EXPLORING VIRTUAL REALITY FLIGHT TRAINING AS A VIABLE ALTERNATIVE TO TRADITIONAL SIMULATOR FLIGHT TRAINING

by

Agata Lawrynczyk

A thesis submitted to

the Faculty of Graduate and Postdoctoral Affairs

in partial fulfillment of

the requirements for the degree of

MASTER OF APPLIED SCIENCE

School of Computer Science

at

CARLETON UNIVERSITY

Ottawa, Ontario

September, 2018

Abstract

Upwards of 60% of preventable aviation accidents are due to human error, suggesting that sufficient training is vital for improved flight safety. The objective of the present work was to inform the use of virtual reality technology as a flight simulation method that supports training of high-risk scenarios and has major cost saving potential compared to traditional flight simulators.

Participants flew three predefined circuits in a traditional flight simulator graphics condition (Broad Angle Display System, (BADS)), and conducted the same flight path and tasks using a VR graphics environment. The exploratory hypotheses in this research are that there would be no differences in the user experience, cognitive workload, and performance, between the two graphics conditions. User experience was gathered by a questionnaire probing motion sickness and other VR usability metrics. Cognitive load was gathered from a subjective rating in a questionnaire, derived from peripheral detection task, and by continuously measuring participant's physiological indices, including heart rate and galvanic skin response. Performance was obtained from flight path and airspeed deviations, and from flying precision.

USER EXPERIENCE. There were no differences in physiological symptoms (i.e., queasiness, dizziness, or disorientation) after the first flying condition, however queasiness ratings appeared greater for participants who flew in the VR graphics condition on second exposure. **COGNITIVE LOAD**. The average subjective mental workload rating, heart rate, and galvanic skin response were statistically higher in the VR graphics condition than in the BADS. The remaining cognitive load metrics did not have any significant differences between the two graphics conditions. **PERFORMANCE**. The average airspeed root-mean-squared error (RMSE) was statistically greater (worse) in the BADS graphics condition than in the VR. There were no differences in the altitude or heading RMSE, or flying precision between the two graphics conditions. The groups experiencing queasiness and dizziness symptoms in VR on first exposure both had a greater heading RMSE in VR. There were no significant effects from queasiness, dizziness, or disorientation on performance measures in the BADS group.

The current research is the first to compare the user experience, cognitive load, and performance metrics in flight training between BADS and VR graphics conditions. The main takeaway of this research suggests that although the user experience and performance metrics were comparable, the VR experience likely causes an increased cognitive load on users compared to the BADS. Although it is premature to conclude it as a viable replacement, the many advantages of VR outweigh its disadvantages, suggesting those interested in flight training could consider it a viable alternative while being cognizant of possible side effects.

Table of Contents

Abstract ii

Table of Contents iii

List of Figures ii

List of Tables iii

1. Introduction 5

1.1. Motivation for the Present Work 5

1.2. Research Questions 7

1.3. Thesis Outline 7

2. Background 8

2.1. User Experience in Virtual Reality 8

2.1.1. An Interaction-Centered Framework of Experience 8

2.1.2. Immersion and Presence 10

2.1.3. Hardware 10

2.1.4. Software Latency 11

2.1.5. Motion Sickness 12

2.1.5.1. Definitions 12

2.1.5.2. Theories of Cybersickness 13

2.1.5.3. Contributing Factors 13

2.1.5.4. Measuring Cyber Sickness 15

2.2. Cognitive Load 16

2.2.1. Performance and Cognitive Load 16

2.2.2. Measuring Cognitive Load 17

2.2.2.1. Subjective Measures 17

2.2.2.2. Task-Performance-Based Measures 18

2.2.2.3. Physiological Measures 19
2.3. Performance Metrics 22
2.4. Present Research 22
3. User Study Methodology 24
3.1. Participants 24
3.2. Technical Set-Up 25
3.3. Procedure 27
3.4. Experimental Flight Measures 27
3.4.1. User Experience 28
3.4.2. Cognitive Load 28
3.4.3. Performance Metrics 30
4. Results 31
4.1. High Level Overview 31
4.2. User Experience 33
4.2.1. Adverse Symptoms after First Exposure 33
4.2.1.1. Adverse Symptoms after Second Exposure 33
4.2.2. Subjective VR Experience 33
4.3. Cognitive Load 36
4.3.1. Subjective Rating 36
4.3.2. Secondary Task - PDT 37
4.3.3. Physiological Measures 38
4.3.3.1. Heart Rate 38
4.3.3.2. Galvanic Skin Response 39
4.4. Performance Metrics 41
4.4.1. Flight Path Deviations 41
4.4.2. Airspeed Deviation 42

4.4.3. Hoop Precision 43
4.4.4. Impact of Adverse Symptoms in VR on Flight Performance 43
5. Discussion and Conclusion 45
5.1. Overview 45
5.2. Research Questions 45
5.2.1. User Experience 45
5.2.2. Cognitive Load 47
5.2.3. Performance Metrics 51
5.3. Takeaways and Recommendations 52
5.4. Limitations and Future Work 53
6. Bibliography 55
Appendix A – Factors that Contribute to Adverse Effects in VR 66
Appendix B – Pre and Post-Test Questionnaire 69

