on both the cameras and ECDIS sensor data to optimize their SA.

3.4 Procedure

Participants first filled out an informed consent form and a demographic form, and were then shown a brief instructional video before starting the study. When the test began, the participant was tasked with monitoring a MASS container ship and its 4 support vessels as they traversed the Davis Strait.

Additionally, the monoscopic panoramic camera was available to observe the surroundings of each vessel via the seven camera perspectives. After two minutes, the participant was prompted to answer questions. When they

accepted, they were asked three questions for the query set:

- One randomized SPAM question
- One self-assessed Trust question
- One Fast MS Scale (FMS) question

This process was repeated two more times at two-minute intervals for a total of three query sets, with 9 questions in total. The participant was then asked to fill out a Raw Task Load Index (RTLX) and System Usability Score (SUS) questionnaires for the immersion level, which was repeated for the other immersion levels.

3.5 Design

As discussed in the procedure section, the study started off with the participant completing an informed consent form, as well as a background questionnaire. The questionnaire asked people the following questions:

- Age
- Occupation and/or Field of Study
- Experience with Video Games (days played weekly) (GamesWk)
- Driving Experience (in years) (DrivingYr)
- Experience Flying Planes and/or Navigating Boats (in years)
- Robotics Experience (in years)
- Handedness

- Gender
- Sex at Birth

The questionnaire served an exploratory research purpose, as values could be correlated against DVs to assess potential impact and relationships.

As previously mentioned, following the questionnaire the participant watched a roughly three-minute long instructional video to acquaint them with the purpose, key terms such as SCCO, their task of remote monitoring, and expectations of the user study. After the video they were free to ask uncompromising questions, and then proceeded to begin testing.

The user study used a within-subjects design, with the primary IVs being levels of immersion. Herein, each participant performed all three immersion level tests. Counterbalancing was used to mitigate the effects of learning and asymmetric skill transfer during the three immersion levels. As such, a 3x3 Latin square was used to place participants into counterbalanced groups, as seen in Figure 3.9.

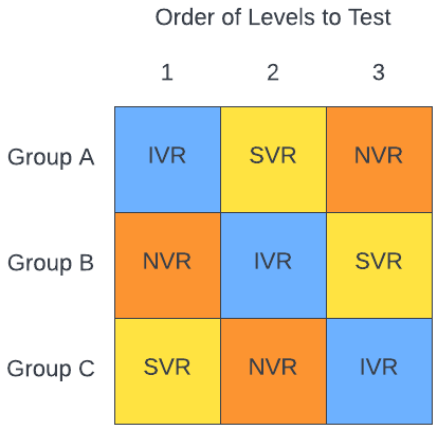


Figure 3.9: The 3x3 latin square for all immersion levels

Three different query sets were asked per immersion level, as seen in Figure 3.10. The length of each immersion level was intended to invoke a sense of routineness, the same kind that might occur in a real SCC where the SCCOs would spend a lot of time monitoring the MASS system's progress. Four primary DVs, SA Activation Time (SA-AT), SA Percent Correct (SA-PC), SA Response Time (SA-RT), and Trust were collected. Automating the recording of these variables increased objectivity by removing the role of experimenter. An additional DV was created using SA-PC and SA-RT, namely the Rate Correct Score (SA-RCS), which combined the two variables to get a better understanding of participant SA. Five extra DVs, namely MS, MWL, US, C-ARS, C-NVSC, were collected for exploratory purposes. The rest of this section will explore these DVs in more detail.

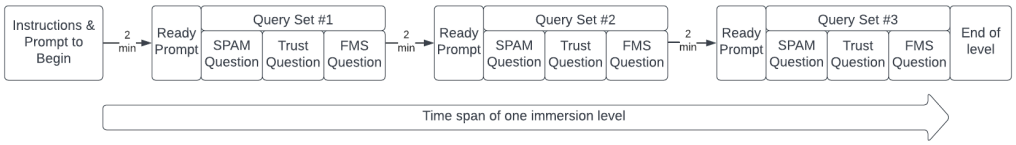


Figure 3.10: Timeline of one immersion level.
Shows Query Set #1 (Q1), Query Set #2 (Q2), and Query Set #3 (Q3).

3.5.1 Situational Awareness

Numerous methods have been developed to assess a user's SA at a given moment. More common methods like the SAGAT rely on pausing a simulation and querying the user based on their memory of the systems state directly preceding the question (Endsley, 2019; Durso et al., 2006; Fraune et al., 2021; Nguyen et al., 2019; Fabris et al., 2021). In contrast, SPAM focuses on one's ability to locate the requisite information, as opposed to using memory to

measure their SA, and is assessed based on reaction time and performance (Durso et al., 1999; Durso and Dattel, 2004; Cunningham et al., 2015). This is a useful feature when assessing SA in C2 scenarios where an operator would not need to memorize all data from moment-to-moment, but would simply need to know where to find said information. Additionally, SPAM does not pause during its set of queries, allowing the simulation to continue operating in the background (Endsley, 2019). This is also useful for C2 scenarios since in real control centres there would be no chance to pause external events.

It should be noted that the use of a ready prompt is standard for SPAM questions, as it allows the participant time attend to the question before response time measurements begin. As such, the time it takes for the participant to accept the question, herein referred to as SA-AT, is sometimes thought of as an SA correlate to MWL (Durso et al., 2006). Thus, a high SA-AT value is viewed as a negative outcome.

Conventionally, SA can be broken down into 3 primary levels, defined by Endsley (1995) as:

- SA Level 1: Perception of Elements in the Environment
- SA Level 2: Comprehension of the Current Situation
- SA Level 3: Projection of Future Status

Adding three different SPAM questions to each immersion level allowed for a question from each SA level to be asked, which further allowed for a more holistic assessment of a participant's SA. Additionally, the study used a modified SPAM design. The query sets began with a pre-recorded audio prompt. This was meant to emulate a radio communication with the SCCO.

The questions were not shown until the participant pressed "Yes" to accept it. An example of this process can be seen in Figure 3.11. The time from when the audio prompt ended until the participant accepted it was recorded as the SA-AT for that particular query set.

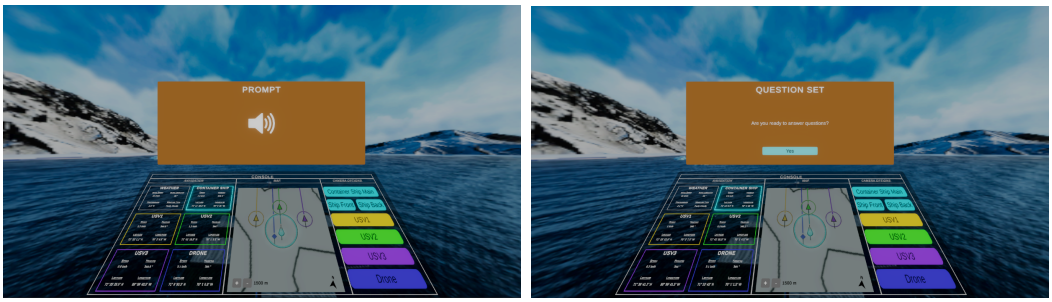


Figure 3.11: An example of how the initial query set prompt appeared. Left: The audio prompt, Right: The written prompt and the response button.

The SPAM question was then automatically assigned a random SA Level from the pool of unused levels, and then randomly assigned a question from that SA Level's pool of five questions. That SA Level was then removed from the pool of unused levels so that the same question could not be asked more than once, and each SA level was assessed once per immersion level. The random assignment of SA Level and SPAM question was used to mitigate any unintentional question biases.

All SPAM questions used pre-recorded audio prompts. Following this, a written version of the same question was displayed, as well as a set of 5 multiple choice responses. An example of this can be seen in Figure 3.12.

When the participant selected a response, the time it took to respond was saved as the SA-RT, wherein a low value was considered indicative of a higher SA. Additionally, the multiple choice selection was compared against the ground truth of the simulation at the moment of response. This saved a Boolean value

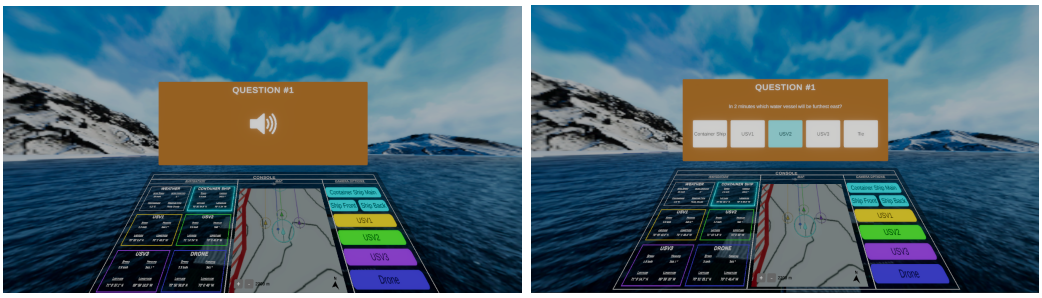


Figure 3.12: An example of a SPAM question in IVR.

of 1 for true and 0 for false into the SA-PC variable, which can be thought of more colloquially as accuracy. As noted previously, SA-RCS is a DV that combines SA-PC and SA-RT. First described by Woltz and Was (2006), and later compared against other combination metrics by Vandierendonck (2016), this method takes the sum of all correct answers and divides it by the sum all response times - this can be seen in Equation 3.1.

$$SA\text{-}RCS = \frac{\sum_{i=1}^{n}[SA\text{-}PC_i = 1]}{\sum_{i=1}^{n} SA\text{-}RT} \tag{3.1}$$

This DV can be thought of as the number of correct answers per second, and has the benefit of integrating the two primary SA metrics.

Finally, a full list of the SPAM question pool can be seen in Table 3.1.

3.5.2 Trust

Trust in automation, or more specifically, human Trust in autonomous robots, is a perceptual construct defined by Lee and See (2004) as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability". Automation is inevitably imperfect and can falter for a variety of reasons - which means it is never completely reliable.

Situational Awareness Level	SPAM Question	Possible Responses
Level 1	What is the Container Ship's heading?	0° - <72°
		72° - <144°
		144° - <216°
		216° - <288°
		288° - <360°
	What is the Container Ship's speed?	0 kn/h - <1 kn/h
		1 kn/h - <2 kn/h
		2 kn/h - <3 kn/h
		3 kn/h - <4 kn/h
		≥4 kn/h
	What is the Container Ship's latitude?	<70°0
		70°0 - <72°0
		72°0 - <74°0
		74°0 - <76°0
		≥76°0
	What is the Container Ship's longitude?	<70°0
		70°0 - <70°30
		70°30 - <71°0
		71°0 - <71°30
		≥71°30
	How many icebergs are in the Container Ship's proximity radius?	0
		1
		2
		3
		4+
Level 2	Which water vessel is furthest north?	Container Ship
		USV1
		USV2
		USV3
		Tie
	Which water vessel is furthest south?	Container Ship
		USV1
		USV2
		USV3
		Tie
	Which water vessel is furthest west?	Container Ship
		USV1
		USV2
		USV3
		Tie
	Which water vessel is furthest east?	Container Ship
		USV1
		USV2
		USV3
		Tie
	Which water vessel has the most obstacles in it's proximity radius?	Container Ship
		USV1
		USV2
		USV3
		Tie
Level 3	In 2 minutes which water vessel will be furthest north?	Container Ship
		USV1
		USV2
		USV3
		Tie
	In 2 minutes which water vessel will be furthest south?	Container Ship
		USV1
		USV2
		USV3
		Tie
	In 2 minutes which water vessel will be furthest west?	Container Ship
		USV1
		USV2
		USV3
		Tie
	In 2 minutes which water vessel will be furthest east?	Container Ship
		USV1
		USV2
		USV3
		Tie
	In 2 minutes which water vessel will have the most obstacles in it's proximity radius?	Container Ship
		USV1
		USV2
		USV3
		Tie

Table 3.1: SPAM question pool with all possible multiple choice responses.

So, to mitigate risk it is important to avoid over-reliance on these systems by maintaining an appropriate level of Trust. Having an appropriate level of Trust means that it is reflective of an autonomous robot's actual capabilities - a concept typically referred to as calibrated Trust. This thesis considers a moderate, or middling, Trust value to be appropriate for this simulation, since it is reasonable for the participant to assume that the MASS vessel and autonomous robots are at a potential risk of collision with ice in the scene. It should be noted that, over-reliance, and in turn inappropriately high Trust, can occur for a variety of reasons including a poor mental model and/or high MWL (Inagaki and Itoh, 2013). As such, it is important to understand how Trust interacts with other cognitive constructs.

Human Trust in robots is made more complicated by the fact that humans tend to anthropomorphize robotic vehicles and systems (Fraune et al., 2021; Hancock et al., 2020). That is to say people tend to erroneously attribute human emotions and features to robots, so understanding how self-assessed Trust develops across immersion levels, and over time, might give insight into this relationship.

Trust was assessed 3 times during each immersion level immediately following each SPAM question. This contemporaneous measurement contextualized how self-assessed Trust and SA correlated, and how that relationship evolved over time. The participant used a Likert-style rating scale to describe their current level of self-assessed Trust. This method was inspired by Xu & Dudek's self-assessed Trust question in their OPTIMo research (Xu and Dudek, 2015). However, they used a Visual Analog Scale as opposed to a rating scale, which I felt was less conducive to a VR setting - with discrete buttons being more evidently clickable. It should also be noted that Xu & Dudek, as well as

other groups like Hald et al. (2020), have attempted to create objective Trust measures, but I chose to use self-assessed Trust for its simplicity in terms of implementation and statistical assessment.

The Trust question asked participants to quantify their Trust in the autonomous MASS system's ability to safely navigate at a given moment. Herein, participants rated this Trust on a 7-point rating scale, with 1 representing the lowest Trust, and 7 representing the highest Trust. An example of this can be seen in Figure 3.13. This method is comparable to the 5-point Trust Likert scale used by Merritt (2011). A 7-point rating scale was chosen as it is relatively easy and fast to use, while still allowing participants more detail to express their sentiment than a 5-point rating scale (Preston and Colman, 2000).

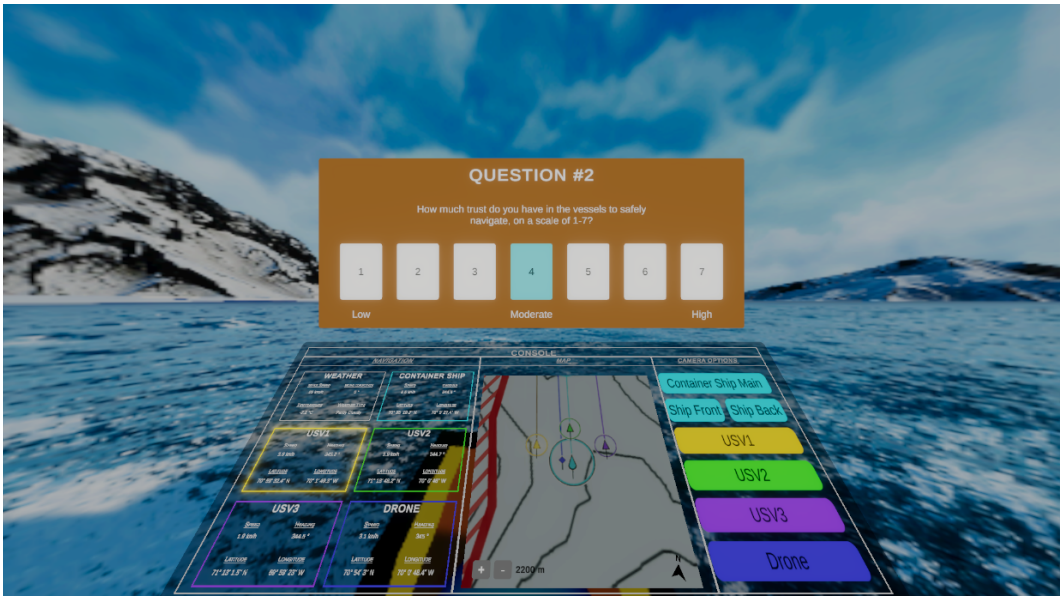


Figure 3.13: An example of a Trust question in IVR.

3.5.3 Motion Sickness

MS can occur during prolonged VR usage, which is often thought to be due to accrual of conflicting sensory information (Palmisano et al., 2020). For example, visual-vestibular conflict during self-motion can occur when the VR user's optic flow indicates movement but the inner ear indicates that the user is stationary (Keshavarz and Hecht, 2011; Palmisano et al., 2020). However, as noted, this is one of many proposed causes for visually induced MS, and there is some disagreement about the exact sensory interactions at the root of this problem. Regardless, MS is still generally acknowledged to increase over time in VR HMDs due to some compounding sensory conflicts (Akiduki et al., 2003). As such, MS should increase over time, especially in higher levels of immersion.

To account for this, a modified 7-point FMS question was asked immediately following the Trust question, with 1 representing the lowest MS, and 7 representing the highest MS. An example of this can be seen in Figure 3.14. The original FMS by Keshavarz and Hecht (2011) used a 21-point scale (0-20) for detailed rating, but since participants respond via buttons, as opposed to verbally, I opted to use a smaller scale to reduce visual clutter. Additionally, to match the other rating scale response measures in the study, a 7-point rating scale was chosen. Asking the FMS as part of the query set helped to contextualize MS with SA and Trust, both through mean comparisons and time series analysis.

3.5.4 Mental Workload

The NASA-TLX is a questionnaire to assess MWL, which was first developed in the 1980s by Hart and Staveland (1988). As this is a well-validated and

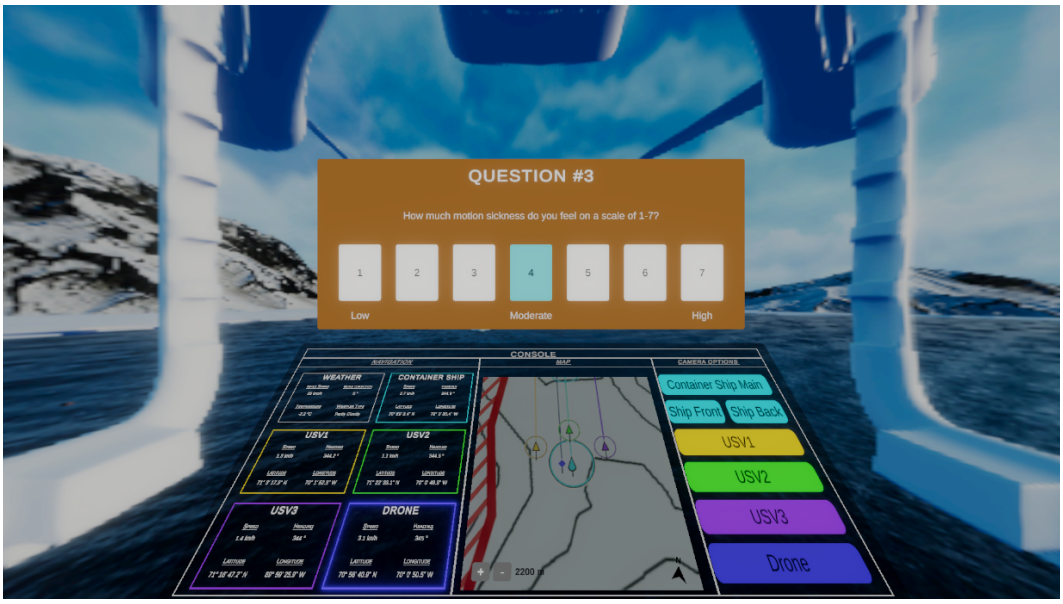


Figure 3.14: An example of an FMS question in IVR.

frequently used assessment for MWL in the literature, a variation of this test was administered in the written post-test questionnaire. Instead of using the pairwise rating system from the original NASA-TLX, I opted to use the version with the pairwise ratings removed. This is typically known as the RTLX score, and there is some evidence of it being more sensitive to changes in MWL (Hart, 2006; Bustamante and Spain, 2008). However, the main reason it was implemented in this user study was to make testing simpler and to avoid overwhelming the participants. Each question in the RTLX was scored out of 100, with the average scores from each question being used as the RTLX value. The application of the RTLX helped to ascertain the relationship between MWL on mean SA and mean Trust.

3.5.5 Usability

The SUS is a questionnaire that was used to determine the participant's opinion of the system's US for each level (Brooke, 1986). This questionnaire was in the form of a written response, and administered on the same page as the RTLX. The scale returns a score out of 40 which can then be represented as percentage. This was useful for understanding how participants perceived the US of the immersion level interfaces, and how it correlated to their mean SA and mean Trust.

3.5.6 Camera Data

Two different variables were recorded from the camera data returned by the simulation - the C-ARS and the C-NVSC. C-ARS was the average angular speed in degrees per second. C-NVSC was the total count of the times the participant switched between panoramic camera views. These variables gave insight into how camera usage relates to mean SA and mean Trust in the monitoring of autonomous robots.

3.6 Statistical Analysis

This section will discuss the methods of data analysis. This will be done by reviewing the data subsets, methods of normalization, methods of parametric analysis, and methods of non-parametric analysis.

3.6.1 Subset Analysis

This research aimed to better understand how SA and Trust are affected by varying levels of immersion when monitoring remote robotics, such as MASS and their navigational aids. MASS has the unique benefit of having hazardous environmental elements that approach more slowly than in other autonomous vehicle domains such as automobiles and planes. Two main problems arose from this research however. The first is that it examined the human factor impact of monitoring a hypothetical technology, in a professional setting, using the general populace instead of Subject-matter Experts (SMEs) like seafarers. The second is that since this technology is still largely in its conception stages - or planning stages in the case of the Yara Birkeland (Kongsberg, 2022) - there appear to be no SCCOs yet, and as such, no true SMEs. So while it would have preferable to test on seafarers, since it is assumed the first batch of SCCOs will have a seafaring background, it is not an exact correlate since they would have a different set of experiences and biases. With these considerations in mind, it is important to consider additional methods of analyzing the output data to better understand how a technologically savvy person might interact with such systems. To accomplish this, I opted to analyze both the full data set, as well as various subset groups within the full set - such as people with consistent experience playing video games, people with some experience driving, and people that can get a minimum number of SA questions correct. Additionally, the Interquartile Range (IQR) method was implemented on two of the larger data sets to examine the effects of removing outliers. The entire list of data sets examined is as follows:

- Full data set

  - 39 people
  - Takes the entire set of participants.

- Passed subset
  - 30 people
  - Takes the participants that managed to get at least one SPAM question right in every immersion level (i.e., showing some basic competence).

- GamingExp subset
  - 18 people
  - Takes the participants with at least some weekly gaming experience (i.e., they listed more than 0 for the gaming question in the pre-test questionnaire).

- DrivingExp subset
  - 30 people
  - Takes the participants with at least some driving experience (i.e., they listed more than 0 for the driving question in the pre-test questionnaire).

- Full data set with no outliers (FullNO)
  - 34-39 people, depending on the DV
  - Takes the entire set of participants, but removes participants who had one or more outlier per variable.

- Passed subset with no outliers (PassedNO)
  - 25-30 people, depending on the DV
  - Takes the participants that managed to get at least one SPAM question right in every immersion level, but removes participants who had one or more outliers per variable.

Notably, no subsets were analyzed for experience boating or experience with robotics, because only 2 participants respectively stated experience with these technologies.

3.6.2 Normalization

All data that was measured was normalized using monotonic power functions so as to satisfy the requirements for parametric testing. These were calculated using pre-built functions from Python's SciPy package, which automatically generated the fitted lambda values for each data set when ran. This was done to the full data set, all three subsets, and the outlier analyses. All DVs with strictly positive values were normalized using a fitted lambda and Box-Cox Transformation (BC)[7] - which, can be seen in Equation 3.2 (Box and Cox, 1964).

$$y = \begin{cases} (x^{\lambda} - 1) & \lambda \neq 0 \\ log(x) & \lambda = 0 \end{cases} \tag{3.2}$$

Conversely, the rest of the DVs were normalized using a fitted lambda and Yeo-Johnson Transformation (YJ)[8] - which, can be seen in Equation 3.3 (Yeo

[7]https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.boxcox.html
[8]https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.yeojohnson.html

and Johnson, 2000).

$$y = \begin{cases} \frac{((x+1)^{\lambda}-1)}{\lambda} & x \geq 0, \lambda \neq 0 \\ log(x+1) & x \geq 0, \lambda = 0 \\ \frac{-((-x+1)^{(2-\lambda)}-1)}{(2-\lambda)} & x < 0, \lambda \neq 0 \\ -log(-x+1) & x < 0, \lambda = 0 \end{cases} \tag{3.3}$$

The normalized data for the Full data set was then tested for normality using the Lilliefors test, which was completed using Scott MacKenzie's GoStats application[9]. This test was performed on the normalized Full data set alone as it was intended to understand which variables in general illustrate normality. Herein, the null hypothesis stipulates that the residuals for the DV will have a Gaussian distribution, so DVs that do not have their null hypotheses rejected are considered to have residuals that have satisfied normality.

3.6.3 Parametric Analysis

For all normalized mean data I performed a repeated measures Analysis of Variance (ANOVA) with a post-hoc Fisher's Least Significant Difference (LSD) test. Additionally, for the normalized data of the time series breakdown I used a two-way ANOVA between immersion level and query set. A post-hoc Fisher's LSD was also used for the time series breakdown, and was split into two groups - comparing differences between immersion levels per query set, and comparing differences between query sets per immersion level.

Another set of two-way ANOVAs were performed on the normalized data for the Full data set, between immersion level and counterbalanced groups.

[9]http://www.yorku.ca/mack/GoStats/

This was to determine whether there was any statistically significant group effect. All of these methods of analysis were performed on all of the data subsets in an exploratory analysis. The results of these analyses are shown in Chapter 4.

Two important points should be highlighted here. Regardless of whether or not the normalized data satisfied normality per the Lilliefors test it was still tested via ANOVA. Glass et al. (1972) note that normality, depending on circumstance, is not always considered a strict requirement for ANOVA testing. As such, using an ANOVA on non-normally distributed data was considered valid for this analysis. However, all DVs that did not satisfy normality were also analyzed using a Friedman test to confirm the robustness of the findings. The second major point to discuss is that core DVs, which are DVs collected from the SPAM and self-assessed Trust measures (i.e., SA-AT, SA-PC, SA-RT, SA-RCS, and Trust), were tested using post-hoc analysis regardless of whether or not they achieved significance.

3.6.4 Non-Parametric Analysis

As discussed, several of the DVs transformed did not satisfy Lilliefors test for normality. While still analyzed using ANOVA tests, comparable non-parametric tests were also required to confirm the findings and better understand these data. Since the user study used a within-subjects design with multiple IV levels, a Friedman test was utilized, along with a Connover's F pairwise analysis for the mean DVs. The Friedman was completed using GoStats, and is a non-parametric correlate for the one-way ANOVA specifically. This method of analysis produced two variables for determining statistical significance, which

were the p value (p), and the p value corrected for ties (p'). For the time series breakdown, I was unable to find an utility or academic source for calculating a Friedman test correlate to the two-way ANOVA. As such, I opted to perform a one-way Friedman test on each query set and each immersion level, respectively. These breakdowns also include Connover's F pairwise comparisons. The outputs of these analyses are listed in Chapter 4.

Chapter 4

Results

Before beginning this chapter, the context of the simulation length should be discussed. Each immersion level simulation took around 8 minutes for participants to complete on average, with IVR taking the longest - the probable reason for this will be discussed later. A comparison of the mean completion times can be seen in Figure 4.1. Standard Error (SE) bars are small and show very little variation from the mean across all immersion levels. Completion times here differentiate based on the time it takes to respond to the questions in the query set.

This chapter is broken down into three major sections - DV analysis, correlation analysis, and a brief summary. DV analysis will look at each of the DVs separately and look at their means as a function of the IVs and time series breakdowns. Correlation Analysis will use the Pearson Correlation Coefficient to compare the strength of correlation between two DVs in all of the data set variations. The first subsection discusses the full correlation table. The second subsection compares all of the pertinent correlations to the research goals. Finally, the summary section contextualizes the core results of this chapter.

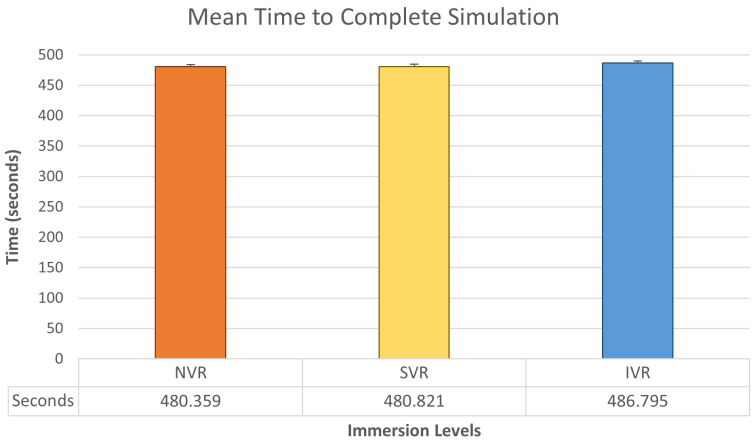


Figure 4.1: Mean immersion level completion time.
Determined using the Full data set.
Error bars represent SE.

4.1 Dependent Variable Analysis

The following section will discuss how the different DVs were affected by the immersion levels. To do this, the methods of normalization, group effect, means, and time series breakdowns of the DVs is examined.

4.1.1 Normality

The results of the Lilliefors test for the normalized Full data set can be seen in Table 4.1. The DVs that had their null hypothesis rejected did not satisfy normality - these were SA-PC, SA-RCS, Trust, MS, MWL, and US.

4.1.2 Group Effect

A two-way ANOVA on group and immersion was used to test for a statistically significant group effect. Herein, none of the DVs had a statistically significant group effect, as can be seen in Table 4.2. This is an encouraging result as it

Dependent Variable (DV)	Power Transformation	Fitted Lambda	Number of Samples (n)	Mean Values	Standard Deviation (SD)	Test Statistic (M)	Critical Value (CV)	Null Hypothesis
SA-AT	BC	-0.44	117	0.414	0.482	0.045	0.082	**Not Rejected**
SA-PC	YJ	0.88	117	40.917	20.137	0.232	0.082	Rejected
SA-RT	BC	-0.06	117	2.064	0.433	0.071	0.082	**Not Rejected**
SA-RCS	YJ	-5.74	117	0.056	0.033	0.084	0.082	Rejected
Trust	BC	1.71	117	9.905	4.170	0.129	0.082	Rejected
MS	BC	-0.58	117	0.437	0.431	0.289	0.082	Rejected
MWL	BC	0.47	117	7.191	3.238	0.084	0.082	Rejected
US	BC	1.54	117	484.632	149.683	0.096	0.082	Rejected
C-ARS	YJ	-0.31	117	0.772	0.462	0.072	0.082	**Not Rejected**
C-NVSC	YJ	0.46	117	5.994	2.758	0.066	0.082	**Not Rejected**

Table 4.1: Results of the Lilliefors normality test for the Full data set.

implied that placement into counterbalanced groups did not have a significant impact on the DVs.

Dependent Variable (DV)	Power Transformation	Fitted Lambda	Independent Variable (IV)	Degrees of Freedom (df)	F Statistic	p Value
SA-AT	BC	-0.44	Group	(2, 36)	0.39	0.678
			Immersion	(2, 72)	19.47	**0.000**
SA-PC	YJ	0.88	Group	(2, 36)	0.23	0.798
			Immersion	(2, 72)	0.21	0.808
SA-RT	BC	-0.06	Group	(2, 36)	2.59	0.089
			Immersion	(2, 72)	0.01	0.993
SA-RCS	YJ	-5.74	Group	(2, 36)	0.51	0.604
			Immersion	(2, 72)	0.49	0.617
Trust	BC	1.71	Group	(2, 36)	2.78	0.075
			Immersion	(2, 72)	2.09	0.131
MS	BC	-0.58	Group	(2, 36)	1.53	0.230
			Immersion	(2, 72)	5.57	**0.006**
MWL	BC	0.47	Group	(2, 36)	0.74	0.487
			Immersion	(2, 72)	4.51	**0.014**
US	BC	1.54	Group	(2, 36)	1.29	0.288
			Immersion	(2, 72)	1.32	0.274
C-ARS	YJ	-0.31	Group	(2, 36)	0.36	0.704
			Immersion	(2, 72)	96.25	**0.000**
C-NVSC	YJ	0.46	Group	(2, 36)	0.95	0.396
			Immersion	(2, 72)	0.23	0.794

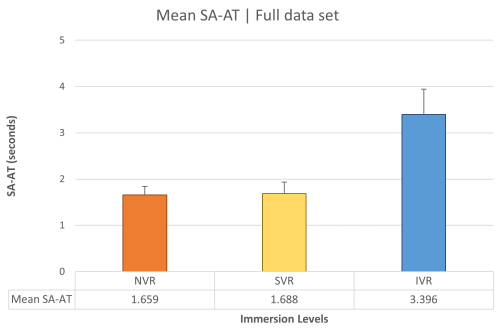
Table 4.2: Results of the two-way ANOVA to verify counterbalancing.

4.1.3 Means

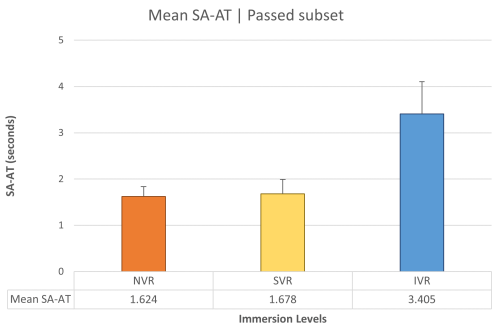
This section will explore the patterns of mean responses across immersion levels for all of the DVs.

Situational Awareness - Activation Time (SA-AT)

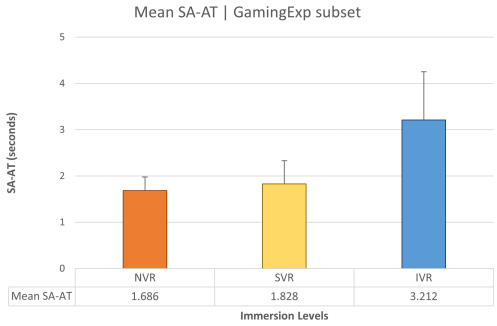
The mean results of SA-AT for all of the data set variations can be seen in Figure 4.2.



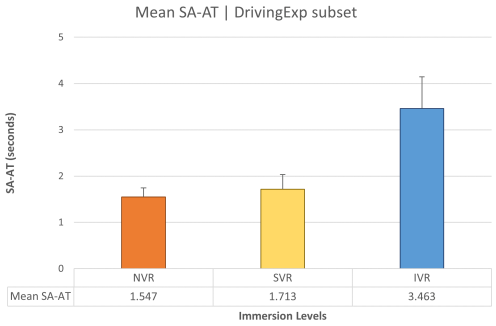
a) Full data set



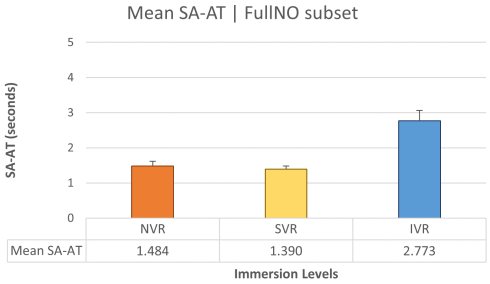
b) Passed subset



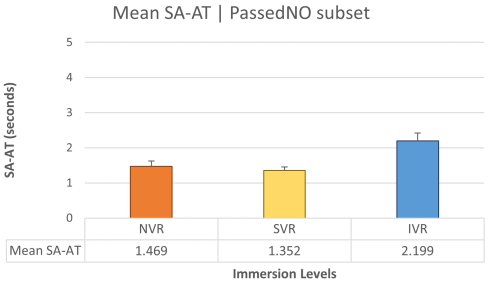
c) GamingExp subset



d) DrivingExp subset



e) FullNO subset



f) PassedNO subset

Figure 4.2: Mean SA-AT comparisons of the three immersion levels. Error bars represent SE.

One can note for the Full data set how mean SA-AT for IVR was more than twice as slow as the other two SA-AT means - which is a probable cause for why the IVR's average completion time was slightly slower. Notably, this pattern did not change very much when examining the subset data. GamingExp showed a minor reduction in SA-AT for IVR relative to other subsets, but it

still showed a large difference with the other two immersion levels.

The repeated measures ANOVA performed on mean SA-AT can be seen in Table 4.3 for each data set. SA-AT had omnibus significance for the Full data set and every subset variation. The post-hoc Fisher's LSD test showed pairwise significance between NVR and IVR, as well as SVR and IVR. So IVR had a significantly higher SA-AT overall, regardless of subset.

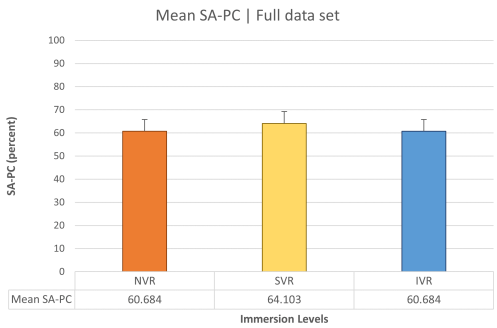
	Power Function	Fitted Lambda	Degrees of Freedom (df)	F Statistic	p Value	Post Hoc Fisher's LSD
Full	BC	-0.44	(2, 76)	18.86	**0.000**	N:I, S:I
Passed	BC	-0.49	(2, 58)	11.70	**0.000**	N:I, S:I
GamingExp	BC	-0.60	(2, 34)	3.94	**0.029**	N:I, S:I
DrivingExp	BC	-0.53	(2, 58)	14.84	**0.000**	N:I, S:I
FullNO	BC	-0.30	(2, 66)	20.14	**0.000**	N:I, S:I
PassedNO	BC	-0.11	(2, 48)	8.47	**0.001**	N:I, S:I

Table 4.3: Mean SA-AT ANOVA analysis across all of the data set variations. Note for the pairwise comparisons: standard letters indicate pairwise significance between immersion levels and omnibus significance; italicized letters indicate pairwise significance between immersion levels and no omnibus significance; light grey shading indicates no pairwise significance between immersion levels and omnibus significance; dark grey shading indicates no pairwise significance between immersion levels and no omnibus significance.

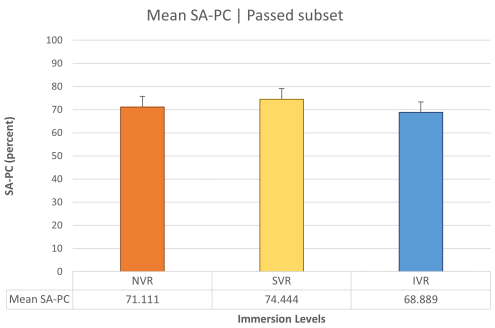
Situational Awareness - Percent Correct (SA-PC)

The mean results of SA-PC for all of the data set variations can be seen in Figure 4.3.

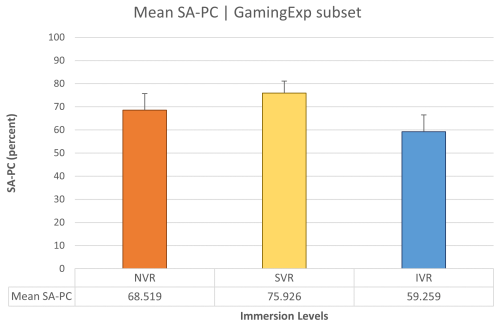
One can note that mean SA-PC for SVR was slightly higher than the other two immersion levels, with NVR and IVR being similar. The fact that SA-PC for NVR and IVR performances are similar in several instances could be due to the fact that the mean results for SA-PC were fairly discrete, with only 4 different results being possible per immersion level. Regardless, SA-PC for SVR was higher than both of the other values across all data set variations; a relationship that was the most pronounced in the GamingExp subset.



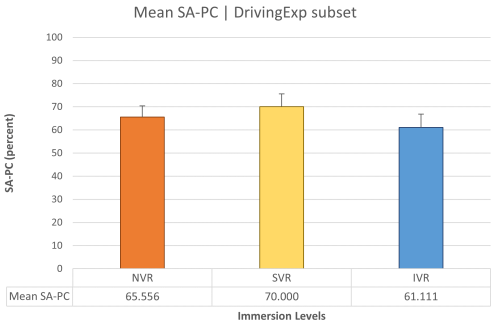
a) Full data set



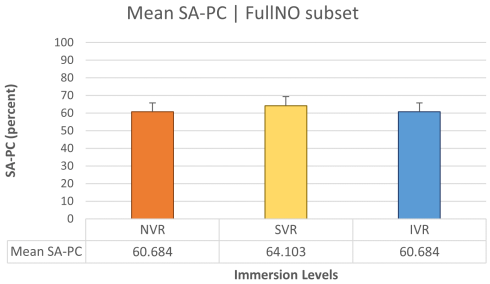
b) Passed subset



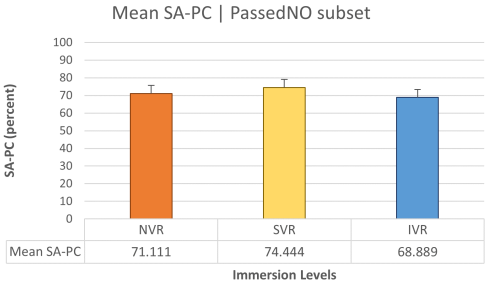
c) GamingExp subset



d) DrivingExp subset



e) FullNO subset



f) PassedNO subset

Figure 4.3: Mean SA-PC comparisons of the three immersion levels. Error bars represent SE.

The ANOVA for mean SA-PC showed no significant effects. The mean SA-PC analysis can be seen in Table 4.4, wherein no significant effect of immersion was present in any of the data set variations.

The Friedman analysis for mean SA-PC can be seen in Table 4.5. Herein, both p and p' (tie corrected) showed no omnibus significance across any of the

	Power Function	Fitted Lambda	Degrees of Freedom (df)	F Statistic	p Value	Post Hoc Fisher's LSD
Full	YJ	0.88	(2, 76)	0.20	0.821	
Passed	YJ	1.11	(2, 58)	0.43	0.650	
GamingExp	YJ	2.66	(2, 34)	2.04	0.146	
DrivingExp	YJ	2.14	(2, 58)	0.92	0.402	
FullNO	YJ	0.88	(2, 76)	0.20	0.821	
PassedNO	YJ	1.11	(2, 58)	0.43	0.650	

Table 4.4: Mean SA-PC ANOVA analysis across all of the data set variations.

data set variations. However, as this was a core planned comparison to the research, a post-hoc analysis was also done using Connover's F, which did show pairwise significance between SVR and IVR in the GamingExp subset.

	df	H	p	H'	p'	Connover's F
Full	2	0.94	0.626	1.38	0.502	
Passed	2	1.22	0.544	1.92	0.383	
GamingExp	2	3.44	0.179	5.51	0.064	S : I
DrivingExp	2	2.40	0.301	3.43	0.180	
FullNO	2	0.94	0.626	1.38	0.502	
PassedNO	2	1.22	0.544	1.92	0.383	

Table 4.5: Mean SA-PC Friedman analysis across all of the data set variations.

Another key factor to consider is how the different SA level questions performed on average. One would expect a gradually decreasing mean SA-PC with increasing SA level. This speculation is based on the premise that as a question requires a higher level of SA it will require the operator to better synthesize data and generate an increasingly accurate mental model, which is a progressively difficult task (Endsley, 1995). As seen in Figure 4.4 this turned out to be the case, which is an encouraging validation of the SPAM question design.

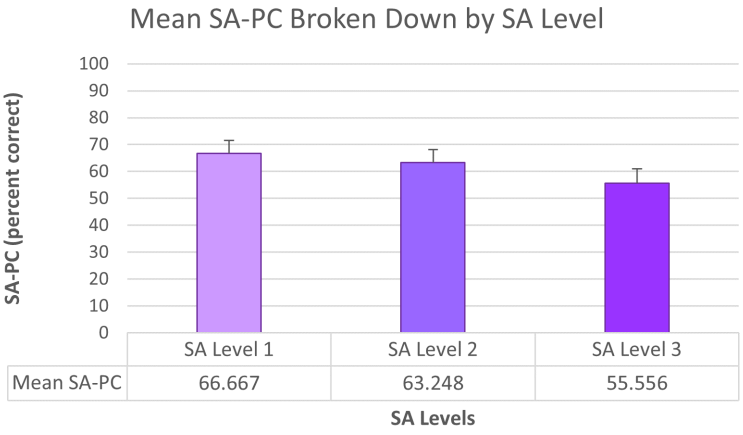


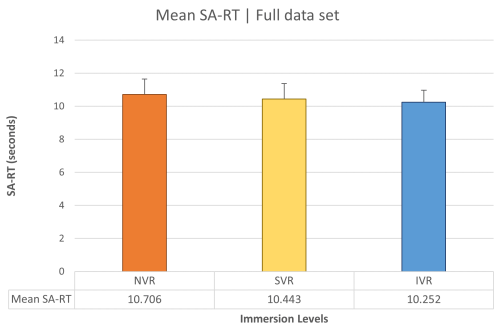
Figure 4.4: Mean SA-PC broken down by SA level.
Error bars represent SE.

Situational Awareness - Response Time (SA-RT)

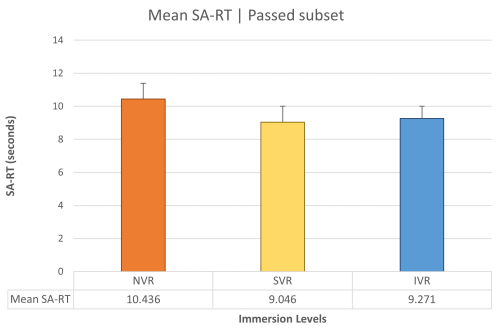
The mean results of SA-RT for all of the data set variations can be seen in Figure 4.5.

For the Full data set mean SA-RT showed a shallow decreasing trend from NVR to SVR to IVR. This indicates that IVR had the shortest mean SA-RT, however this trend does not hold for the data subsets that suggest technical proficiency. In every data subset, outside of FullNO, SVR had the fastest mean SA-RT, with IVR having the second fastest.

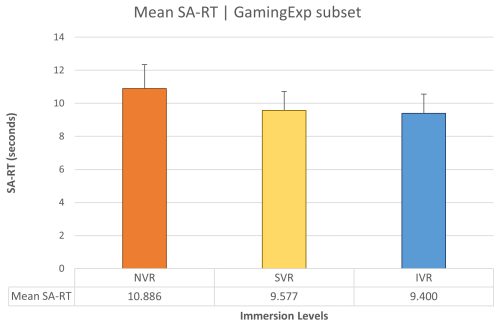
As with mean SA-PC in the previous section, mean SA-RT's ANOVA did not indicate differences between immersion levels except in the PassedNO subset where there was some statistical significance. The Fisher's LSD showed pairwise significance between NVR and SVR, which based on the means indicates that SVR had consistently faster SA-RT for that data subset. The results of this can be seen in Table 4.6.