List of Figures

Figure 2-1: Summary of a framework of user experience as it relates to interactive systems (Forlizzi & Battarbee, 2004) 9
Figure 3-1: Virtual reality product usage (left) and frequency playing video games (right) 24
Figure 3-2: Queasiness from playing video games (left) or with VR products (right) 25
Figure 3-3: Oculus Rift headset, used for VR condition 25
Figure 3-4: Graphics for BADS/non-VR and VR condition (Left to Right) 26
Figure 3-5: Empatica E4 wrist watch to measure heart rate and GSR, and thumb switch for PDT 26
Figure 3-6: Overview of measures captured during/after each graphics condition 28
Figure 3-7: Empatica E4 PPG sensor (Empatica, 2016) 29
Figure 3-8: Pendleton Airfield flight circuit with hoops 30
Figure 4-4: Comfort rating Oculus Rift at onset and towards end of flying experience 34
Figure 4-5: Perceived heaviness of Oculus Rift 34
Figure 4-6: Perceived crispness of visuals in Oculus Rift 35
Figure 4-7: Perceived realness of visuals in Oculus Rift 35
Figure 4-8: "How would you describe the fact that you could not see your hands on the yoke?" 36
Figure 4-9: "How would you describe the fact that you could not see your hands on the throttle or flaps?" 36
Figure 4-10: Self-reported mental effort in BADS vs VR. Error bars indicate 95% credible intervals. 37
Figure 4-11: Mean heart rate in BADS vs VR. Error bars indicate 95% Credible Intervals. 38
Figure 4-12: Average # of SCR Peaks between BADS and VR graphics condition (left) and average amplitude sum of SCR between BADS and VR (right). Error bars indicate 95% Credible Intervals. 39
Figure 4-13: Main effect of gender on # of SCR peaks. Error bars indicate a 95% confidence interval. 41
Figure 4-14: Altitude RMSE and heading RMSE in the BADS versus VR. Error bars indicate 95% credible interval. 42
Figure 4-15: Airspeed RMSE in BADS and VR condition. Error bars indicate 95% credible interval. 42
Figure 4-17: Flying precision as indicated by the number of hoops flown through the center (left), and near the periphery (right) in BADS and VR condition. Error bars indicate 95% credible interval. 43

List of Tables

TABLE 4-1: HIGH LEVEL OVERVIEW OF RESULTS 32
TABLE 4-3: PEARSON CORRELATIONS TO INVESTIGATE MODERATING EFFECTS OF GRAPHICS CONDITION ON MENTAL EFFORT 37
TABLE 4-4: PEARSON CORRELATIONS TO INVESTIGATE MODERATING EFFECTS OF GRAPHICS CONDITION ON HEART RATE 39
TABLE 4-5: PEARSON CORRELATIONS TO INVESTIGATE MODERATING EFFECTS OF GRAPHICS CONDITION ON SCR AMPLITUDE AND SCR PEAKS 40
TABLE 4-6: PEARSON CORRELATIONS TO INVESTIGATE MODERATING EFFECTS OF GRAPHICS CONDITION ON AIRSPEED RMSE 43

1. Introduction

1.1. Motivation for the Present Work

Human error can lead to catastrophic results in aviation. Between 1959 and 2015, upwards of 60% of preventable aviation incidents were attributed to pilot error (Boeing, 2015), suggesting that access to sufficient training for pilots is integral to improving flight safety for all aboard. Cost and safety concerns limit the use of live flight training for many procedures, enabling the case for leveraging flight simulators. The present work was conducted to inform the use of virtual reality (VR) technology as a flight simulation method that supports training of high-risk scenarios and has major cost saving potential. Specifically, the present research examined the user experience, cognitive load, and performance outcomes between a traditional flight simulator and a VR graphics environment.

Simulation is a general term applied to reproducing real world tasks, especially in the context of training or research, when the actual task is too dangerous, difficult or costly to conduct (Salas, 2017), making aviation an ideal domain for engaging simulation. Aviation procedures can be extremely dangerous to practice in live flight (e.g., engine failures, etc.) and the cost of practicing routine procedures is often prohibitive where an hour of live flight can cost well over $10,000 including cost of aircraft rental, instructor, fuel, etc. (Pausch, Crea, & Conway, 1992). A full-flight simulator is an enclosed unit, typically on a six-degrees of freedom motion platform (although may also be on a stationary platform), that contains a cockpit and controls identical to an actual aircraft, and contains realistic sounds and motions associated with flying.

Full-flight simulators can be an effective and efficient way to provide pilot training (Johnson, 2005). They permit pilots to experience simulated conditions, including high-risk scenarios that would be hazardous to attempt in a live aircraft (e.g., system problems, poor weather conditions, wind sheer, conflicting traffic), without risk to themselves or the aircraft. Their availability is not dependent on environmental weather conditions, which is often a limiting factor with live flight training. Full-flight simulators may provide advanced instructional training features, such as replay mode for debriefing, and the ability to practice mid-flight maneuvers without full flight progression. This makes them a valuable, cost-effective, and environmentally friendly (i.e., no fuel consumption) training medium that is employed extensively across civil and military airlines. Recognizing the vital roles these devices play in aviation, the International Civil Aviation Organization (ICAO), a specialized United Nations agency representing 192 member civil aviation groups worldwide, developed standards documentation for aircraft simulator quality and evaluation (ICAO, 2009).