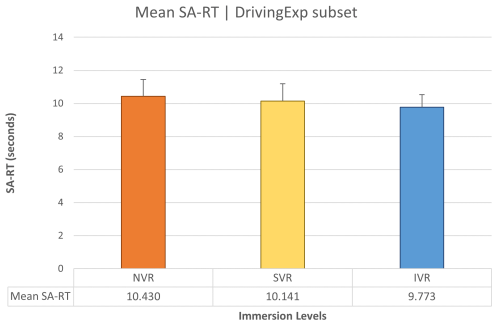
a) Full data set



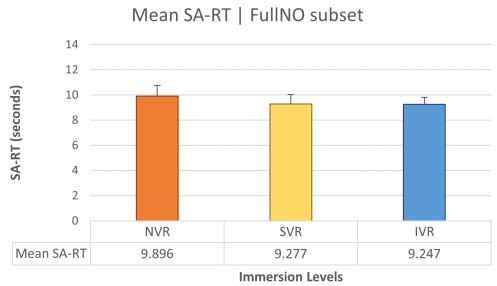
b) Passed subset



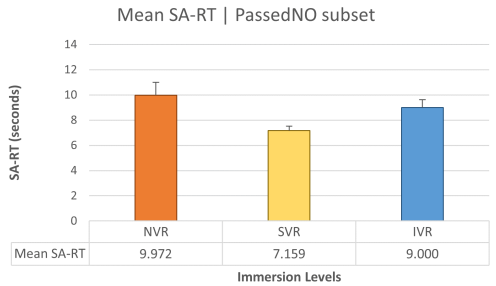
c) GamingExp subset



d) DrivingExp subset



e) FullNO subset



f) PassedNO subset

Figure 4.5: Mean SA-RT comparisons of the three immersion levels. Error bars represent SE.

Situational Awareness - Rate Correct Score (SA-RCS)

The mean results of SA-RCS for all of the data set variations can be seen in Figure 4.6.

Notably, SVR showed a trend of having the best mean SA-RCS in every data set variation, with NVR and IVR trading places in the different data sets.

	Power Function	Fitted Lambda	Degrees of Freedom (df)	F Statistic	p Value	Post Hoc Fisher's LSD
Full	BC	-0.06	(2, 76)	0.01	0.993	
Passed	BC	-0.14	(2, 58)	0.76	0.474	
GamingExp	BC	-0.09	(2, 34)	0.67	0.518	
DrivingExp	BC	-0.04	(2, 58)	0.01	0.988	
FullNO	BC	0.01	(2, 68)	0.06	0.945	
PassedNO	BC	-0.32	(2, 48)	3.44	**0.040**	N:S

Table 4.6: Mean SA-RT ANOVA analysis across all of the data set variations.

However, the ANOVA for mean SA-RCS, which can be seen in Table 4.7, found no statistically significant effects for immersion in any of the data set variations.

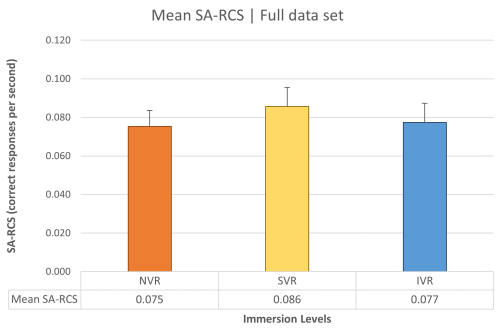
	Power Function	Fitted Lambda	Degrees of Freedom (df)	F Statistic	p Value	Post Hoc Fisher's LSD
Full	YJ	-5.74	(2, 76)	0.44	0.645	
Passed	YJ	-6.58	(2, 58)	1.00	0.375	
GamingExp	YJ	-4.40	(2, 34)	0.73	0.490	
DrivingExp	YJ	-5.92	(2, 58)	0.72	0.489	
FullNO	YJ	-5.38	(2, 72)	1.05	0.356	
PassedNO	YJ	-5.52	(2, 56)	1.69	0.193	

Table 4.7: Mean SA-RCS ANOVA analysis across all of the data set variations.

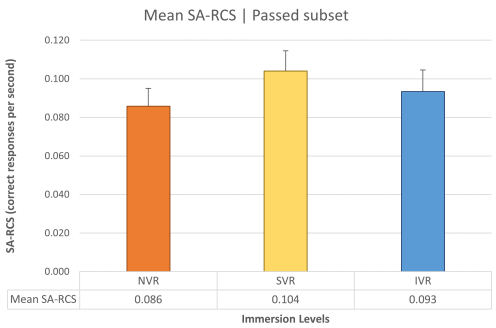
The Friedman results for mean SA-RCS can be seen in Table 4.8. Here one can note there was no omnibus significance for any of the data set variations. However, as this was a core planned comparison to the research, a Connover's F was performed, which showed pairwise significant differences between SVR and IVR for both FullNO and PassedNO. This indicates that in these specific scenarios SVR had a consistently better SA-RCS than IVR.

	df	H	p	H'	p'	Connover's F
Full	2	3.13	0.209	3.19	0.203	
Passed	2	4.07	0.131	4.07	0.131	
GamingExp	2	4.19	0.123	4.25	0.119	
DrivingExp	2	3.32	0.191	3.35	0.188	
FullNO	2	4.92	0.086	5.02	0.081	*S : I*
PassedNO	2	5.24	0.073	5.24	0.073	*S : I*

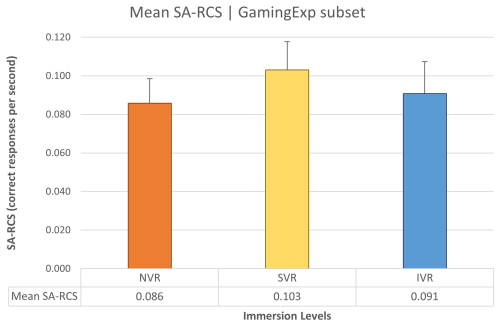
Table 4.8: Mean SA-RCS Friedman analysis across all of the data set variations.



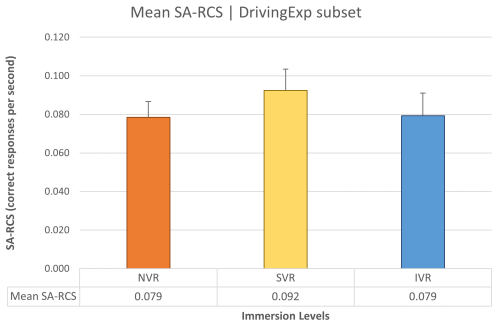
a) Full data set



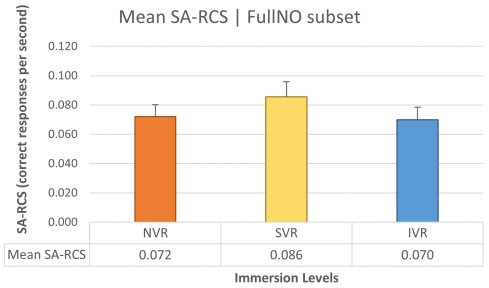
b) Passed subset



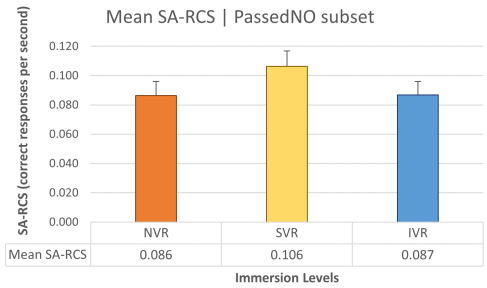
c) GamingExp subset



d) DrivingExp subset



e) FullNO subset



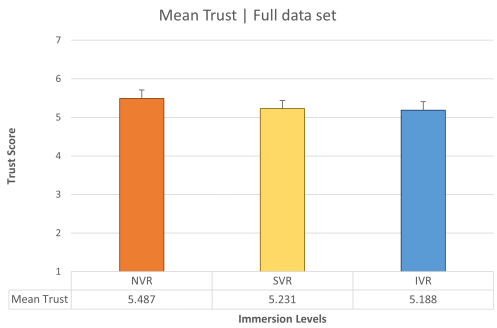
f) PassedNO subset

Figure 4.6: Mean SA-RCS comparisons of the three immersion levels. Error bars represent SE.

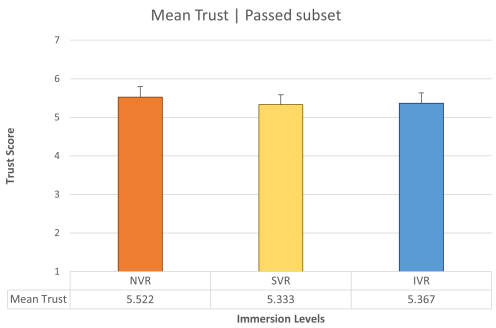
Trust

The results of mean self-assessed Trust for all of the data set variations can be seen in Figure 4.7.

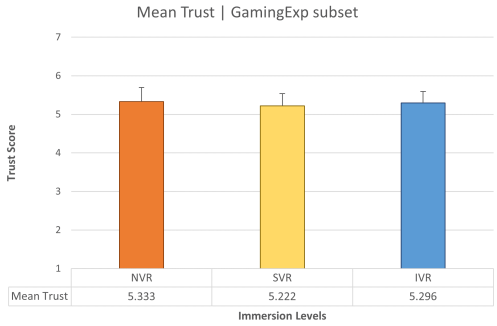
Mean self-assessed Trust appeared to decline for higher immersion levels, as NVR showed a trend of being consistently higher than either SVR or IVR



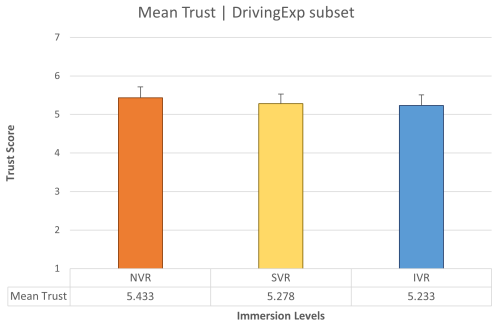
a) Full data set



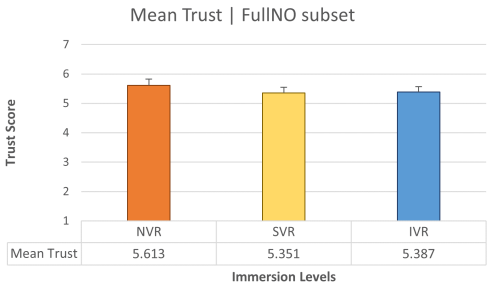
b) Passed subset



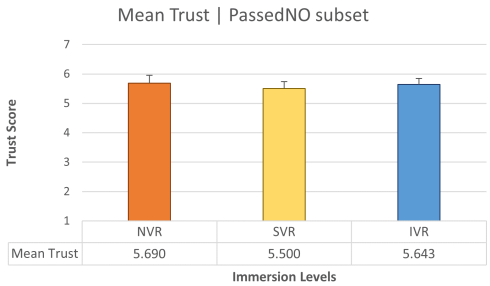
c) GamingExp subset



d) DrivingExp subset



e) FullNO subset



f) PassedNO subset

Figure 4.7: Mean Trust comparisons of the three immersion levels. Error bars represent SE.

in every data set variation.

The ANOVA for mean Trust can be seen in Table 4.9. Notably, no statistically significant relationships between Trust and immersion could be discerned in either the Full data set or any of the various subsets.

The results of the Friedman test for mean Trust can be seen in Table 4.10.

	Power Function	Fitted Lambda	Degrees of Freedom (df)	F Statistic	p Value	Post Hoc Fisher's LSD
Full	BC	1.71	(2, 76)	1.99	0.144	
Passed	BC	1.95	(2, 58)	0.68	0.512	
GamingExp	BC	1.49	(2, 34)	0.12	0.888	
DrivingExp	BC	1.75	(2, 58)	0.60	0.553	
FullNO	BC	1.61	(2, 72)	1.66	0.196	
PassedNO	BC	2.10	(2, 54)	0.68	0.509	

Table 4.9: Mean Trust ANOVA analysis across all of the data set variations.

Trust showed no significant effect of immersion for any of the data set variations. However, Trust was also considered one of the core variables to this research, and as such, a Connover's F was performed. Herein, pairwise significance between NVR and SVR could be observed for both the Full data set and the FullNO data set. This indicates that Trust for NVR was consistently higher than Trust for SVR in those data sets.

	df	H	p	H'	p'	Connover's F
Full	2	3.86	0.145	5.15	0.076	N : S
Passed	2	1.85	0.397	2.68	0.263	
GamingExp	2	1.19	0.550	1.56	0.458	
DrivingExp	2	0.65	0.723	0.87	0.648	
FullNO	2	3.50	0.174	4.75	0.093	N : S
PassedNO	2	2.09	0.352	3.12	0.210	

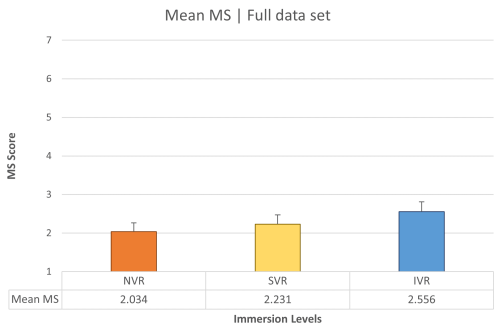
Table 4.10: Mean Trust Friedman analysis across all of the data set variations.

Motion Sickness (MS)

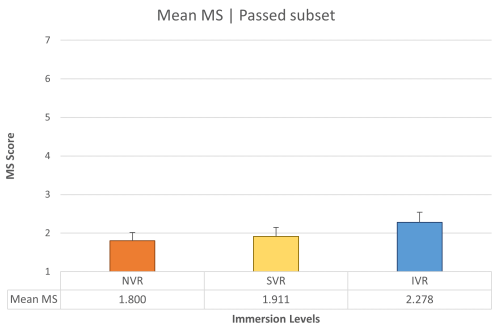
Mean self-assessed MS results for the data set variations can be seen in Figure 4.8.

These results are not very stark, but they are consistent. Mean MS increased in the order NVR $<$ SVR $<$ IVR across all of the data set variations.

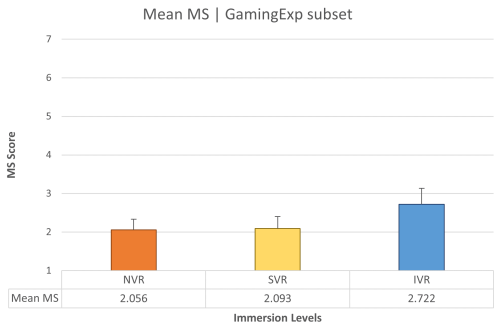
The ANOVA for mean MS can be seen in Table 4.11. Herein, there was significance for every data set variation. Interestingly, no pairwise significance



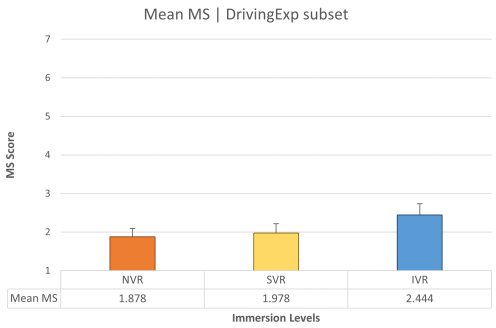
a) Full data set



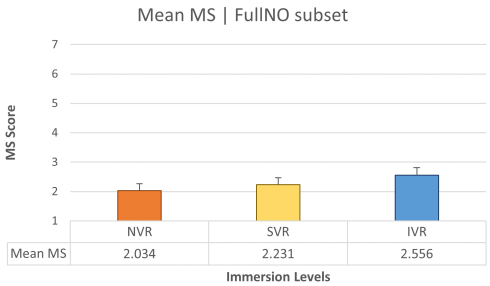
b) Passed subset



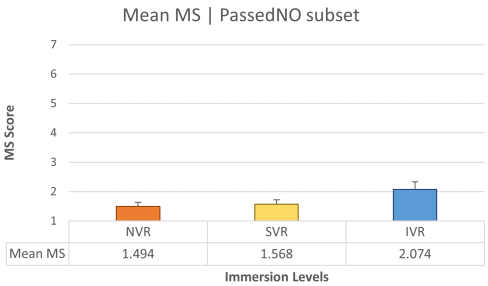
c) GamingExp subset



d) DrivingExp subset



e) FullNO subset



f) PassedNO subset

Figure 4.8: Mean MS comparisons of the three immersion levels. Error bars represent SE.

was observed using Fisher's LSD. This indicates that while the immersion level effect was present, pairwise differences were too small to be detected by this post-hoc analysis.

The Friedman analysis for mean MS can be seen in Table 4.12. Herein, p showed no significance for any of the data set variations, but p' did show

	Power Function	Fitted Lambda	Degrees of Freedom (df)	F Statistic	p Value	Post Hoc Fisher's LSD
Full	BC	-0.58	(2, 76)	5.17	**0.008**	
Passed	BC	-0.94	(2, 58)	3.42	**0.040**	
GamingExp	BC	-0.44	(2, 34)	3.50	**0.042**	
DrivingExp	BC	-0.75	(2, 58)	4.20	**0.020**	
FullNO	BC	-0.58	(2, 76)	5.17	**0.008**	
PassedNO	BC	-1.40	(2, 52)	3.35	**0.043**	

Table 4.11: Mean MS ANOVA analysis across all of the data set variations.

significance for PassedNO. There was also pairwise significance between NVR and IVR per Connover's F for PassedNO.

	df	H	p	H'	p'	Connover's F
Full	2	2.71	0.259	4.44	0.109	
Passed	2	2.22	0.330	3.91	0.141	
GamingExp	2	3.25	0.197	4.88	0.087	
DrivingExp	2	3.15	0.207	5.48	0.065	
FullNO	2	2.71	0.259	4.44	0.109	
PassedNO	2	3.69	0.158	6.86	**0.032**	N:I

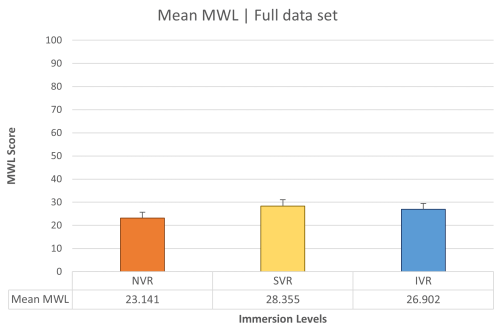
Table 4.12: Mean MS Friedman analysis across all of the data set variations.

Mental Workload (MWL)

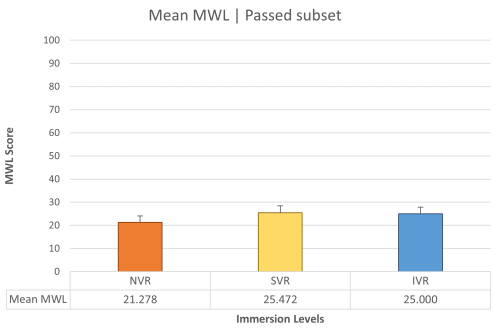
The mean MWL results for the data set variations can be seen in Figure 4.9.

Notably, mean MWL showed a trend of being lowest for NVR and highest for SVR, though this difference was relatively small. This order remained the same in most of the data set variations, except for DrivingExp where MWL was highest for IVR.

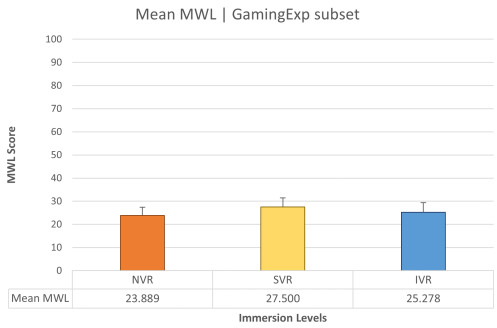
The ANOVA for mean MWL, which can be seen in Table 4.13, showed significance for both the Full data set and the FullNO subset. However, none of the pairwise comparisons between immersion levels were significant. Also, ANOVA did not show significant effects of immersion in the other subset variations.



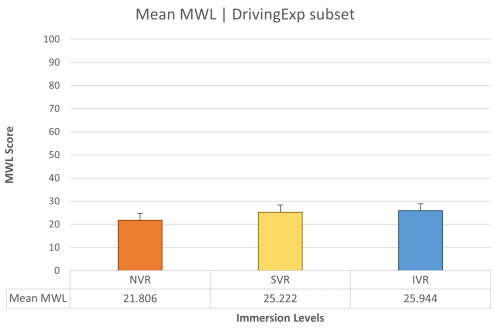
a) Full data set



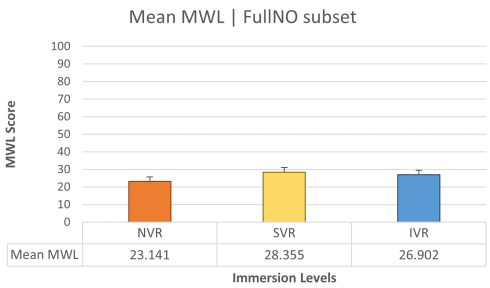
b) Passed subset



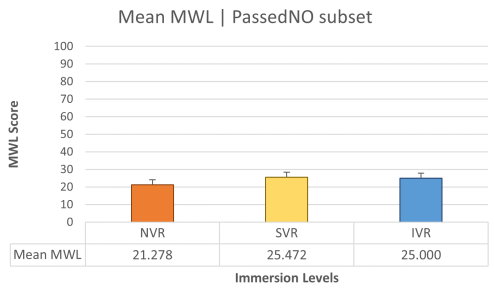
c) GamingExp subset



d) DrivingExp subset



e) FullNO subset



f) PassedNO subset

Figure 4.9: Mean MWL comparisons of the three immersion levels. Error bars represent SE.

The results of the Friedman test for mean MWL can be seen in Table 4.14. Notably, p and p' were significant for the Full data set, as well as the DrivingExp and FullNO subsets. All of these showed pairwise significance between NVR and SVR, and the Full and FullNO data sets also showed pairwise significance between NVR and IVR. This indicates that in these data sets NVR

	Power Function	Fitted Lambda	Degrees of Freedom (df)	F Statistic	p Value	Post Hoc Fisher's LSD
Full	BC	0.47	(2, 76)	4.10	**0.020**	
Passed	BC	0.43	(2, 58)	2.58	0.084	
GamingExp	BC	0.30	(2, 34)	0.99	0.380	
DrivingExp	BC	0.39	(2, 58)	2.92	0.062	
FullNO	BC	0.47	(2, 76)	4.10	**0.020**	
PassedNO	BC	0.43	(2, 58)	2.58	0.084	

Table 4.13: Mean MWL ANOVA analysis across all of the data set variations.

had consistently lower MWL than IVR, and more specifically, in the Full and FullNO data sets NVR had a consistently lower MWL than both of the other immersion levels.

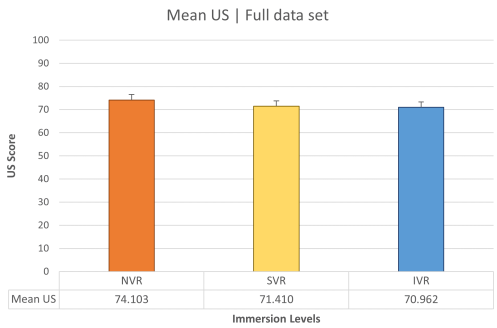
	df	H	p	H'	p'	Connover's F
Full	2	6.55	**0.038**	6.72	**0.035**	N:S, N:I
Passed	2	5.15	0.076	5.33	0.070	
GamingExp	2	1.19	0.550	1.23	0.541	
DrivingExp	2	7.02	**0.030**	7.26	**0.027**	N:I
FullNO	2	6.55	**0.038**	6.72	**0.035**	N:S, N:I
PassedNO	2	5.15	0.076	5.33	0.070	

Table 4.14: Mean MWL Friedman analysis across all of the data set variations.

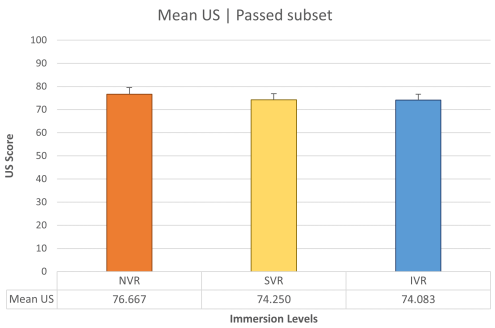
Usability (US)

Mean US for NVR was notably the highest, followed by SVR and then IVR can be seen in Figure 4.10. This trend holds for most of the data set variations except for two. In FullNO IVR overtook SVR to have the second highest US score, and in GamingExp IVR overtook NVR to have the highest US score overall. However, it should be noted that all of these differences are relatively small. Additionally, all of the SUS scores ranged between 70 to 76, which indicates they would all be considered to be between average and good SUS scores (Lewis, 2018).

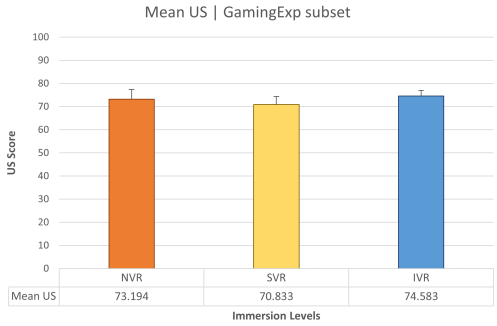
The ANOVA results for mean US can be seen in Table 4.15 and were not



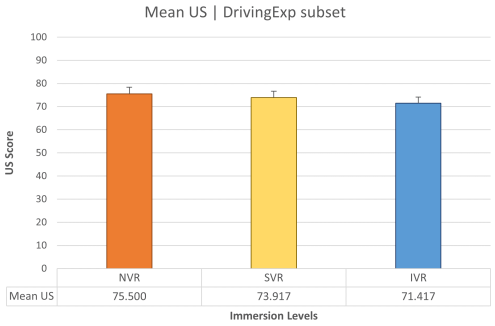
a) Full data set



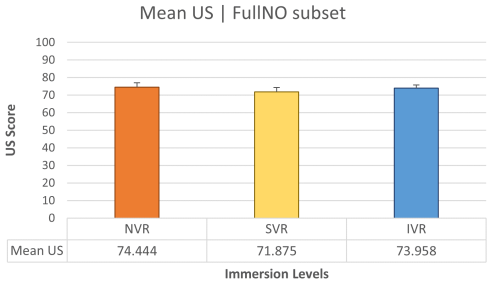
b) Passed subset



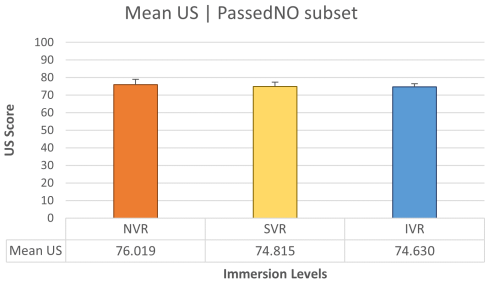
c) GamingExp subset



d) DrivingExp subset



e) FullNO subset



f) PassedNO subset

Figure 4.10: Mean US comparisons of the three immersion levels. Error bars represent SE.

significant for any of the data set variations.

Similarly, the Friedman analysis for mean US in Table 4.16 did not find any significant relationships, which indicates that the trends observed were not statistically validated.

	Power Function	Fitted Lambda	Degrees of Freedom (df)	F Statistic	p Value	Post Hoc Fisher's LSD
Full	BC	1.54	(2, 76)	1.37	0.260	
Passed	BC	1.68	(2, 58)	0.93	0.401	
GamingExp	BC	1.33	(2, 34)	0.79	0.461	
DrivingExp	BC	1.49	(2, 58)	1.48	0.237	
FullNO	BC	1.23	(2, 70)	0.97	0.385	
PassedNO	BC	1.07	(2, 52)	0.30	0.739	

Table 4.15: Mean US ANOVA analysis across all of the data set variations.

	df	H	p	H'	p'	Connover's F
Full	2	0.46	0.794	0.51	0.775	
Passed	2	0.32	0.854	0.36	0.836	
GamingExp	2	1.33	0.513	1.41	0.494	
DrivingExp	2	0.80	0.670	0.88	0.644	
FullNO	2	0.22	0.895	0.25	0.883	
PassedNO	2	0.24	0.887	0.27	0.872	

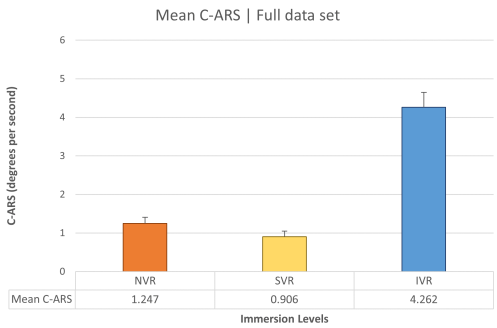
Table 4.16: Mean US Friedman analysis across all of the data set variations.

Camera's Average Rotational Speed (C-ARS)

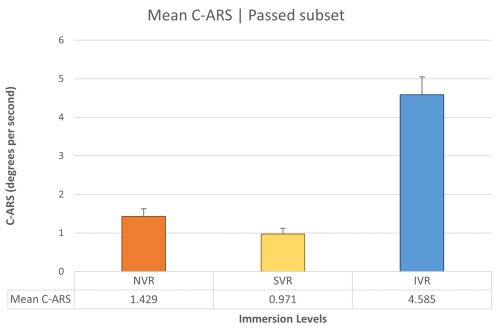
The mean C-ARS output for the data set variations can be seen in Figure 4.11.

Mean C-ARS for IVR was noticeably much higher than the other two immersion levels. This is unsurprising as the camera was affixed to the user's head so the camera movement is much more accessible and common. Even when IVR users are stationary they are still unwittingly performing some minute movement that will be registered by the simulation. Compare this to NVR and SVR where, when users stop moving, the camera movement will be registered as zero. With this in mind it becomes unremarkable that C-ARS for IVR remained significantly higher for every subset variation as well. Additionally, C-ARS was slightly lower for SVR than NVR, which is a trend that can be seen in every data set variation.

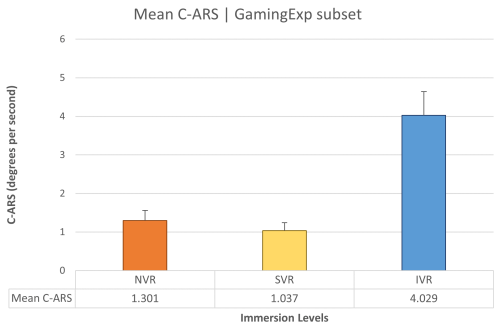
The results of the ANOVA for mean C-ARS can be seen in Table 4.17. There was strong significance for all of the data set variations, with all of them



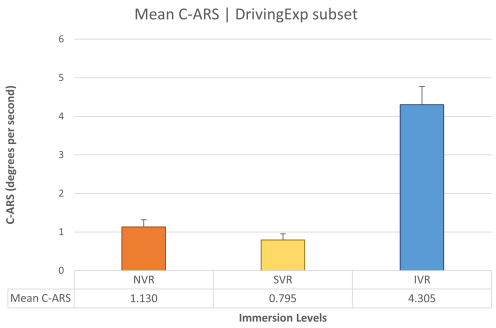
a) Full data set



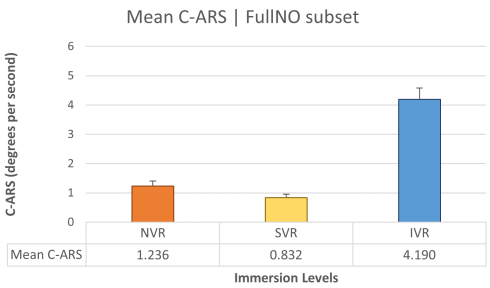
b) Passed subset



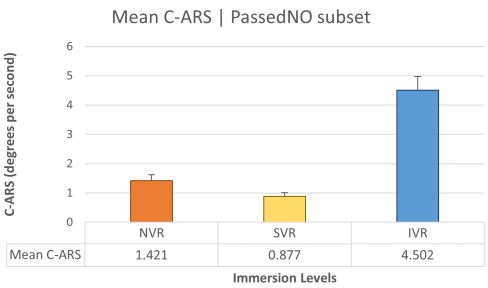
c) GamingExp subset



d) DrivingExp subset



e) FullNO subset



f) PassedNO subset

Figure 4.11: Mean C-ARS comparisons of the three immersion levels. Error bars represent SE.

also having pairwise significance between NVR and IVR, as well as SVR and IVR. As noted in the mean analysis, this is somewhat unsurprising considering IVR has some degree of constant rotational movement, whereas NVR and SVR can have zero movement.

	Power Function	Fitted Lambda	Degrees of Freedom (df)	F Statistic	p Value	Post Hoc Fisher's LSD
Full	YJ	-0.31	(2, 76)	89.07	**0.000**	N:I, S:I
Passed	YJ	-0.25	(2, 58)	62.69	**0.000**	N:I, S:I
GamingExp	YJ	-0.35	(2, 34)	28.52	**0.000**	N:I, S:I
DrivingExp	YJ	-0.41	(2, 58)	72.99	**0.000**	N:I, S:I
FullNO	YJ	-0.33	(2, 74)	88.18	**0.000**	N:I, S:I
PassedNO	YJ	-0.28	(2, 56)	61.71	**0.000**	N:I, S:I

Table 4.17: Mean C-ARS ANOVA analysis across all of the data set variations.

Camera's Number of Vantage State Changes (C-NVSC)

The differences between mean C-NVSC for each of the immersion levels were extremely shallow, as can be seen in Figure 4.12. Herein, the main apparent trend was that C-NVSC appeared to be higher for IVR in every data set variation outside of the outlier analysis.

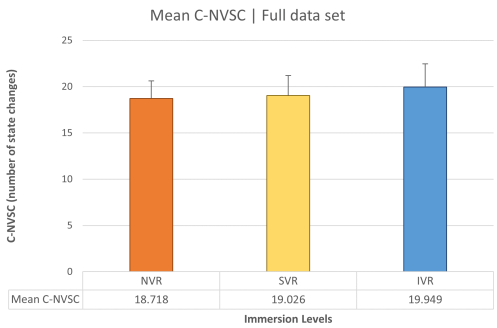
The ANOVA results for mean C-NVSC can be seen in Table 4.18. This analysis showed no significance for any of the data set variations or pairwise comparisons, which indicates that the trends observed were not statistically validated.

	Power Function	Fitted Lambda	Degrees of Freedom (df)	F Statistic	p Value	Post Hoc Fisher's LSD
Full	YJ	0.46	(2, 76)	0.22	0.803	
Passed	YJ	0.45	(2, 58)	0.12	0.891	
GamingExp	YJ	0.74	(2, 34)	0.15	0.863	
DrivingExp	YJ	0.46	(2, 58)	0.17	0.847	
FullNO	YJ	0.70	(2, 70)	0.16	0.855	
PassedNO	YJ	0.72	(2, 52)	0.06	0.939	

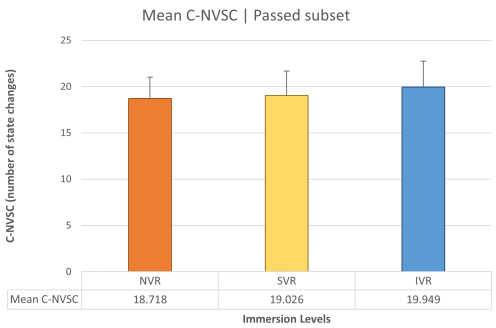
Table 4.18: Mean C-NVSC ANOVA analysis across all of the data set variations.

4.1.4 Time Series Breakdowns

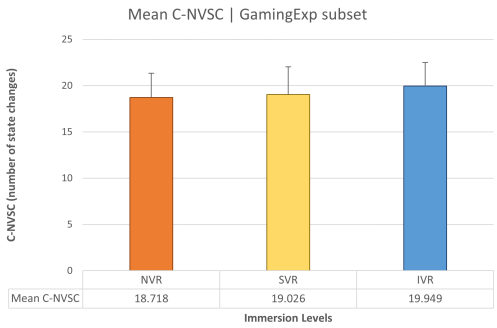
Having analyzed the means one can now look at the time series breakdowns to better understand how these DVs changed over time within each immersion level. This can only be performed on data that was recorded repeatedly as



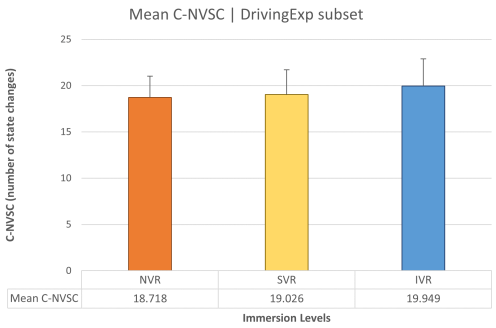
a) Full data set



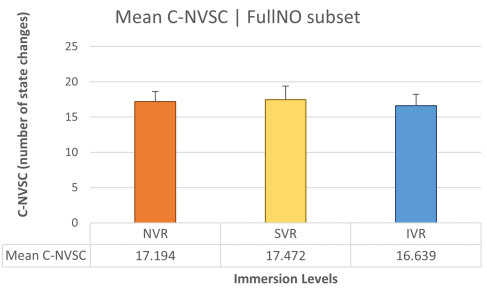
b) Passed subset



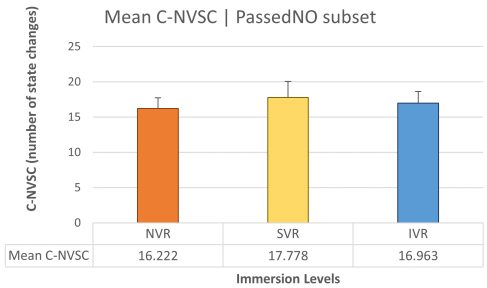
c) GamingExp subset



d) DrivingExp subset



e) FullNO subset



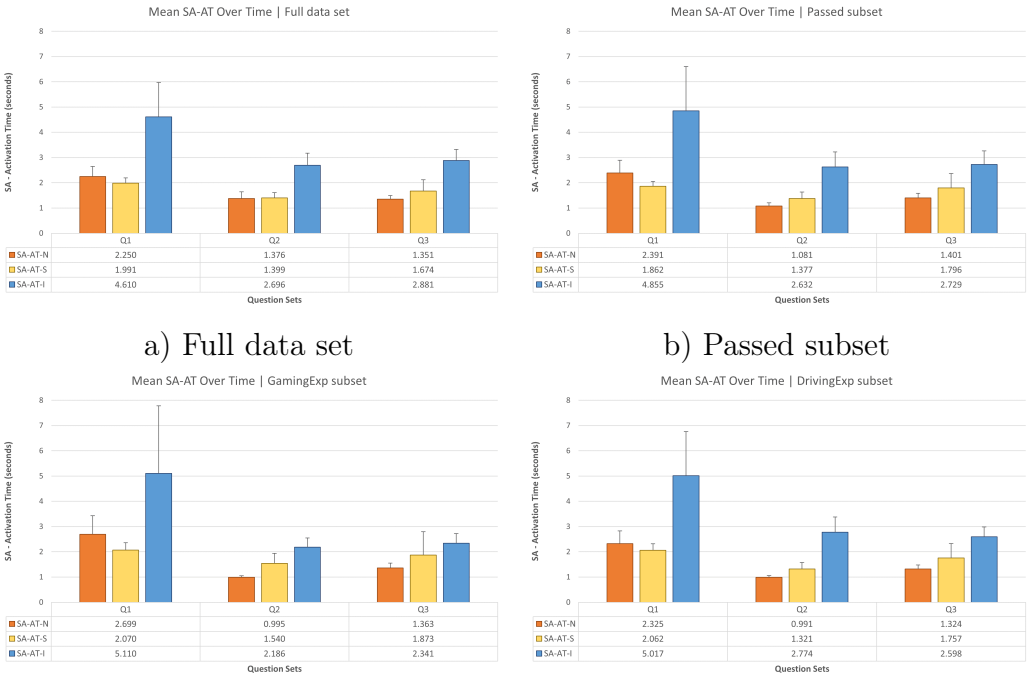
f) PassedNO subset

Figure 4.12: Mean C-NVSC comparisons of the three immersion levels. Error bars represent SE.

part of the query set, so only SA-AT, SA-PC, SA-RT, Trust, and MS will be explored in this section.

Situational Awareness - Activation Time

The time series breakdowns of SA-AT for the data set variations can be seen in Figure 4.13.



a) Full data set

b) Passed subset

c) GamingExp subset

d) DrivingExp subset

Figure 4.13: SA-AT's time series breakdown.
Error bars represent SE.

All of the immersion levels appear to show a trend of developing faster SA-AT over time, but IVR still remained markedly slower than NVR and SVR at each query period. This difference varied in the data subsets, with GamingExp showing the biggest reduction over time. Herein, by Q3, SA-AT for IVR was close to SA-AT for SVR, but was still around half a second slower. It is worth noting that by Q3 SA-AT for NVR was the fastest in every data set variation.

The two-way ANOVA for the time series breakdown can be seen in Table

4.19. Here significance was present in every subset for both Question and immersion. Notably, the post-hoc analysis showed that for the Question breakdown IVR had pairwise significance with both NVR and SVR in almost every data set variation except GamingExp. Herein, GamingExp had no significance for Q1, and only significance between SVR and IVR for Q3. In the immersion breakdown the post-hoc also showed that Q1 and Q2 had a significant relationship for NVR and SVR in almost every data set variation - except for the GamingExp subset in SVR. Q1 and Q3 also had a significant relationship for SVR in all of the data sets. IVR showed a significant relationship between Q1 and Q2 for the Full and DrivingExp data sets, but then showed no pairwise significance in the other data set variations.

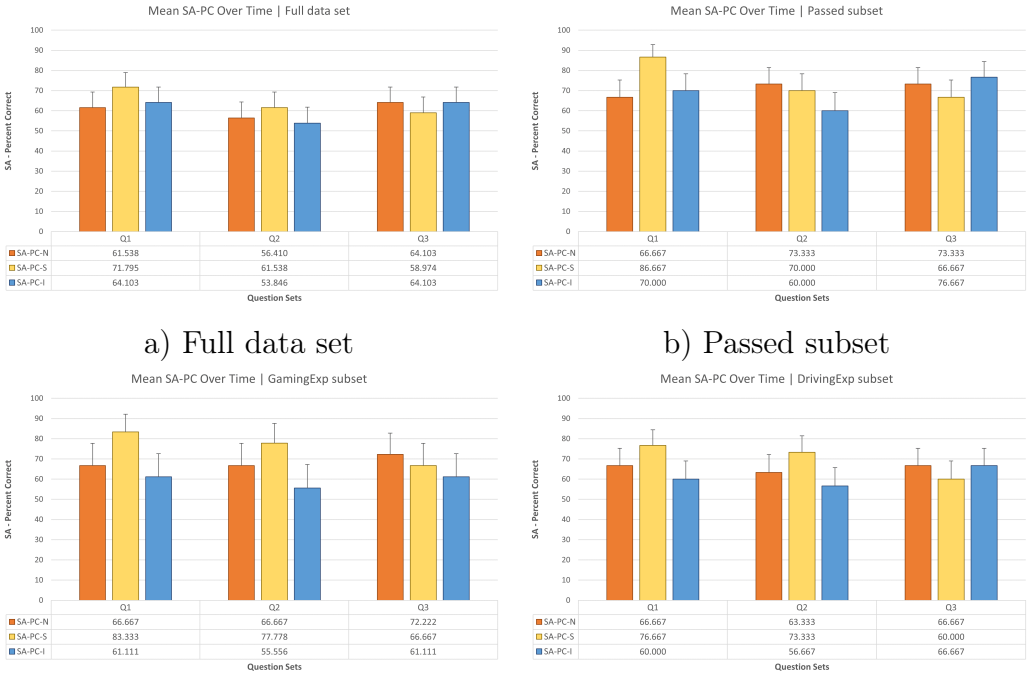
	Power Transformation	Fitted Lambda	Independent Variable (IV)	Degrees of Freedom (df)	F Statistic	p Value	Post Hoc Fisher's LSD					
							Q1	Q2	Q3	NVR	SVR	IVR
Full	BC	-0.43	Question	(2, 76)	29.51	**0.000**	N:I, S:I	N:I, S:I	N:I, S:I			
			Immersion	(2, 76)	20.18	**0.000**				Q1:Q2	Q1:Q2, Q1:Q3	Q1:Q2
Passed	BC	-0.46	Question	(2, 58)	17.19	**0.000**	N:I, S:I	N:I, S:I	N:I, S:I			
			Immersion	(2, 58)	19.00	**0.000**				Q1:Q2	Q1:Q2, Q1:Q3	
GamingExp	BC	-0.47	Question	(2, 34)	8.15	**0.001**		N:I, S:I	S:I			
			Immersion	(2, 34)	14.63	**0.000**				Q1:Q2	Q1:Q3	
DrivingExp	BC	-0.46	Question	(2, 58)	23.19	**0.000**	N:I, S:I	N:I, S:I	N:I, S:I			
			Immersion	(2, 58)	20.71	**0.000**				Q1:Q2	Q1:Q2, Q1:Q3	Q1:Q2

Table 4.19: Two-way ANOVA analysis of SA-AT's time series breakdown across all of the data set variations.

Situational Awareness - Percent Correct

The time series breakdowns of SA-PC for the data set variations can be seen in Figure 4.14.

Notably, for the Full, GamingExp, and DrivingExp data set variations, SA-PC for SVR remained higher than the other two immersion levels for the first two query sets, but dropped below them for Q3. This trend where SA-PC for SVR started off with the best results, but eventually fell below at least one of the other immersion levels by Q3, is present in every data set variation.



a) Full data set

b) Passed subset

c) GamingExp subset

d) DrivingExp subset

Figure 4.14: SA-PC's time series breakdown.
Error bars represent SE.

However, the two-way ANOVA for the time series breakdown, which can be seen in Table 4.20, showed there was no significance for either Question or immersion in any of the data set variations.

	Power Transformation	Fitted Lambda	Independent Variable (IV)	Degrees of Freedom (df)	F Statistic	p Value	Post Hoc Fisher's LSD Q1	Q2	Q3	NVR	SVR	IVR
Full	YJ	2.12	Question	(2, 76)	0.22	0.807						
			Immersion	(2, 76)	1.19	0.309						
Passed	YJ	4.22	Question	(2, 58)	0.42	0.657						
			Immersion	(2, 58)	0.52	0.597						
GamingExp	YJ	3.37	Question	(2, 34)	2.25	0.121						
			Immersion	(2, 34)	0.13	0.877						
DrivingExp	YJ	2.87	Question	(2, 58)	0.89	0.418						
			Immersion	(2, 58)	0.20	0.824						

Table 4.20: Two-way ANOVA analysis of SA-PC's time series breakdown across all of the data set variations.

The results of the Friedman test for the query set time series breakdown can be seen in Table 4.21. Here again there was notably no significance for any of the data sets.

		df	H	p	H'	p'	Connover's F
Full	Q1	2	0.50	0.779	1.04	0.595	
	Q2	2	0.27	0.874	0.54	0.764	
	Q3	2	0.15	0.926	0.36	0.834	
Passed	Q1	2	1.55	0.461	3.88	0.144	
	Q2	2	0.65	0.723	1.24	0.539	
	Q3	2	0.35	0.840	0.82	0.663	
GamingExp	Q1	2	1.08	0.582	2.36	0.307	
	Q2	2	1.00	0.607	2.18	0.336	
	Q3	2	0.25	0.883	0.55	0.761	
DrivingExp	Q1	2	0.95	0.622	2.00	0.368	
	Q2	2	0.95	0.622	1.90	0.387	
	Q3	2	0.20	0.905	0.40	0.819	

Table 4.21: Friedman analysis of SA-PC's query set time series breakdown across all of the data set variations.

The Friedman test results for the immersion level time series breakdown can be seen in Table 4.22. Once more there was no significance for any of the immersion levels in any of the data set variations, which indicates that none of the observed trends were statistically validated.

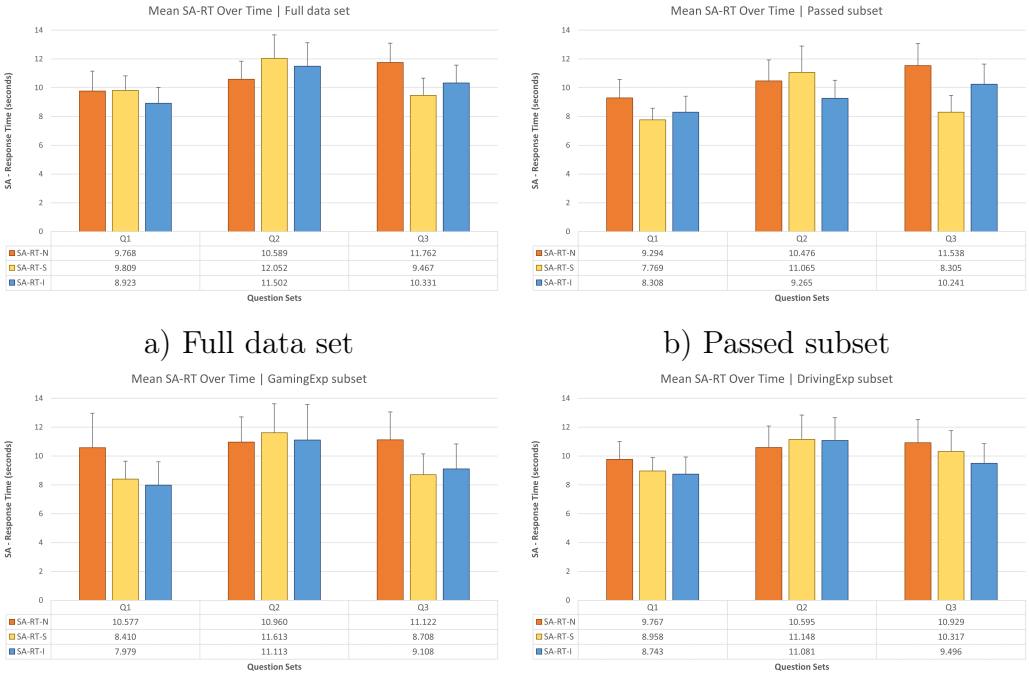
		df	H	p	H'	p'	Connover's F
Full	NVR	2	0.27	0.874	0.58	0.747	
	SVR	2	0.81	0.668	1.91	0.385	
	IVR	2	0.62	0.735	1.33	0.513	
Passed	NVR	2	0.20	0.905	0.42	0.810	
	SVR	2	1.55	0.461	3.65	0.162	
	IVR	2	0.95	0.622	1.81	0.405	
GamingExp	NVR	2	0.08	0.959	0.20	0.905	
	SVR	2	0.58	0.747	1.27	0.529	
	IVR	2	0.08	0.959	0.17	0.920	
DrivingExp	NVR	2	0.05	0.975	0.10	0.954	
	SVR	2	1.05	0.592	2.63	0.269	
	IVR	2	0.35	0.840	0.74	0.692	

Table 4.22: Friedman analysis of SA-PC's immersion level time series breakdown across all of the data set variations.

Situational Awareness - Response Time

The time series breakdowns of SA-RT for the data set variations can be seen in Figure 4.15.

SA-RT for NVR and IVR showed a trend of increasing slowness across the



a) Full data set

b) Passed subset

c) GamingExp subset

d) DrivingExp subset

Figure 4.15: SA-RT's time series breakdown.
Error bars represent SE.

data set variations. On the other hand, SA-RT for SVR appeared to show a trend of peaking in slowness during Q2 of every data set variation.

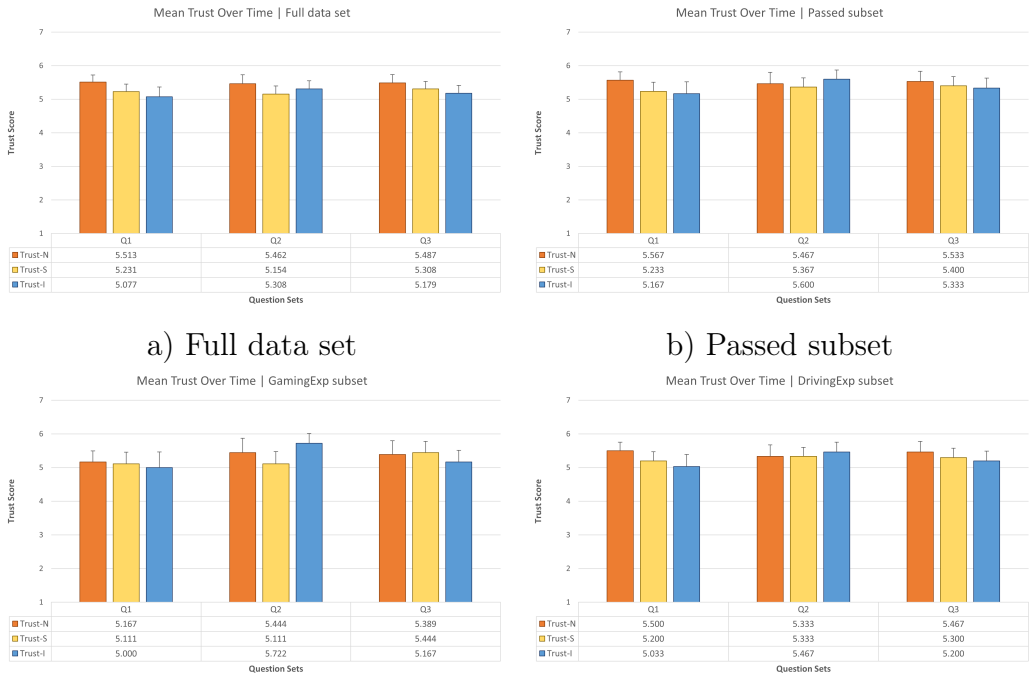
Notably, for the two-way ANOVA, there was no statistically significant relationships for either Question or immersion, as seen in Table 4.23. This indicates that none of the observed trends were statistically validated.

	Power Transformation	Fitted Lambda	Independent Variable (IV)	Degrees of Freedom (df)	F Statistic	p Value	Post Hoc Fisher's LSD					
							Q1	Q2	Q3	NVR	SVR	IVR
Full	BC	0.02	Question	(2, 76)	0.31	0.738						
			Immersion	(2, 76)	1.24	0.295						
Passed	BC	0.02	Question	(2, 58)	0.35	0.708						
			Immersion	(2, 58)	1.08	0.346						
GamingExp	BC	-0.11	Question	(2, 34)	1.21	0.311						
			Immersion	(2, 34)	1.23	0.305						
DrivingExp	BC	0.02	Question	(2, 58)	0.25	0.783						
			Immersion	(2, 58)	0.79	0.458						

Table 4.23: Two-way ANOVA analysis of SA-RT's time series breakdown across all of the data set variations.

Trust

The time series breakdowns of Trust for the data set variations can be seen in Figure 4.16.



a) Full data set

b) Passed subset

c) GamingExp subset

d) DrivingExp subset

Figure 4.16: Trust's time series breakdown.
Error bars represent SE.

There was a trend of Trust decreasing in the order NVR > SVR > IVR on Q1, which held for every data set variation. This trend was also present on Q3 for most of the data set variations except for GamingExp. Trust for Q2 was much less clear, and did not have an obvious pattern, although SVR always had the lowest Trust score for Q2 in every data set except for in DrivingExp where it was tied with NVR.

The two-way ANOVA for the time series breakdown showed no statistically significant relationships for either Question or immersion. The results of this

can be seen in Table 4.24.

	Power Transformation	Fitted Lambda	Independent Variable (IV)	Degrees of Freedom (df)	F Statistic	p Value	Post Hoc Fisher's LSD					
							Q1	Q2	Q3	NVR	SVR	IVR
Full	BC	1.76	Question	(2, 76)	1.90	0.157						
			Immersion	(2, 76)	0.10	0.906						
Passed	BC	1.95	Question	(2, 58)	0.71	0.496						
			Immersion	(2, 58)	0.84	0.438						
GamingExp	BC	1.64	Question	(2, 34)	0.13	0.878						
			Immersion	(2, 34)	1.93	0.161						
DrivingExp	BC	1.78	Question	(2, 58)	0.57	0.570						
			Immersion	(2, 58)	0.64	0.529						

Table 4.24: Two-way ANOVA analysis of Trust's time series breakdown across all of the data set variations.

The Friedman test for the query set time series breakdown proved to be a more notable analysis - the results of which can be seen in Table 4.25. Herein, none of the p values proved statistically significant, but p' for the Full data set was statistically significant. The Connover's F for the Full data set showed that pairwise significance was present between NVR and SVR, as well as NVR and IVR. This indicates that, for NVR, Trust was generally higher for the Full data set. Additionally, the Passed subset showed pairwise significance between NVR and SVR for Q1 - and the DrivingExp subset showed Q1 pairwise significance between NVR and SVR, as well as NVR and IVR. This indicates that NVR had a consistently higher Trust score for NVR on Q1.

		df	H	p	H'	p'	Connover's F
Full	Q1	2	4.67	0.097	6.56	**0.038**	N:S, N:I
	Q2	2	1.78	0.410	2.60	0.273	
	Q3	2	3.32	0.190	5.45	0.066	N:I
Passed	Q1	2	3.82	0.148	5.52	0.063	N:S
	Q2	2	0.20	0.905	0.32	0.850	
	Q3	2	1.82	0.403	3.03	0.220	
GamingExp	Q1	2	1.44	0.486	1.93	0.382	
	Q2	2	1.36	0.506	2.04	0.360	
	Q3	2	1.69	0.429	2.98	0.226	
DrivingExp	Q1	2	3.62	0.164	5.43	0.066	N:S, N:I
	Q2	2	0.07	0.967	0.10	0.951	
	Q3	2	2.15	0.341	3.44	0.179	

Table 4.25: Friedman analysis of Trust's query set time series breakdown across all of the data set variations.

The results of the Friedman test for the immersion level time series break-

down can be seen in Table 4.26. Herein, neither p nor p' showed any significance for any of the immersion levels across the data set variations. However, pairwise significance was observed between Q1 and Q2 in IVR for GamingExp. This indicates that for that specific data set's immersion level, Trust typically decreased from Q1 to Q2.

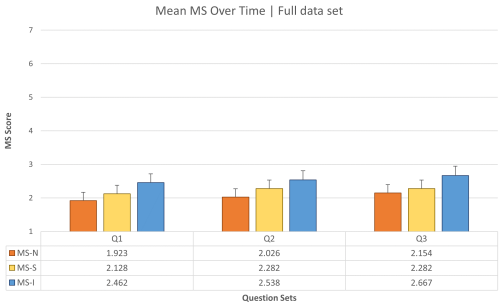
		df	H	p	H'	p'	Connover's F
Full	NVR	2	0.12	0.944	0.19	0.910	
	SVR	2	0.27	0.874	0.45	0.800	
	IVR	2	0.78	0.676	1.39	0.500	
Passed	NVR	2	0.05	0.975	0.09	0.958	
	SVR	2	0.62	0.735	1.21	0.545	
	IVR	2	1.62	0.446	3.08	0.214	
GamingExp	NVR	2	1.86	0.394	3.12	0.211	
	SVR	2	1.03	0.598	1.76	0.414	
	IVR	2	2.53	0.283	4.33	0.115	*Q1 : Q2*
DrivingExp	NVR	2	0.27	0.875	0.44	0.803	
	SVR	2	0.60	0.741	1.13	0.570	
	IVR	2	1.95	0.377	3.60	0.165	

Table 4.26: Friedman analysis of Trust's immersion level time series breakdown across all of the data set variations.

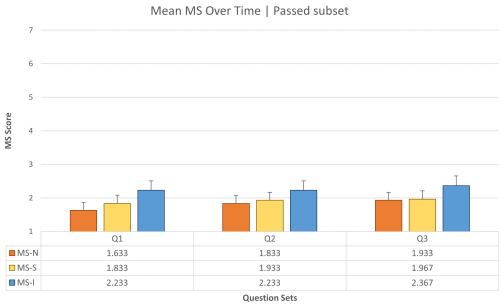
Motion Sickness

The time series breakdowns of MS for the data set variations can be seen in Figure 4.17.

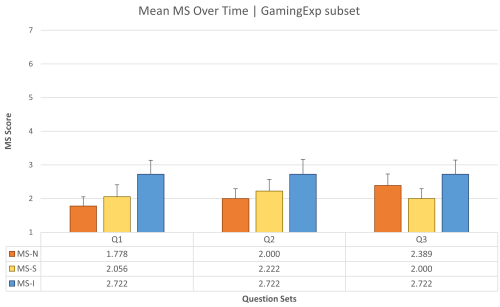
As one can note, for MS there was generally a clear ordering per question of NVR $<$ SVR $<$ IVR. There was also an increase in MS over time, although this upward trend was perhaps shallower than one might anticipate. These trends held for the most part across data subsets, except in a few instances. Firstly, for GamingExp, SVR and IVR did not show a trend of increasing. In fact, MS for SVR showed a trend of slightly decreasing, with SVR having had the lowest MS score for Q3 in that data subset. Secondly, for DrivingExp, MS for SVR was tied with MS for NVR as the lowest MS score for Q3.



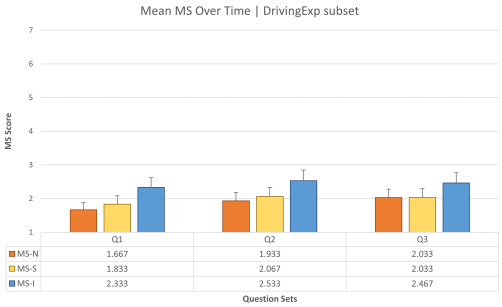
a) Full data set



b) Passed subset



c) GamingExp subset



d) DrivingExp subset

Figure 4.17: MS' time series breakdown.
Error bars represent SE.

The two-way ANOVA for the time series breakdown sheds some light on how MS was affected over time. The results of this can be seen in Table 4.27. Question had statistical significance for the Full, GamingExp, and DrivingExp data sets - and immersion had statistical significance for the DrivingExp subset. However, none of these instances showed pairwise significance.

	Power Transformation	Fitted Lambda	Independent Variable (IV)	Degrees of Freedom (df)	F Statistic	p Value	Post Hoc Fisher's LSD Q1	Q2	Q3	NVR	SVR	IVR
Full	BC	-0.71	Question	(2, 76)	4.81	**0.011**						
			Immersion	(2, 76)	2.44	0.094						
Passed	BC	-1.12	Question	(2, 58)	2.68	0.077						
			Immersion	(2, 58)	1.82	0.171						
GamingExp	BC	-0.55	Question	(2, 34)	3.30	**0.049**						
			Immersion	(2, 34)	1.13	0.336						
DrivingExp	BC	-0.91	Question	(2, 58)	4.17	**0.020**						
			Immersion	(2, 58)	3.50	**0.037**						

Table 4.27: Two-way ANOVA analysis of MS' time series breakdown across all of the data set variations.

The Friedman analysis for the query set time series breakdown can be seen

in Table 4.28. One can note how p showed no significance for any of the data set variations, but there was significance for p' on Q1 of every data set except for GamingExp. These values showed pairwise significance between NVR and IVR, suggesting that IVR had consistently higher MS for Q1.

		df	H	p	H'	p'	Connover's F
Full	Q1	2	3.28	0.194	6.17	**0.046**	N:I
	Q2	2	2.58	0.276	4.79	0.091	
	Q3	2	1.89	0.390	3.77	0.152	
Passed	Q1	2	3.32	0.191	6.75	**0.034**	N:I
	Q2	2	1.55	0.461	2.95	0.229	
	Q3	2	1.27	0.531	2.82	0.245	
GamingExp	Q1	2	2.78	0.249	5.00	0.082	
	Q2	2	2.53	0.283	4.44	0.109	
	Q3	2	2.53	0.283	4.67	0.097	
DrivingExp	Q1	2	3.52	0.172	7.03	**0.030**	N:I
	Q2	2	2.82	0.245	5.63	0.060	
	Q3	2	1.25	0.535	2.59	0.274	

Table 4.28: Friedman analysis of MS' query set time series breakdown across all of the data set variations.

The Friedman analysis for the immersion level time series breakdown can be seen in Table 4.29. This showed that none of the data set variations had discernible significance from either p or p'.

		df	H	p	H'	p'	Connover's F
Full	NVR	2	0.40	0.820	1.27	0.531	
	SVR	2	0.32	0.852	0.86	0.650	
	IVR	2	0.40	0.820	0.93	0.630	
Passed	NVR	2	0.47	0.792	1.70	0.428	
	SVR	2	0.22	0.897	0.74	0.690	
	IVR	2	0.02	0.992	0.04	0.979	
GamingExp	NVR	2	2.33	0.311	5.79	0.055	
	SVR	2	0.69	0.707	1.52	0.469	
	IVR	2	0.11	0.946	0.25	0.883	
DrivingExp	NVR	2	1.32	0.518	3.95	0.139	
	SVR	2	0.35	0.840	1.11	0.575	
	IVR	2	0.52	0.772	1.38	0.502	

Table 4.29: Friedman analysis of MS' immersion level time series breakdown across all of the data set variations.

4.2 Correlation Analysis

I analyzed the relationships between different DVs using the Pearson Correlation Coefficient. This was visualized through two primary methods in this section. The first is the analysis of the Pearson Correlation Coefficients in a full table, the second is the break down of relevant and interesting correlations by their relationship per immersion level.

4.2.1 Correlation Coefficients

This section looks at the various Pearson Correlation Coefficients associated with all of the data sets. A full list of these coefficient values across immersion levels for each of the data set variations can be seen in Table 4.30.

The correlation coefficients for NVR across the data set variations show some interesting results. MWL & US showed a very strong negative correlation across all NVR data sets. MWL had sporadic but frequent correlation with the other self-assessed values like Trust, MS, and US. Although, Trust, outside of correlating with MWL, had very little correlation with other DVs, except for SA-RT in the Full data set. In contrast to Trust, MS had much more correlation to the other DVs than one might anticipate from NVR, since NVR should not evoke a lot of MS - and this is the case across all data set variations.

Moving on to the SVR data, there are several points of note. There was once again a strong negative correlation between MWL & US across all data set variations, with MWL once more showing strong correlation with the other self-assessed DVs. Furthermore, US & Trust showed strong positive correlation across all data set variations for SVR. Trust had correlation with more DVs in this instance, although in the DrivingExp subset there is only correlation with

Correlated DVs			Full			Passed			GamingExp			DrivingExp		
			NVR	SVR	IVR	NVR	SVR	IVR	NVR	SVR	IVR	NVR	SVR	IVR
MWL	vs	US	**-0.625**	**-0.652**	**-0.578**	**-0.591**	**-0.597**	**-0.592**	**-0.766**	**-0.538**	**-0.802**	**-0.586**	**-0.577**	**-0.545**
MWL	vs	Trust	**-0.366**	**-0.519**	**-0.447**	-0.342	**-0.526**	**-0.427**	-0.235	**-0.490**	-0.376	**-0.380**	**-0.517**	**-0.420**
MWL	vs	MS	**0.409**	**0.594**	**0.505**	0.218	**0.514**	**0.408**	**0.503**	**0.617**	**0.629**	0.342	**0.583**	**0.472**
MWL	vs	SA-AT	0.178	0.076	**0.396**	0.240	0.075	**0.412**	0.325	0.154	**0.535**	0.231	0.072	0.355
MWL	vs	SA-PC	**-0.417**	-0.257	**-0.394**	-0.350	-0.184	-0.243	**-0.776**	**-0.558**	**-0.552**	**-0.391**	-0.074	-0.294
MWL	vs	SA-RT	0.290	0.089	0.242	0.138	-0.229	0.068	0.265	0.249	0.370	0.117	-0.069	0.164
MWL	vs	SA-RCS	**-0.401**	-0.289	**-0.323**	-0.304	-0.143	-0.221	**-0.483**	**-0.490**	**-0.529**	-0.322	-0.157	-0.283
MWL	vs	C-ARS	-0.281	-0.043	**-0.333**	-0.334	-0.171	-0.278	**-0.607**	0.115	-0.321	**-0.411**	-0.220	-0.322
MWL	vs	C-NVSC	-0.017	-0.314	-0.024	0.014	-0.293	-0.061	-0.291	-0.462	**-0.545**	-0.010	-0.300	-0.111
MWL	vs	Age	0.069	-0.019	0.084	0.152	0.098	0.162	-0.241	0.024	-0.115	0.075	0.049	0.107
MWL	vs	GamesWk	-0.074	-0.147	-0.136	-0.182	-0.180	-0.266	-0.246	-0.248	-0.142	-0.046	-0.111	-0.087
MWL	vs	DrivingYr	-0.114	-0.150	-0.049	-0.006	0.010	0.053	-0.418	-0.241	-0.347	-0.071	-0.031	-0.011
US	vs	Trust	0.235	**0.470**	**0.455**	0.193	**0.458**	**0.410**	0.008	**0.479**	0.133	0.218	**0.470**	**0.417**
US	vs	MS	-0.280	**-0.414**	**-0.324**	-0.238	-0.309	-0.178	-0.352	-0.425	-0.447	-0.252	-0.285	-0.235
US	vs	SA-AT	-0.193	0.234	-0.255	-0.271	0.265	-0.285	-0.343	0.302	-0.349	-0.199	0.284	-0.209
US	vs	SA-PC	0.263	0.200	0.294	0.073	0.054	0.283	**0.564**	0.356	**0.600**	0.109	0.046	0.219
US	vs	SA-RT	**-0.342**	-0.104	**-0.460**	-0.261	0.151	-0.238	**-0.475**	**-0.482**	**-0.544**	-0.160	0.121	-0.338
US	vs	SA-RCS	0.289	0.254	**0.389**	0.195	0.084	0.300	0.382	0.465	**0.574**	0.131	0.107	0.335
US	vs	C-ARS	0.312	-0.010	0.189	0.262	0.137	0.126	0.467	-0.129	0.128	**0.410**	0.189	0.166
US	vs	C-NVSC	-0.045	0.079	**-0.433**	-0.052	-0.033	**-0.467**	0.158	0.046	0.409	-0.003	-0.002	-0.357
US	vs	Age	-0.053	0.050	-0.050	-0.129	0.014	-0.150	0.246	0.278	0.233	-0.118	-0.004	-0.013
US	vs	GamesWk	-0.043	0.008	0.171	-0.007	-0.015	0.150	-0.006	0.074	0.019	-0.073	-0.032	0.114
US	vs	DrivingYr	-0.010	0.121	-0.042	-0.093	0.033	-0.196	**0.472**	**0.563**	0.401	-0.087	0.007	-0.076
Trust	vs	MS	-0.156	**-0.374**	**-0.349**	-0.046	-0.358	-0.231	-0.194	-0.467	**-0.495**	-0.231	-0.360	-0.326
Trust	vs	SA-AT	-0.008	-0.220	0.013	0.031	-0.220	0.056	0.048	-0.344	-0.175	-0.006	-0.204	0.063
Trust	vs	SA-PC	0.046	0.315	**0.354**	0.046	**0.378**	0.339	0.210	**0.674**	0.285	-0.043	0.283	0.289
Trust	vs	SA-RT	**-0.356**	-0.142	-0.278	-0.305	-0.090	-0.138	-0.131	-0.186	-0.151	-0.359	-0.067	-0.226
Trust	vs	SA-RCS	0.314	**0.406**	**0.383**	0.313	**0.418**	0.319	0.271	**0.544**	0.326	0.333	0.358	**0.371**
Trust	vs	C-ARS	0.201	0.132	0.280	0.218	0.235	0.247	0.330	0.174	**0.503**	0.139	0.234	0.263
Trust	vs	C-NVSC	-0.266	-0.090	**-0.370**	-0.276	-0.122	**-0.422**	0.147	0.104	0.310	-0.265	-0.100	-0.359
Trust	vs	Age	0.138	-0.040	0.032	0.028	-0.117	-0.009	0.433	0.084	0.214	0.155	-0.081	0.035
Trust	vs	GamesWk	0.159	0.308	0.257	0.228	0.320	0.268	0.448	**0.630**	**0.486**	0.168	0.304	0.235
Trust	vs	DrivingYr	0.128	0.016	0.129	0.023	-0.060	0.070	0.266	0.073	0.223	0.172	-0.011	0.121
MS	vs	SA-AT	**0.383**	0.280	0.262	**0.504**	0.256	0.282	**0.496**	0.282	0.262	0.342	0.277	0.192
MS	vs	SA-PC	**-0.368**	**-0.407**	-0.174	**-0.455**	**-0.412**	-0.007	-0.438	-0.354	-0.306	-0.206	-0.259	0.013
MS	vs	SA-RT	0.280	0.234	**0.465**	0.157	-0.056	0.317	0.133	0.426	**0.469**	0.185	0.078	**0.439**
MS	vs	SA-RCS	**-0.412**	**-0.407**	-0.286	**-0.383**	-0.294	-0.188	-0.405	-0.436	-0.372	-0.289	-0.236	-0.170
MS	vs	C-ARS	-0.104	-0.001	-0.204	-0.113	-0.080	-0.157	**-0.503**	0.026	-0.200	**-0.362**	-0.173	-0.145
MS	vs	C-NVSC	-0.151	**-0.325**	-0.123	-0.091	-0.309	-0.218	-0.338	**-0.680**	**-0.591**	-0.097	-0.352	-0.249
MS	vs	Age	-0.248	-0.267	-0.130	-0.275	-0.209	0.028	0.183	0.168	-0.082	**-0.362**	-0.286	-0.113
MS	vs	GamesWk	-0.026	-0.135	-0.080	-0.031	-0.166	-0.135	-0.096	-0.175	-0.296	0.124	-0.045	0.003
MS	vs	DrivingYr	**-0.320**	**-0.343**	-0.156	-0.268	-0.221	0.043	0.045	-0.042	-0.030	**-0.378**	-0.316	-0.137
SA-AT	vs	SA-PC	-0.300	-0.125	-0.191	**-0.422**	-0.120	-0.242	-0.097	-0.136	-0.405	-0.207	-0.114	-0.114
SA-AT	vs	SA-RT	-0.038	0.060	0.023	-0.028	0.003	-0.050	-0.108	-0.073	0.038	-0.008	0.044	-0.053
SA-AT	vs	SA-RCS	-0.202	-0.138	-0.187	-0.241	-0.141	-0.205	0.003	-0.114	-0.324	-0.152	-0.113	-0.145
SA-AT	vs	C-ARS	-0.127	0.271	-0.218	-0.171	0.331	-0.223	-0.326	0.289	-0.156	-0.206	0.311	-0.162
SA-AT	vs	C-NVSC	-0.161	-0.228	-0.239	-0.143	-0.267	-0.357	-0.262	-0.423	**-0.515**	-0.149	-0.248	-0.321
SA-AT	vs	Age	0.049	-0.032	0.185	0.129	0.015	0.195	0.388	0.463	0.365	0.128	-0.040	0.189
SA-AT	vs	GamesWk	0.116	-0.104	0.033	0.178	-0.095	0.018	0.196	-0.257	0.114	0.252	-0.097	0.062
SA-AT	vs	DrivingYr	-0.089	0.034	0.124	-0.005	0.087	0.155	-0.128	**0.732**	-0.100	-0.027	0.025	0.124
SA-PC	vs	SA-RT	-0.095	**-0.484**	-0.218	0.038	**-0.397**	-0.176	-0.382	-0.335	-0.358	0.190	**-0.462**	-0.091
SA-PC	vs	SA-RCS	**0.648**	**0.835**	**0.676**	**0.540**	**0.810**	**0.579**	**0.657**	**0.785**	**0.616**	**0.465**	**0.818**	**0.651**
SA-PC	vs	C-ARS	0.302	0.215	0.130	0.222	0.118	-0.034	0.368	0.159	0.292	**0.376**	0.289	0.081
SA-PC	vs	C-NVSC	0.140	0.189	0.042	0.139	0.112	-0.124	-0.033	0.143	0.413	0.195	0.152	0.125
SA-PC	vs	Age	-0.017	0.034	-0.193	-0.072	0.134	-0.155	0.146	0.016	-0.092	-0.150	-0.052	-0.234
SA-PC	vs	GamesWk	0.221	**0.336**	-0.033	0.214	0.219	-0.064	0.127	0.289	-0.008	0.145	0.230	-0.056
SA-PC	vs	DrivingYr	0.078	0.103	-0.073	-0.018	0.152	-0.044	0.025	0.083	0.085	-0.038	-0.028	-0.104
SA-RT	vs	SA-RCS	**-0.648**	**-0.708**	**-0.658**	**-0.690**	**-0.654**	**-0.730**	**-0.740**	**-0.766**	**-0.758**	**-0.658**	**-0.683**	**-0.633**
SA-RT	vs	C-ARS	-0.048	-0.038	0.086	-0.108	-0.077	0.277	-0.271	0.199	-0.085	-0.025	-0.131	0.212
SA-RT	vs	C-NVSC	**0.443**	0.004	0.172	**0.431**	0.152	-0.013	0.186	-0.122	-0.325	**0.541**	0.092	-0.029
SA-RT	vs	Age	-0.141	0.004	0.032	-0.030	0.143	0.201	-0.324	-0.255	-0.396	-0.162	0.045	0.019
SA-RT	vs	GamesWk	-0.062	-0.086	**-0.324**	-0.151	-0.075	**-0.385**	-0.168	0.036	-0.394	-0.075	-0.056	**-0.391**
SA-RT	vs	DrivingYr	-0.152	0.038	-0.055	-0.033	0.224	0.176	-0.193	-0.122	-0.446	-0.158	0.091	0.023
SA-RCS	vs	C-ARS	0.237	0.070	0.104	0.199	-0.004	-0.029	0.338	0.042	0.377	0.242	0.136	0.062
SA-RCS	vs	C-NVSC	-0.217	0.121	-0.001	-0.258	0.000	-0.070	-0.142	0.200	0.416	**-0.368**	0.093	0.067
SA-RCS	vs	Age	-0.019	0.099	-0.164	-0.125	0.125	-0.181	0.169	0.152	0.191	-0.081	0.043	-0.183
SA-RCS	vs	GamesWk	0.200	0.306	**0.367**	0.199	0.259	**0.392**	0.133	0.244	0.423	0.079	0.225	**0.417**
SA-RCS	vs	DrivingYr	0.032	0.119	-0.073	-0.105	0.090	-0.109	-0.070	0.072	0.418	-0.014	0.053	-0.114
C-ARS	vs	C-NVSC	**0.498**	0.133	0.231	**0.541**	0.104	0.233	**0.791**	-0.209	**0.615**	**0.585**	0.175	0.297
C-ARS	vs	Age	-0.124	-0.218	0.232	-0.182	-0.255	0.253	0.285	0.413	0.285	-0.117	-0.202	0.234
C-ARS	vs	GamesWk	**0.354**	0.277	0.304	**0.368**	0.260	0.357	**0.627**	0.384	**0.727**	**0.395**	0.277	0.311
C-ARS	vs	DrivingYr	-0.107	-0.222	0.209	-0.139	-0.221	0.197	0.235	0.305	0.073	-0.035	-0.171	0.224
C-NVSC	vs	Age	-0.078	-0.216	-0.128	-0.050	-0.265	-0.197	0.141	-0.316	0.035	-0.051	-0.263	-0.194
C-NVSC	vs	GamesWk	**0.343**	0.173	0.124	0.346	0.175	0.197	**0.599**	0.222	**0.490**	0.312	0.184	0.152
C-NVSC	vs	DrivingYr	-0.017	-0.092	-0.140	-0.021	-0.143	-0.110	0.113	-0.034	0.036	-0.031	-0.176	-0.126
Age	vs	GamesWk	-0.233	-0.233	-0.233	-0.231	-0.231	-0.231	0.411	0.411	0.411	-0.290	-0.290	-0.290
Age	vs	DrivingYr	**0.922**	**0.922**	**0.922**	**0.934**	**0.934**	**0.934**	**0.638**	**0.638**	**0.638**	**0.946**	**0.946**	**0.946**
GamesWk	vs	DrivingYr	-0.263	-0.263	-0.263	-0.275	-0.275	-0.275	0.173	0.173	0.173	-0.350	-0.350	-0.350

Table 4.30: Correlation coefficients across IVs and data set variations. Bold values indicate statistical significance per linear regression.

MWL & US. MS had much more frequent strong correlation with other DVs in the Full data set, whereas in the subsets it was much more limited - and in the DrivingExp subset it was negligible outside of MWL. One of the striking things in the SVR analysis was that SA-PC & SA-RT show a consistent negative correlation across most of the data set variations outside of GamingExp.

IVR also had several notable correlations. Again, MWL & US showed correlation across all data set variations, with MWL showing correlation with the other self-assessed DVs in most instances. Moreover, usability and Trust showed positive correlation in every instance except for the GamingExp subset. Trust also showed frequent instances of correlation with other DVs across most data set variations. What is especially striking about IVR is that it had some consistent correlations for MS, but less than one might anticipate. What was especially notable was that GamesWk was negatively correlated with SA-RT, but positively correlated with SA-RCS, for the Full, Passed, and DrivingExp data set variations. Given the fact that these two SA DVs had inverse value judgments this is unsurprising, but their relative consistency across data sets is notable.

Given the results of this section, one can note 15 correlations worth analyzing. These are correlations that are either directly involving the core DVs, or values that had such stark contrast or consistency across immersion levels that they necessitate further discussion. Namely these are:

- Trust vs SA-PC
- Trust vs SA-RT
- Trust vs SA-RCS
- SA-PC vs SA-RT
- MWL vs US
- MWL vs Trust
- MWL vs MS
- MWL vs SA-AT
- US vs Trust
- MS vs SA-RT
- SA-RT vs GamesWk
- SA-RCS vs GamesWk

4.2.2 Correlation Comparisons

In this section I will break down the notable Pearson Correlation Coefficient relationships across immersion levels.

Trust vs SA-PC

The correlation between Trust & SA-PC across data set variations can be seen in Figure 4.18. The relationship between these DVs was very limited, with a significant positive correlation for IVR in the Full data set, SVR in the Passed subset, and SVR in the GamingExp subset.

Trust vs SA-RT

The correlation between Trust & SA-RT across data set variations can be seen in Figure 4.19. Only NVR in the Full data set had a statistically significant negative correlation.

Trust vs SA-RCS

The correlation between Trust & SA-RCS across data set variations can be seen in Figure 4.20. These two DVs show the only consistent relationship between Trust and an SA variable, but only in the upper levels of immersion.

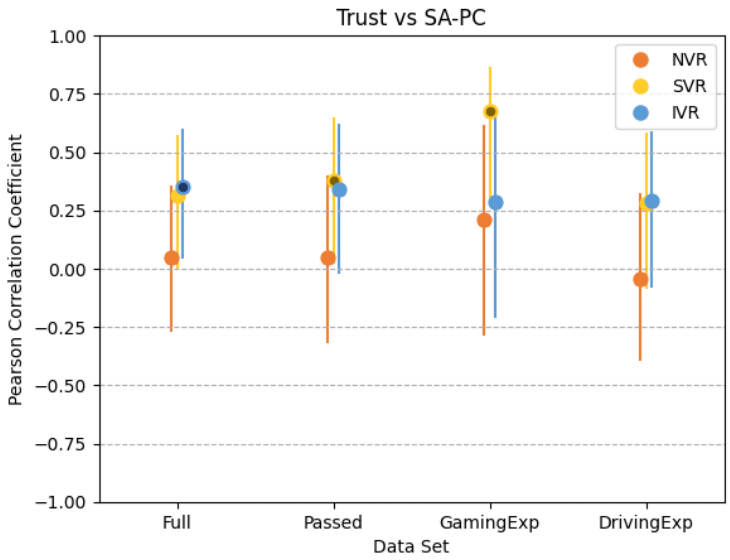


Figure 4.18: Correlation for Trust vs SA-PC.
Dark centres indicate statistical significance, and error bars indicate 95% confidence intervals.

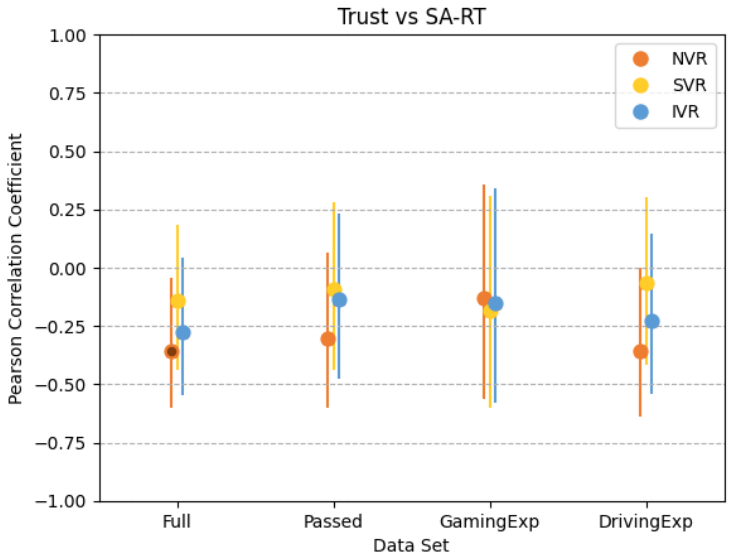


Figure 4.19: Correlation for Trust vs SA-RT.
Dark centres indicate statistical significance, and error bars indicate 95% confidence intervals.

Notably, there was a significant positive relationship between Trust & SA-RCS in both SVR and IVR in the Full data set. Additionally, there was a significant

correlation between the DVs in IVR for the DrivingExp subset, and in SVR for both the Passed and GamingExp subsets.

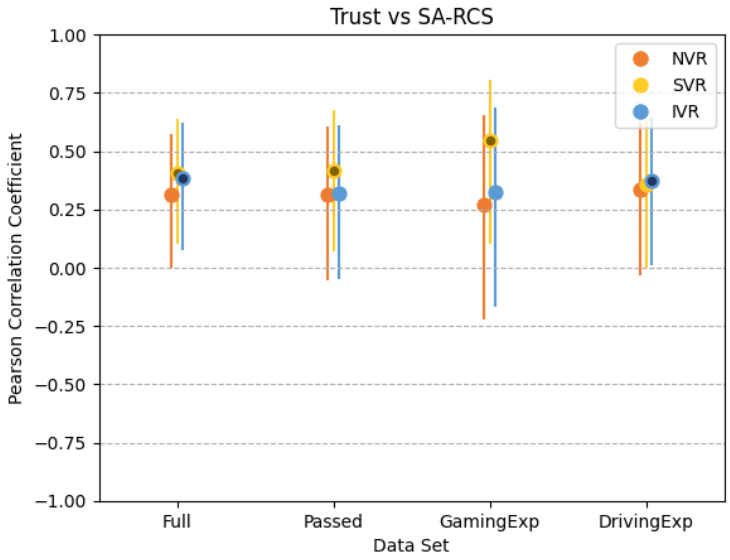


Figure 4.20: Correlation for Trust vs SA-RCS.
Dark centres indicate statistical significance, and error bars indicate 95% confidence intervals.

SA-PC vs SA-RT

The results of the correlation between SA-PC & SA-RT across data set variations can be seen in Figure 4.21. Considering SA-PC & SA-RT are both metrics of SA, one might expect a strong negative correlation across all data sets and immersion levels. Notably, significant negative correlation is only found in SVR, specifically for the Full, Passed, and DrivingExp data set variations.

MWL vs US

The results of the correlation between MWL & US across data set variations can be seen in Figure 4.22. One can note a significant negative correlation

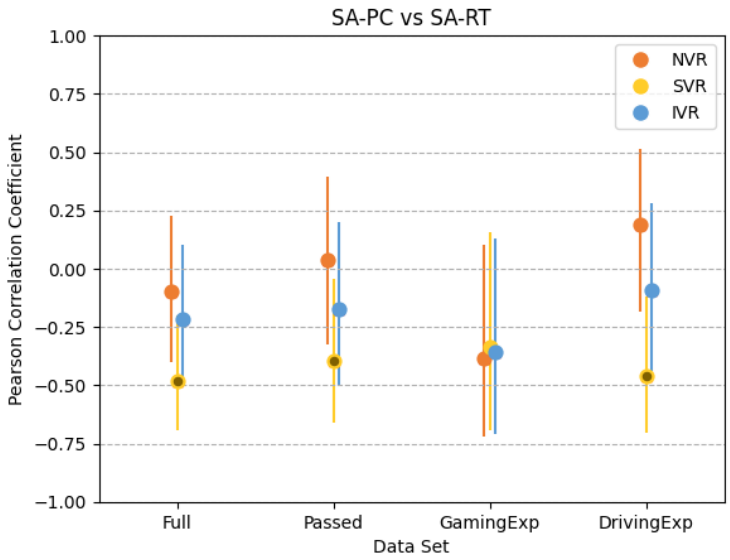


Figure 4.21: Correlation for SA-PC vs SA-RT.
Dark centres indicate statistical significance, and error bars indicate 95% confidence intervals.

for all of the immersion levels. This means that in general, both post-test questionnaires, MWL & US, had an inverse relationship.

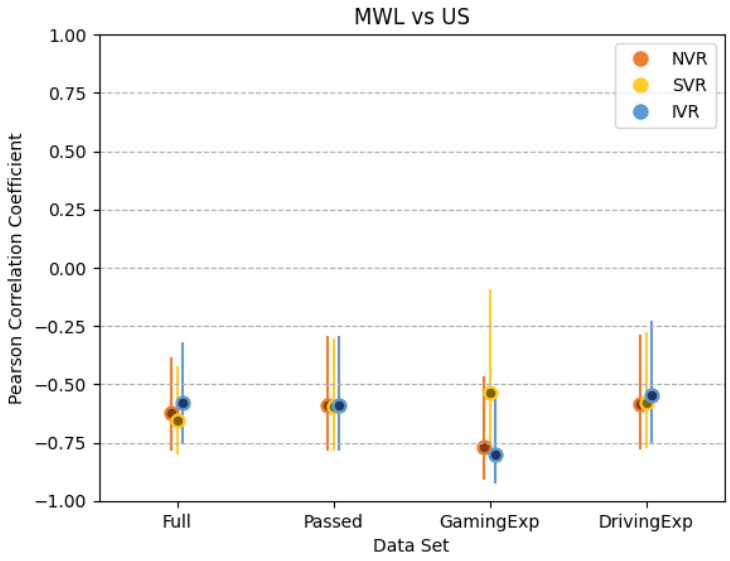


Figure 4.22: Correlation for MWL vs US.
Dark centres indicate statistical significance, and error bars indicate 95% confidence intervals.

MWL vs Trust

The results of the correlation between MWL & Trust across data set variations can be seen in Figure 4.23. Much like the relationship between US & MWL, Trust showed a strong negative correlation with MWL for the higher levels of immersion, and weaker, but still significant, negative correlation for NVR.

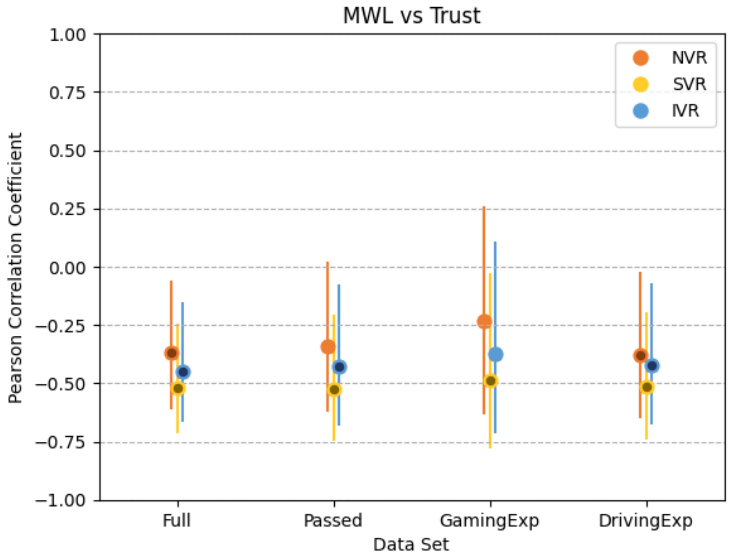


Figure 4.23: Correlation for MWL vs Trust.
Dark centres indicate statistical significance, and error bars indicate 95% confidence intervals.

MWL vs MS

The results of the correlation between MWL & MS across data set variations can be seen in Figure 4.24. For MS & MWL there was a very strong positive correlation, with all values having $p < .01$. Notably, the correlation coefficient was the strongest for SVR in all subsets. Given these results, and the results from the previous two subsections, MWL appears to have had correlation with all self-assessed DVs.

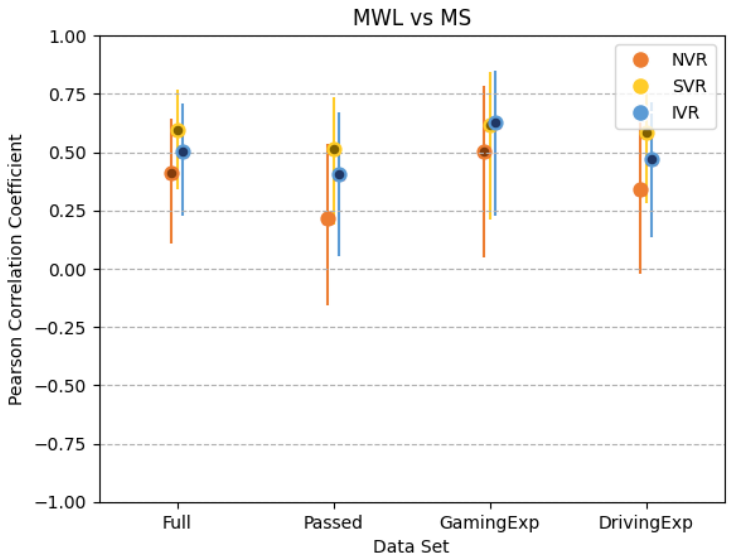


Figure 4.24: Correlation for MWL vs MS.
Dark centres indicate statistical significance, and error bars indicate 95% confidence intervals.

MWL vs SA-AT

The results of the correlation between MWL & SA-AT across data set variations can be seen in Figure 4.25. The correlation coefficients for both NVR and SVR were not significant. However, there is a fair amount of significance between these two DVs in IVR.

US vs Trust

The results of the correlation between US & Trust across data set variations can be seen in Figure 4.26. US & Trust had a significant positive correlation for the higher levels of immersion, but was not significant for NVR.

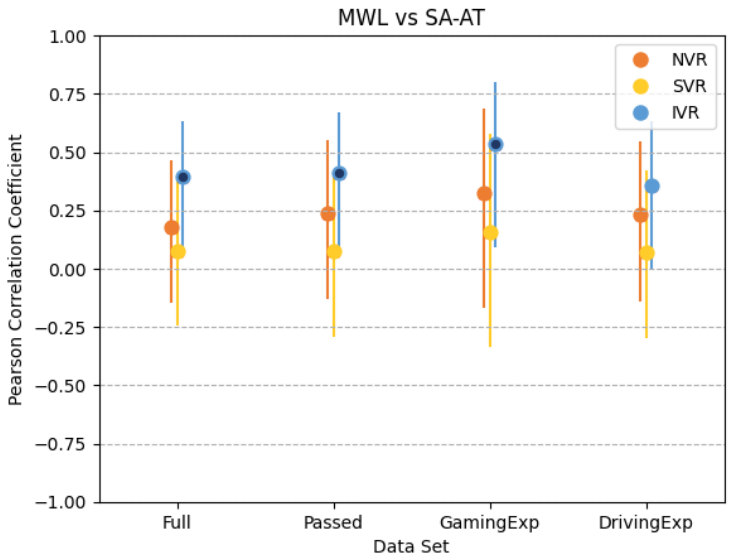


Figure 4.25: Correlation for MWL vs SA-AT.
Dark centres indicate statistical significance, and error bars indicate 95% confidence intervals.

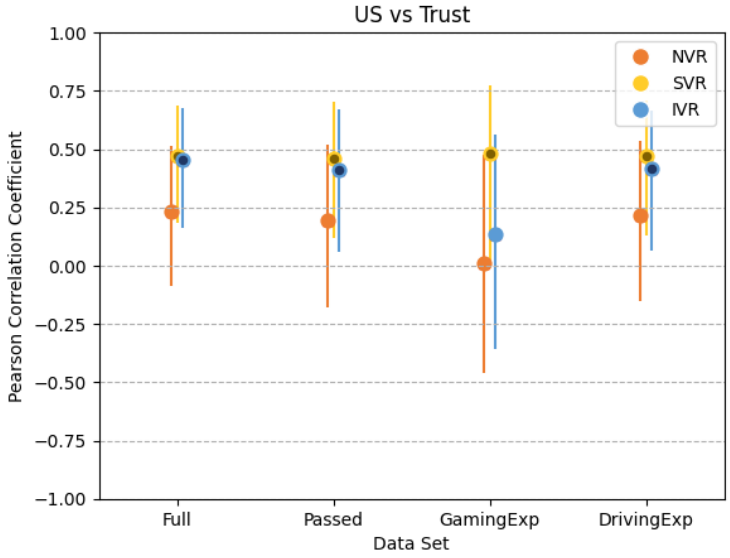


Figure 4.26: Correlation for US vs Trust.
Dark centres indicate statistical significance, and error bars indicate 95% confidence intervals.

MS vs SA-RT

The results of the correlation between MS & SA-RT across data set variations can be seen in Figure 4.27. The relationship between MS & SA-RT is notable

because it is only significant for IVR. Herein, there is a significant positive correlation in the Full, GamingExp, and DrivingExp data set variations.

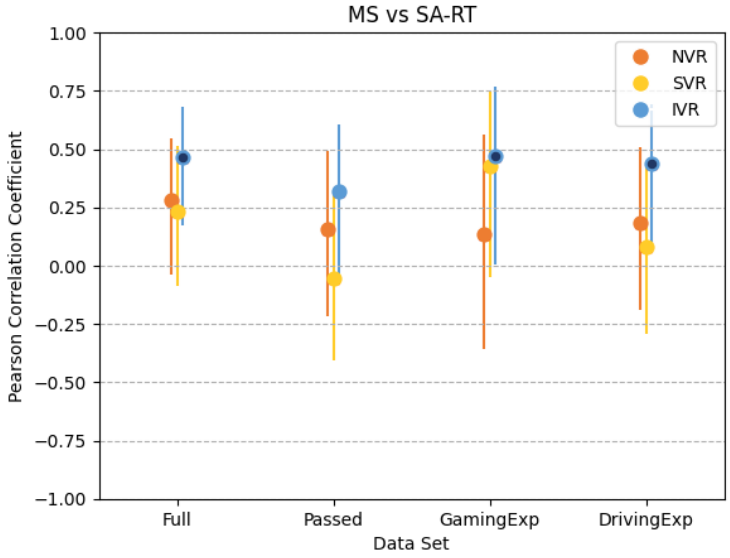


Figure 4.27: Correlation for MS vs SA-RT.
Dark centres indicate statistical significance, and error bars indicate 95% confidence intervals.

SA-RT vs GamesWk

The results of the correlation between SA-RT & GamesWk across data set variations can be seen in Figure 4.28. These two DVs only have statistically significant correlation for IVR. Specifically, IVR showed significant negative correlation in the Full, Passed, and DrivingExp data set variations.

SA-RCS vs GamesWk

The results of the correlation between SA-RCS & GamesWk across data set variations can be seen in Figure 4.29. Given that SA-RT & SA-RCS are directly related DVs it is somewhat unsurprising that SA-RCS & GamesWk

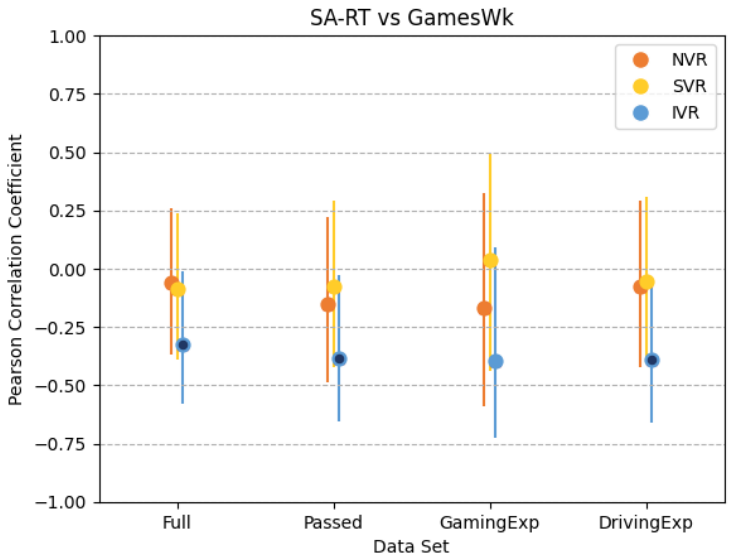


Figure 4.28: Correlation for SA-RT vs GamesWk.
Dark centres indicate statistical significance, and error bars indicate 95% confidence intervals.

showed a similar relationship to the correlation between SA-RT & GamesWk. Herein, IVR showed a significant positive correlation for the Full, Passed, and DrivingExp data set variations.

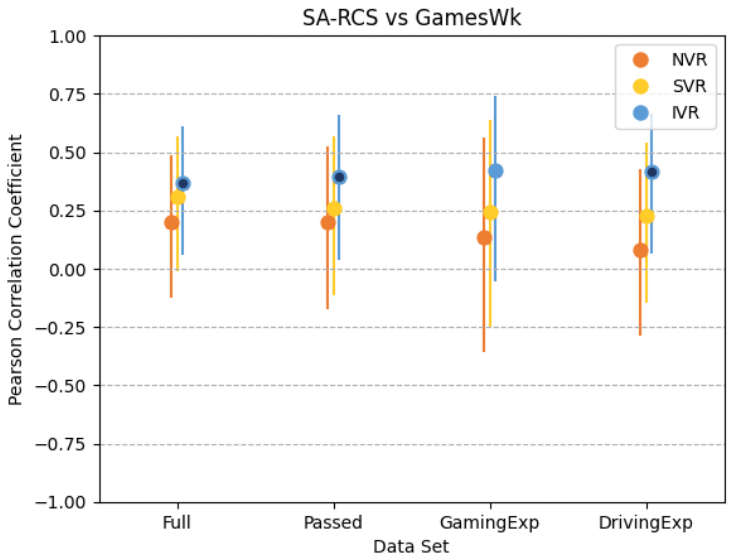


Figure 4.29: Correlation for SA-RCS vs GamesWk.
Dark centres indicate statistical significance, and error bars indicate 95% confidence intervals.

4.3 Summary

The principal manipulation in this study was the level of immersion used for the monitoring and control interface. Level of immersion most consistently impacted SA measures, Trust, MS, and MWL. This section will summarize the core outcomes of the Results chapter in terms of the impact of immersion level.

Immersion level affected SA measures. SVR showed a trend of having faster mean SA-RT than NVR in every data set. The parametric tests showed a significant effect of immersion level and a significant difference between NVR and SVR for the PassedNO subset. Accuracy, as measured by SA-PC, showed a trend of being higher for SVR, wherein the non-parametric tests showed a significantly higher SA-PC in SVR over IVR in the GamingExp subset. Further, non-parametric tests showed immersion level differences in the combined time-accuracy score of SA-RCS, which was significantly higher in SVR than IVR in

the FullNO and PassedNO subsets.

There was also a trend of higher mean Trust for NVR, with non-parametric pairwise significant differences between NVR and SVR in the Full and FullNO data sets. This held over time: Trust for NVR was generally higher than the other immersion levels over time with omnibus significance and significant differences between NVR and either SVR, IVR, or both, on Q1 and Q3 in the Full data set and some subsets.

Mean MS was significantly affected by immersion level across all data sets as indicated by omnibus parametric tests. Herein, MS increased in the order NVR $<$ SVR $<$ IVR. The non-parametric tests showed omnibus significance for MS in the PassedNO subset, and additional pairwise analysis showed that MS in NVR was significantly lower than IVR. The time series breakdown maintained the same general trend as the means, and showed parametric omnibus significance in the Full, GamingExp, and DrivingExp data sets. Additionally, the non-parametric tests showed omnibus significance on Q1 for the Full, Passed, and DrivingExp data sets - with pairwise significance between NVR and IVR.

MWL tended to be generally higher in both IVR and SVR compared to NVR. This relationship showed parametric omnibus significance in the Full and FullNO data sets, however these were essentially the same result since MWL contained no outliers. There was also non-parametric pairwise significance between what is essentially the Full and DrivingExp data sets based on the aforementioned rationale. It also showed pairwise significance between NVR and SVR in the Full data set, and pairwise significance between NVR and IVR in the Full and DrivingExp data set.

In terms of correlation between the DVs:

- Trust vs SA-PC for SVR showed a significant positive correlation for the Passed and GamingExp data sets
- Trust vs SA-RCS showed a significant positive correlation for SVR (Full, Passed, and GamingExp), and IVR (Full and DrivingExp)
- SA-PC vs SA-RT showed a significant negative correlation for SVR (Full, Passed, and DrivingExp)
- MWL vs US showed a significant negative correlation for NVR (all datasets), SVR (all datasets), and IVR (all datasets)
- MWL vs Trust showed a significant negative correlation for NVR (Full and DrivingExp), SVR (all datasets), and IVR (Full, Passed, and DrivingExp)
- MWL vs MS showed a significant positive correlation for NVR (Full and GamingExp), SVR (all datasets), and IVR (all datasets)
- MWL vs SA-AT for IVR showed a significant positive correlation for the Full, Passed, and GamingExp data sets
- US vs Trust showed a significant positive correlation for SVR (all datasets) and IVR (Full, Passed, and DrivingExp) data sets
- MS vs SA-RT for IVR showed a significant positive correlation for the Full, GamingExp, and DrivingExp data sets
- SA-RT vs GamesWk for IVR showed a significant negative correlation for the Full, Passed, and DrivingExp data sets
- Finally, SA-RCS vs GamesWk for IVR showed a significant positive correlation for the Full, Passed, and DrivingExp data sets

Chapter 5

Discussion

Having analyzed the results, the impact of these findings can now be discussed. This section will explore key findings, core takeaways, and avenues of future research.

5.1 Key Findings

This section discusses the key points of the results to better contextualize the data.

Trust had a trend of being generally higher in NVR, but only showed significance in some pairwise comparisons

I had originally hypothesized that at higher immersion levels, there should be a decrease in Trust towards a moderate, or a more calibrated, level. This is predicated on the idea that when monitoring automation, a lack of SA can lead to overconfidence, or complacency, and an inappropriately high degree of Trust (Nguyen et al., 2019). Consistent with this hypothesis, mean Trust did show a

gradual decrease as immersion increased in the full data set. However, none of the data set variations showed significance in either the parametric or non-parametric tests. The non-parametric post-hoc did show pairwise significance between NVR and SVR for some data sets, indicating that there are scenarios where Trust is consistently higher for NVR. Additionally, in the time series breakdown the full data set did show non-parametric significance for Q1, with further post-hoc analysis showing pairwise significance between NVR and SVR, as well as NVR and IVR. This indicates NVR was significantly higher in this scenario. It also appears to indicate a much more nuanced relationship, one where self-assessed Trust is generally higher for NVR over the higher immersion levels, but mostly just when it is measured at the start of the simulation. Furthermore, the fact that mean Trust in SVR and IVR are very close, but only SVR showed significantly lower values than NVR in some data sets, appears to indicate that Trust does decrease for higher immersion levels, but that Trust does not decrease strictly in tandem with immersion level. As noted earlier, calibrated Trust is the preferred outcome, as it avoids overtrust and distrust, which can both be detrimental to the safety of the system (Lee and See, 2004). As such, SVR having the only statistically significant lower Trust score, which was also closest to a moderate Trust score, makes it likely to be a preferred outcome for Trust.

Trust & SA had correlation in numerous scenarios

I had hypothesized that there would be an inverse relationship between Trust & SA, which would mean SA increased as Trust decreased - and vice versa. Herein, DVs like SA-RT would have a positive correlation with Trust - while SA-PC and SA-RCS would have negative correlation. Trust & SA did show

sporadic correlation across immersion levels, with the most notable correlation occurring between Trust & SA-RCS. Herein, there was a consistent statistically significant positive correlation at the higher levels of immersion (ie. SVR and IVR), with Trust & SA-RCS showing the strongest correlation across the most data set variations. This appears to indicate that immersion fosters some increased relationship between Trust & SA-RCS - however, there is a positive, not negative, correlation between these DVs. Regardless, the positive correlation means there is still a notable relationship between Trust & SA, which corroborates Lee and See's statement that there is a relationship between these constructs (Lee and See, 2004).

Mean SA-RT showed a shallow decrease across immersion levels, but SVR was potentially faster for technologically savvy participants

I had hypothesized that as more spatial information became available at higher immersion levels, there should be a correlative increase in SA as the participant's mental model becomes closer to the ground truth of the simulation. However, this relationship, wherein the parametric analysis only showed omnibus significance in the PassedNO data set, likewise only showed pairwise significance between NVR and SVR for the same subset. Herein, SVR showed a trend of having better SA-RT than NVR, with this relationship being statistically significant for one of the subsets that indicate technical proficiency. Thus, there is some inclination that SA for SVR improves with technical proficiency.

SA-AT was consistently worse in IVR

One of the most notable and consistent results from the study was how much slower participants accepted query set prompts in IVR. This can be partly

attributed to HMD controllers being substantially different from the mice used in the NVR and SVR setups. However, the fact that IVR was still significantly slower in the data subsets that indicate higher technical proficiency could mean this would remain a consistent issue even for trained industry professionals. Although, MWL & SA-AT showed a significant correlation in three of the data sets, which could indicate that mitigating the impact of MWL for SCCOs may improve SA-AT for IVR. A major limitation of this study was that to maintain continuity with the other immersion levels the console had to be stationary in the world space. SA-AT could potentially become faster if the console tracked with the users line of sight. However, this could bring its own host of issues as fixating a pointer on a button that has the potential to move could make acquiring a target more difficult.

MS increased with immersion level

These data indicated that mean MS increased with immersion level. The non-parametric tests showed omnibus significance in the PassedNO subset, and pairwise significance was detected between NVR and IVR in the same subset. As such, MS was significantly higher for IVR in this instance. The parametric tests showed omnibus significance in every data set variation, but no pairwise significance in any of these same instances. The failure to detect pairwise differences in the parametric tests is likely attributable to the relatively short time of the simulations. Essentially, as the simulations were roughly 8 minutes long there may not have been enough time for the effects of sensory conflict to compound and be reflected by self-assessed MS - resulting in shallow differences between immersion levels for MS. As such, one would expect the difference in MS between immersion levels to become more clear and detectable in longer

simulations. Regardless, IVR appears to be the worst option when it comes to MS. This outcome relates to the observations of Martirosov et al. (2021), who saw MS increase in tandem with immersion level.

MS & SA-RT were consistently correlated for IVR

MS & SA-RT only showed significant positive correlation for IVR. This appears to indicate that as MS increases participants tend to respond to SPAM questions slower. Given that MS is expected to increase, over time and at higher levels of immersion, this could pose a problem for SCCOs in IVR - as SA-RT, and in turn their SA, could be severely impacted by MS.

US & Trust only had correlation for higher immersion levels

US & Trust exhibited some surprisingly clear correlation. SVR having persistent significant positive correlation for these two DVs, and IVR having significant positive correlation in every data set variation except GamingExp, shows that the participants' sense of US was generally correlated to their level of Trust for the higher immersion levels. Understanding that US & Trust have a potential relationship for immersive displays could prove to be a boon for SCC designers. However, for SVR and IVR, there is also a potential risk that users who find the interface too usable may be prone to overtrust the system - which numerous authors in the literature note is a dangerous state for an operator to be in when teaming with an autonomous robot (Lee and See, 2004; Inagaki and Itoh, 2013; Hoff and Bashir, 2014). Ideally, US would be very high and Trust would be calibrated based on the situation, but never at the highest level. This means an extremely positive correlation could be a poor outcome that SCC designers will need to look out for.

All SA accuracy metrics showed better mean performance in SVR

SA-PC & SA-RCS are generally inconsistent across data set variations, except for the fact that every mean analysis of SA-PC & SA-RCS shows a higher score for SVR. However, no omnibus significance was shown in any of the parametric or non-parametric tests for either SA-PC or SA-RCS. With that said, the non-parametric post-hoc test did show pairwise significance between SVR and IVR for SA-PC in the GamingExp subset. Additionally, there was pairwise significance for SA-RCS between SVR and IVR in both of the outlier removed subsets. This finding possibly indicates that participants with technical proficiency have higher SA-PC & SA-RCS for SVR, but the lack of omnibus significance means this needs to be explored further.

SA-PC & SA-RT were consistently negatively correlated, but only in SVR

Given that both SA-PC & SA-RT are metrics for SA performance that have inverse value judgements, one might anticipate that SA-PC & SA-RT have consistent inverse correlation across immersion levels. However, significant negative correlation only appears to be present for SVR, and even then is only present in three of the data set variations. A strong negative correlation between these DVs is a preferable outcome, as the SCCO would need to be both fast and accurate in their assessments of ship safety (Yoshida et al., 2020). This would seem to suggest that SVR is preferable for SA from this perspective. However, it should be noted that in GamingExp all of the immersion levels have a negative correlation, but none of them show statistical significance. This could indicate that the relationship between SA-PC & SA-RT being stronger

in SVR would become negligible at higher levels of technical proficiency, or that the GamingExp subset is too small.

MWL was worse for higher levels of immersion, but this difference became negligible in SVR for technologically savvy participants

Immersion level had a significant effect on MWL for several data set variations, wherein, it was lower for NVR than SVR and IVR. However, this relationship appears to become negligible for SVR in all of the subsets that select for at least some level of technical proficiency. This likely indicates that risk of higher MWL for SVR, relative to NVR, would become a mitigated factor for trained professionals.

MWL was a strong indicator of self-assessed values

MWL had strong correlation with all of the other self-assessed DVs across all immersion levels. These relationships were both positive and negative depending on the value judgement of the DV compared to MWL. As such, the RTLX appears to be a powerful self-assessed sentiment tool that would be useful for assessing any future SCC interface designs, regardless of the immersion level used in SCCs.

5.2 Takeaways

These data have illustrated a lot of interesting factors regarding immersion, and with said factors it becomes easier to anticipate what level of immersion would be the most useful in industrial scenarios. However, the key takeaway from this research becomes not "What level of immersion is the most useful?",

but "Which level of immersion is the least useful?". Herein, I would say that IVR, so HMDs, are likely the least useful level of immersion for this application, which was indicated by a variety of factors.

MS proved to be a consistent issue for IVR in both the mean and the time series breakdown. While this issue was anticipated to be a much more severe issue between immersion levels than what was observed, it was still a persistent one for IVR. Moreover, for variables that account for SA accuracy, SA-PC & SA-RCS, IVR consistently performs worse than SVR. Most notable however is the impact of IVR on SA-AT, which was significantly slower across all data sets. The fact that this DV was relatively unaffected by the participants' technical proficiency would seem to indicate that HMD remote response is unavoidably slower than a mouse. This makes sense for two reasons, the first being that most people interact with computers everyday, and as such have far more experience using mice than they do HMD controllers. The second reason is the time it takes the cursor to fixate on the target, a concept that can be thought of as the time to acquire a target. This is very similar to the concept of "motion time" used in the original HCI paper by English et al. (1967), wherein they explored the effectiveness of different types of computer controllers - including the time it took for the controllers to reach an on-screen target. For this thesis, in NVR and SVR, the cursor is always somewhere on the main console, and as such, never particularly far away from the question box. Conversely, to enable freedom of movement, the IVR pointer could be anywhere within the panoramic range, and as such, could be comparatively far away from the question box. As mentioned, having an interface that follows the users' gaze could mitigate this - but also an interface in motion would likely bring its own set of target acquisition issues for SA-AT, not to mention SA-RT.

This a potential drawback for IVR, especially in an industrial environment where every second could be valuable for mitigating risk. It is important to note that MWL & SA-AT showed a statistically significant correlation in most of the data sets, which may indicate that mitigating the impact of MWL for SCCOs would improve SA-AT for IVR. However, MWL for IVR appeared to be relatively high, specifically compared to NVR, a relationship which showed statistical significance in two data sets. Given that MWL for IVR also showed significant correlations with Trust, MS, US, and SA-AT, it appears to be a variable that has strong relationships with numerous other human factors variables, and as such is a potential risk if it becomes too high. In terms of potential benefits, it was anticipated that IVR would have the lower, more moderate, Trust consistently across data set variations. There did appear to be a general trend of Trust being lower for IVR, at least compared to NVR, across the data sets in both the mean and time series breakdown. This trend did show statistical significance in the time series breakdown specifically, mostly for Q1 across three data sets. As such, some indication of this relationship was present, although it might breakdown with time or system exposure. Notably, IVR did have some statistically significant correlations between Trust & SA metrics. Herein, IVR had a significant correlation between Trust & SA-RCS in the two data set variations. In addition to this, US & Trust showed a significant relationship in three of the data sets, which, as long it does not become too high, is a potentially valuable relationship to be aware of for SCC designers. The biggest potential benefit for IVR was that it showed strong positive correlations between SA-RT & GamesWk, as well as SA-RCS & GamesWk - which may indicate that SA for IVR would benefit greatly from system exposure and training. However, IVR's strong correlation between MS & SA-RT, especially

in the GamingExp and DrivingExp subsets, may also indicate that SA-RT for IVR would become progressively slower in the event of increasing MS - despite technical proficiency. As such, the SA of SCCOs could see a negative impact from prolonged IVR usage. With all of that said, IVR appears to be a system with minimal benefits and numerous points of concern for maritime remote monitoring.

Moving on, NVR had a number of interesting factors that made it stand out from the other immersion levels - most notably it had consistently lower MWL and MS. However, given that NVR is very similar to the desktop computer setup an individual could have at home, it is unsurprising that participants seemed to find it less taxing. One of the potential issues for NVR was that it had a higher level of Trust than both of the other immersion levels, which showed pairwise significance in various scenarios from the mean and time series breakdown - although, for the latter, these were predominately on Q1. This appears to indicate that NVR is prone to overtrust relative to the other immersion levels, at least for the start of the simulation. Another potential issue for NVR was that SA-RT appeared to be significantly slower than it was for SVR in the PassedNO subset. This could indicate that individuals with at least a base level of SA awareness, which one would expect an SCCO to have, would then have improved SA in SVR over NVR. When one considers this, along with the fact that Trust only had significant correlation with SA-RT and none with any of the SA accuracy metrics, a picture begins to emerge - one where a stand-in SCCO may feel more comfortable due to the setup's familiarity, which can lead them to have higher Trust in general. However, this Trust is potentially representative of overconfidence and complacency. Herein, their level of SA is worse but their Trust is greater.

SVR on the other hand was quite different from NVR. Much the opposite, it had the high MWL and a middling level of MS. As such, it at first seems the worse option between SVR and NVR. However, SVR had a trend of having higher mean SA-PC & SA-RCS values, which showed some pairwise significance in various data sets. Moreover, SA-RT for SVR showed omnibus and pairwise significance between NVR and SVR for the PassedNO subset. Building on this, SA-PC & SA-RT were only significantly correlated for SVR, which poses a problem for the other immersion levels - as one would want both fast response times and high accuracy in systems with high risk like an SCC. With these factors in mind it would appear that participant SA is generally stronger for SVR. Trust showed a trend of being lower for SVR than NVR, a relationship which showed pairwise significance in two data sets. This appears to indicate that participants were perhaps more skeptical when using SVR. It also turned out that SA and Trust were shown to have a relationship in numerous scenarios. Specifically, Trust & SA-PC, along with Trust & SA-RCS, had a significant positive correlation in SVR for several data set variations. I had originally anticipated a negative correlation between these two DVs, but the fact that Trust and SA accuracy metrics for SVR showed a relationship here at all is important to note. However, much like other correlations with Trust, an overtly high SA might lead to relative overtrust for SVR. If this is true, future SCC designers would need to be cognizant of that relationship when designing SCC interfaces. Moving on to another Trust correlation for SVR - US & Trust showed a significant correlation, which for the same reasons with IVR, is an interesting and valuable point to consider for future SCC design. However, it might also be a factor to keep an eye on, since a very high positive correlation could be indicative of a situation were SCCOs who the find system very usable

will be prone to overtrust. Finally, MWL for SVR showed a trend of being higher than for NVR, but this only showed omnibus and pairwise significance in, what is essentially, the Full data set. This appears to indicate that the issue of relatively high MWL may become negligible at higher levels of technical proficiency. If true, this would be a positive since MWL showed significant correlation with Trust, MS, and US for SVR - so keeping MWL mitigated would be beneficial to avoid distrust, high MS, and a low sense of US. With all of these factors considered, one can note a system that is generally more taxing to users compared to NVR, but also one that holds potential benefits for the core variables of SA and Trust. Given that the eLumens projector for SVR had the lowest luminance, poor contrast, and was easily the oldest of all the immersion systems used, at over 20 years old, it is surprising that it performed as well as it it did. Also, it might mean that a newer CAVE, or other immersive projection display system, would show improved results for the core variables.

As can be noted, the differences between NVR and SVR were much more nuanced since both systems have pros and cons that contrast with each other to a degree. However, this appears to boil down to a preference between a familiar and easy to learn system versus one with better SA and a closer to calibrated Trust given the scenario. I feel that in an industrial remote monitoring environment higher SA and calibrated Trust should be paramount, and as such would advocate for the use of SVR style systems in the future implementation of SCCs.

5.3 Future Research

The long-term goal of this research project is to work toward the development of Intelligent Adaptive Interface (IAI) for teleoperated robotics - ones you might expect to find in SCCs. There are a number of avenues that should be explored to build upon this research in the direction of IAI development. However, there are four specifically that will be crucial.

The first is the impact of augmented reality in SVR and/or NVR. This will be important for understanding if other aspects of MR are beneficial for teleoperation. Technologies such as the Magic Leap 2 and the HoloLens 2 could prove a useful way to improve SA. This could work by having a background SVR or NVR system with the SCCO using an AR headset for data overlay. It would allow the SCCO to have interactive console components such as maps and LiDAR readings that are malleable and interactive. This is not dissimilar from the system discussed by Heffner and Rødseth (2019).

The second is integration with actual robotics and artificial intelligence (AI) systems. Building for real world systems means that this work needs to make the jump to integration with real world robotics systems. This will necessitate work with a robotics group, wherein the ideal system for bridging between robotics systems and the software interface can be determined.

Third is the testing of the system in a correlate to a real world icy water environment. Even if MASS does not get implemented in the arctic, any sort of implementation in Canada would require the ability to navigate icy waters, since many of Canada's major shipping routes, such as the St. Lawrence Seaway, have significant ice coverage for large portions of the year. To do this sort of testing, the aforementioned NRC-OCRE ice tank in St. John's,

Newfoundland & Labrador could be used to create a scaled MASS simulation. This would entail creating a miniature MASS vessel that would relay data to an SCC as it navigated the ice tank.

Finally, there is the testing and incorporation of Explainable AI (XAI) or Interpretable Machine Learning (IML) technologies. When a relationship with a robotics group has been established, it will be paramount to integrate with their systems so as to convey abstract decisions on the part of the autonomous vessels in a way that is transparent and comprehensible to the SCCO. This will serve to mitigate risk by giving the SCCO better SA. Exploring these avenues should enable the research group to begin preliminary designs of an IAI system.

It should be noted however that even with research into these areas, it will be difficult to make any sort of substantial advances in designing IAIs for MASS systems specifically until consistently active MASS vessels operating at an IMO automation level of 3 exist.

5.4 Conclusion

As automation becomes more omnipresent, there is a growing requirement to understand human-robot relationships, especially with autonomous robots that requires some level of monitoring.

This research illustrated the relative pros and cons to various levels of immersion for use in the remote monitoring of MASS systems. Here one can see that IVR introduces a number of strong negative factors and only minimal benefits to SA and Trust. NVR represents a system that invokes less MWL and MS, but one prone to overconfidence and lower SA. Conversely, SVR represents a system with moderate, or a more calibrated, Trust and better SA in general,

but one that can invoke marginally higher MS, and higher MWL in users with less technical experience. It appears that determining which system is optimal is a much more nuanced decision, especially for SVR and NVR, and will depend largely on a designer's needs and requirements. However, in an industrial teleoperation scenario, where the mitigation of risk is paramount, having a higher SA and calibrated Trust is likely more important than mitigating a marginal increase in MWL and MS - especially when MWL can potentially be reduced by improved technical proficiency. As such, I would advocate for the use of an SVR-style system in future SCCs.

All in all, this work represents a requisite step for the construction of future SCC stations. While limited by nature of being a user study on a simulation, this research helps to narrow down the optimal level of immersion for increasing the safety and efficiency of MASS systems - and will ideally aid in the development of SCC interfaces and IAIs. As such, this research holds potential value for various MASS stakeholders in both academia and the private sector.

References

Hironori Akiduki, Suetaka Nishiike, Hiroshi Watanabe, Katsunori Matsuoka, Takeshi Kubo, and Noriaki Takeda. 2003. Visual-vestibular conflict induced by virtual reality in humans. *Neuroscience Letters* 340, 3 (2003), 197–200. https://doi.org/10.1016/S0304-3940(03)00098-3

G. E. P. Box and D. R. Cox. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society* 26, 2 (1964), 211–252. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

John Brooke. 1986. SUS: A 'Quick and Dirty' Usability Scale. In *Usability Evaluation In Industry*, Patrick W. Jordan, B. Thomas, Ian Lyall McClelland, and Bernard. Weerdmeester (Eds.). CRC Press, Boca Raton, Florida, 189–194. https://doi.org/10.1201/9781498710411-35

Ernesto A. Bustamante and Randall D. Spain. 2008. Measurement Invariance of the Nasa TLX. *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting* 52, 19 (2008), 1522–1526. https://doi.org/10.1177/154193120805201946

Canada. 2005. *Manual of Ice (MANICE).*

https://www.canada.ca/en/environment-climate-change/services/weather-manuals-documentation/manice-manual-of-ice.html

Canada. 2017. Arctic Shipping Safety and Pollution Prevention Regulations. (2017). https://laws-lois.justice.gc.ca/eng/regulations/SOR-2017-286/index.html

Canada. 2020. Enhanced Satellite Communication Project - Polar. (2020). http://dgpaapp.forces.gc.ca/en/defence-capabilities-blueprint/project-details.asp?id=1279

Canada. 2022. Ocean, Coastal and River Engineering Research Centre. (2022). https://nrc.canada.ca/en/research-development/research-collaboration/research-centres/ocean-coastal-river-engineering-research-centre

James C. Cunningham, Henri Battiste, Sam Curtis, Elyse C. Hallett, Martin Koltz, Summer L. Brandt, Joel Lachter, Vernol Battiste, and Walter W. Johnson. 2015. Measuring Situation Awareness with Probe Questions: Reasons for not Answering the Probes. *Procedia Manufacturing* 3 (2015), 2982–2989. https://doi.org/10.1016/j.promfg.2015.07.840

Francis T. Durso, M. Kathryn Bleckley, and Andrew R. Dattel. 2006. Does Situation Awareness Add to the Validity of Cognitive Tests? *Human Factors: The Journal of the Human Factors and Ergonomics Society* 48, 4 (2006), 721–733. https://doi.org/10.1518/001872006779166316

Francis T. Durso and Andrew R. Dattel. 2004. SPAM: the real-time assessment of SA. In *A Cognitive Approach to Situation Awareness: Theory and Appli-*

cation, S. Tremblay and S. Banbury (Eds.). Ashgate Publishing, Aldershot, UK, 137–154.

Francis T. Durso, Carla. A. Hackworth, Todd R. Truitt, Jerry Crutchfield, Danko Nikolic, and C. A. Manning. 1999. Situation awareness as a predictor of performance in en route air traffic controllers. *Scientific and technical aerospace reports* 37 (1999), 1–11. https://www.faa.gov/data_research/research/med_humanfacs/oamtechreports/1990s/media/am99-03.pdf

H. Dybvik, E. Veitch, and M. Steinert. 2020. Exploring Challenges With Designing And Developing Shore Control Centers (SCC) For Autonomous Ships. *Proceedings of the Design Society: DESIGN Conference* 1 (2020), 847–856. https://doi.org/10.1017/dsd.2020.131

Mica R. Endsley. 1995. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37, 1 (1995), 32–64. https://doi.org/10.1518/001872095779049543

Mica R. Endsley. 2019. A Systematic Review and Meta-Analysis of Direct Objective Measures of Situation Awareness: A Comparison of SAGAT and SPAM. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 63, 1 (2019), 124–150. https://doi.org/10.1177/0018720819875376

William K. English, Douglas C. Engelbart, and Melvyn L. Berman. 1967. Display-Selection Techniques for Text Manipulation. *IEEE Transactions on Human Factors in Electronics* HFE-8, 1 (1967), 5–15. https://doi.org/10.1109/THFE.1967.232994

Eduardo Jose Fabris, Vicenzo Abichequer Sangalli, Leonardo Pavanatto Soares, and Marcio Sarroglia Pinho. 2021. Immersive telepresence on the operation of unmanned vehicles. *International Journal of Advanced Robotics Systems* 18, 1 (2021), 1–13. https://doi.org/10.1177/1729881420978544

Andrzej Felski and Karolina Zwolak. 2020. The Ocean-Going Autonomous Ship—Challenges and Threats. *Journal of Marine Science and Engineering* 8, 1 (2020), 1–16. https://doi.org/10.3390/jmse8010041

Marlena R. Fraune, Ahmed S. Khalaf, Mahlet Zemedie, Poom Pianpak, Zahra NaminiMianji, Sultan A. Alharthi, Igor Dolgov, Bill Hamilton, Son Tran, and Z. O. Toups. 2021. Developing Future Wearable Interfaces for Human-Drone Teams through a Virtual Drone Search Game. *International Journal of Human-Computer Studies* 147 (2021), 1–16. https://doi.org/10.1016/j.ijhcs.2020.102573

Robert M. Gash, Kevin A. Murrant, Jason W. Mills, and David E. L. Millan. 2020. Machine Vision Techniques for Situational Awareness and Path Planning in Model Test Basin Ice-Covered Waters. *Global Oceans 2020: Singapore – U.S. Gulf Coast* (2020), 1–8. https://doi.org/10.1109/IEEECONF38699.2020.9389165

Gene V. Glass, Percy D. Peckham, and James R. Sanders. 1972. Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research* 42, 3 (1972), 237–288. https://doi.org/10.3102/00346543042003237

Kasper Hald, Matthias Rehmn, and Thomas B. Moeslund. 2020. Human-Robot Trust Assessment Using Motion Tracking & Gal-

vanic Skin Response. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2020), 6282–6287. https://doi.org/10.1109/IROS45743.2020.9341267

P. A. Hancock, Theresa T. Kessler, Alexandra D. Kaplan, John C. Brill, and James L. Szalma. 2020. Evolving Trust in Robots: Specification Through Sequential and Comparative Meta-Analyses. *Human Factors: The Journal of the Human Factors and Ergonomics Society* (2020), 1–34. https://doi.org/10.1177/0018720820922080

Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909

Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology* 52 (1988), 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Kevin Heffner and Ørnulf Jan Rødseth. 2019. Enabling Technologies for Maritime Autonomous Surface Ships. *Journal of Physics: Conference Series* 1357 (2019), 1–12. https://doi.org/10.1088/1742-6596/1357/1/012021

Kevin Anthony Hoff and Masooda Bashir. 2014. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 3 (2014), 407–434. https://doi.org/10.1177/0018720814547570

IMO. 2021. Autonomous ships: regulatory scoping exercise completed. (2021).

https://www.imo.org/en/MediaCentre/PressBriefings/pages/MASSRSE2021.aspx

Toshiyuki Inagaki and Makoto Itoh. 2013. Human's Overtrust in and Overreliance on Advanced Driver Assistance Systems: A Theoretical Framework. *International Journal of Vehicular Technology* 2013 (2013), 1–8. https://doi.org/10.1155/2013/951762

Behrang Keshavarz and Heiko Hecht. 2011. Validating an Efficient Method to Quantify Motion Sickness. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53, 4 (2011), 415–426. https://doi.org/10.1177/0018720811403736

Kongsberg. 2022. Autonomous Ship Project, Key Facts About Yara Birkeland. https://www.kongsberg.com/maritime/support/themes/autonomous-ship-project-key-facts-about-yara-birkeland/

John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

James R. Lewis. 2018. Measuring Perceived Usability: The CSUQ, SUS, and UMUX. *International Journal of Human-Computer Interaction* 34, 12 (2018), 1148–1156. https://doi.org/10.1080/10447318.2017.1418805

Shayne Loft, Daniel B. Morrell, and Samuel Huf. 2013. Using the situation present assessment method to measure situation awareness in simulated submarine track management. *International Journal of Human Factors and Ergonomics* 2, 1 (2013), 33–48. https://doi.org/10.1504/IJHFE.2013.055975

Shayne Loft, Daniel B. Morrell, Kate Ponton, Janelle Braithwaite, Vanessa Bowden, and Samuel Huf. 2016. The Impact of Uncertain Contact Location on Situation Awareness and Performance in Simulated Submarine Track Management. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 58, 7 (2016), 1052–1068. https://doi.org/10.1177/0018720816652754

Sergo Martirosov, Marek Bureš, and Tomáš Zítka. 2021. Cyber sickness in low-immersive, semi-immersive, and fully immersive virtual reality. *Virtual Reality* 26, 1 (2021), 15–32. https://doi.org/10.1007/s10055-021-00507-4

Laura R. Marusich, Jonathan Z. Bakdash, Emrah Onal, Michael S. Yu, James Schaffer, John O'Donovan, Tobias Höllerer, Norbou Buchler, and Cleotilde Gonzalez. 2016. Effects of Information Availability on Command-and-Control Decision Making: Performance, Trust, and Situation Awareness. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 58, 2 (2016), 301–321. https://doi.org/10.1177/0018720815619515

Stephanie M. Merritt. 2011. Affective Processes in Human–Automation Interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53, 4 (2011), 356–370. https://doi.org/10.1177/0018720811411912

Ziaul Haque Munim. 2019. Autonomous ships: a review, innovative applications and future maritime business models. *Supply Chain Forum: An International Journal* 20, 4 (2019), 266–279. https://doi.org/10.1080/16258312.2019.1631714

Abdeldjallil Naceri, Dario Mazzanti, Joao Bimbo, Yonas T. Tefera, Domenico Prattichizzo, Darwin G. Caldwell, Leonardo S. Mattos, and Nikhil Deshpande. 2021. The Vicarios Virtual Reality Interface for Remote Robotic

Teleoperation. *Journal of Intelligent & Robotic Systems* 101, 80 (2021), 1–16. https://doi.org/10.1007/s10846-021-01311-7

Ho Namgung and Joo-Sung Kim. 2021. Collision Risk Inference System for Maritime Autonomous Surface Ships Using COLREGs Rules Compliant Collision Avoidance. *IEEE Access* 9 (2021), 7823–7835. https://doi.org/10.1109/ACCESS.2021.3049238

Thanh Nguyen, Chee Peng Lim, Ngoc Duy Nguyen, Lee Gordon-Brown, and Saeid Nahavandi. 2019. A Review of Situation Awareness Assessment Approaches in Aviation Environments. *IEEE Systems Journal* 13, 3 (2019), 3590–3603. https://doi.org/10.1109/JSYST.2019.2918283

Stephen Palmisano, Robert S. Allison, and Juno Kim. 2020. Cybersickness in Head-Mounted Displays Is Caused by Differences in the User's Virtual and Physical Head Pose. *Frontiers in Virtual Reality* 12 (2020), 1–24. https://doi.org/10.3389/frvir.2020.587698

Gerben Peeters, Gökay Yayla, Tim Catoor, Senne Van Baelen, Muhammad Raheel Afzal, Christos Christofakis, Stijn Storms, René Boonen, and Peter Slaets. 2020. An Inland Shore Control Centre for Monitoring or Controlling Unmanned Inland Cargo Vessels. *Journal of Marine Science and Engineering* 8, 10 (2020), 1–27. https://doi.org/10.3390/jmse8100758

Kimberly A. Pollard, Ashley H. Oiknine, Benjamin T. Files, Anne M. Sinatra, Debbie Patton, Mark Ericson, Jerald Thomas, and Peter Khooshabeh. 2020. Level of immersion afects spatial learning in virtual environments: results of a three-condition within-subjects study with long intersession intervals. *Virtual Reality* 24 (2020), 783–796. https://doi.org/10.1007/s10055-019-00411-y

Carolyn C. Preston and Andrew M. Colman. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica* 104, 1 (2000), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

A. Rizzo and S. T. Koenig. 2017. Is clinical virtual reality ready for primetime? *Neuropsychology* 31, 8 (2017), 877–899. https://doi.org/10.1037/neu0000405

Rolls-Royce. 2016. Rolls-Royce Reveals Future Shore Control Centre. https://www.rolls-royce.com/media/press-releases/2016/pr-2016-03-22-rr-reveals-future-shore-control-centre.aspx

Yuri A. Romanov, Nina A. Romanova, and Peter Romanov. 2011. Shape and size of Antarctic icebergs derived from ship observation data. *Antarctic Science* 24, 1 (2011), 77–87. https://doi.org/10.1017/S0954102011000538

Rana Saha. 2021. Mapping competence requirements for future shore control center operators. *Maritime Policy & Management* (2021), 1–13. https://doi.org/10.1080/03088839.2021.1930224

Nikolai Smolyanskiy and Mar Gonzalez-Franco. 2017. Stereoscopic First Person View System for Drone Navigation. *Frontiers in Robotics and AI* 20 (2017), 1–10. https://doi.org/10.3389/frobt.2017.00011

Christoph Alexander Thieme and Ingrid Bouwer Utne. 2017. A risk model for autonomous marine systems and operation focusing on human–autonomy collaboration. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability* 231, 4 (2017), 446–464. https://doi.org/10.1177/1748006x17709377

Hans Van Den Broek, Jaco Griffioen, and Monique Van Der Drift. 2020. Meaningful Human Control in Autonomous Shipping: An Overview. *IOP Conference Series: Materials Science and Engineering* 929 (2020), 1–12. https://doi.org/10.1088/1757-899x/929/1/012008

André Vandierendonck. 2016. A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods* 49, 2 (2016), 653–673. https://doi.org/10.3758/s13428-016-0721-5

Dan J. Woltz and Christopher A. Was. 2006. Availability of related long-term memory during and after attention focus in working memory. *Memory Cognition* 34, 3 (2006), 668–684. https://doi.org/10.3758/BF03193587

Anqi Xu and Gregory Dudek. 2015. OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (2015), 221–228. https://doi.org/10.1145/2696454.2696492

Yara. 2021. Yara to start operating the world's first fully emission-free container ship. https://www.yara.com/corporate-releases/yara-to-start-operating-the-worlds-first-fully-emission-free-container-ship/

In-Kwon Yeo and Richard A. Johnson. 2000. A new family of power transformations to improve normality or symmetry. *Biometrika* 87, 4 (2000), 954–959. https://doi.org/10.1093/biomet/87.4.954

Masanori Yoshida, Etsuro Shimizu, Masashi Sugomori, and Ayako Umeda. 2020. Regulatory Requirements on the Competence of Remote Operator in Maritime Autonomous Surface Ship: Situation Awareness, Ship

Sense and Goal-Based Gap Analysis. *Applied Sciences* 10, 23 (2020), 1–27. https://doi.org/10.3390/app10238751

Masanori Yoshida, Etsuro Shimizu, Masashi Sugomori, and Ayako Umeda. 2021. Identification of the Relationship between Maritime Autonomous Surface Ships and the Operator's Mental Workload. *Applied Sciences* 11, 5 (2021), 1–23. https://doi.org/10.3390/app11052331