Although the benefits of flight simulators for pilot training are well established (Johnson, 2005; Moroney & Moroney, 2010), their high initial build cost and ongoing operating and maintenance costs decrease their accessibility and may also render them prohibitive as an “on demand” training tool. Further, the traditional large hardware systems that make up full-flight simulators make any interface/modular updates resource intensive.

Technological advancements in simulation and VR are making rich virtual environments appealing as an alternative approach to simulator pilot training (Kinciad & Westerlund, 2009). VR environments are three-dimensional, computer-generated, multisensory environments in which the user is immersed (Barfield, Zeltzer, Sheridan, & Slater, 1995). These relatively low-cost alternatives provide an immersive

environment, replicate interface and external conditions, and much unlike full flight simulators, are easily modifiable (i.e., software updates rather than hardware). VR is increasingly being leveraged for training of personnel across many sectors, including in nuclear power plants, manufacturing and industrial applications, military applications, emergency disaster training, and even medical surgeries (Chung, Shewchuk, & Williges, 2002; Duarte, Rebelo, & Wogalter, 2010; Farra, Miller, & Hodgson, 2015; Fuhua, Duffy, & Su, 2002; Grantcharov, et al., 2004; Lehmann, et al., 2005; Nathanael, Mosialos, Vosniakos, & Tsagkas, 2016; Wu, Mu, Yang, & Gu, 2012).

The prospect of using VR for pilot training, in-lieu or in addition to full flight simulators, is a recent topic of interest. Companies such as Bohemia Interactive Simulations (www.bisimulations.com) and Future Visual (www.futurevisual.com) have leveraged VR technology and brought flight training to market. However, little research is available that justifies replacing traditional flight simulation training with VR.

Flight path deviations (FPDs) are a key indicator of flying performance (Causse, Dehais, Arexis, & Pastor, 2011; Leirer, Yesavage, & Morrow, 1989; Van Benthem & Herdman, 2016). Some deviations from the ideal trajectory may be expected, however to justify using VR as a viable alternative to the traditional flight training approach, the FPDs in a full-flight simulator and VR flying environment should be comparable. Flying performance metrics are insufficient on their own to validate this new training approach. With older VR technologies, undesirable secondary effects from VR exposure are a notable concern. Indeed, users reporting feeling sickness symptoms severe enough to discontinue use had been found to account for up to 30% of VR users in the 1990s (Harm, 2002), while as many as 80%-95% of users reported some sickness symptoms (Stanney, Mourant, & Kennedy, 1998; Cobb, Nichols, Ramsey, & Wilson, 1999). Symptoms can persist for up to six hours after exposure (Nichols S. , 1999; Regan & Ramsey, 1994).

While VR systems have improved tremendously since their inception, incidents of nausea, eye strain, and headaches continue to be reported after VR exposure (Jerald, 2016a). In the context of flight training however, there is concern that users may adapt their behaviour in a way that reduces these symptoms, which results in ineffective training of the intended real-world tasks (Kennedy & Fowlkes, 1992). These behavioral adaptations may ultimately compromise the usefulness of the VR-based training. It is important, therefore, to ensure that there are no unforeseen secondary adverse effects that occur in VR flight training (compared to traditional flight simulator training) before incorporating VR technology into training programs. To this end, a holistic evaluation of VR is necessary before it can be considered a viable alternative to a full-flight simulator. This approach would consider not only any differences in performance metrics, but also in the *human factors* components, such as any adverse physiological effects and impact on cognitive load.

The present research examined the user experience, the cognitive load, and the performance metrics in a traditional flight simulator[1] graphics condition against the same flight path and tasks using a VR graphics environment. Both subjective and objective metrics were obtained to ensure a holistic

[1] Due to the constraints of the study, participants sat inside the flight simulator for both conditions, however they were exposed to either a Broad Angle Display Screen (BADS) in the traditional flight environment, or the VR head mounted display in the VR condition. As such, the main differences being considered are between the two visual environments: BADS and VR.

approach in the evaluation. The interaction-centered framework of experience, put forth by Forlizzi and Battarbee (2004), was leveraged as the overarching theory framing this research.

1.2. Research Questions

This research explores whether VR is a viable alternative to a traditional flight simulator graphics. The study design and the results are interpreted in the context of a theory of immersion put forth by Forlizzi and Battarbee (2004) and how these modalities compare across three broad categories:

(1) the User Experience;

(2) the Cognitive Load; and

(3) the Performance.

The User Experience and Cognitive Load components were assessed with both subjective and objective measures. Performance was measured objectively. The explorative hypotheses in this research were that:

(1) There are no differences in the user experience between the two graphics conditions;

(2) There are no differences in the cognitive workload experienced by participants between the two graphics conditions; and

(3) There are no differences in performance between the two graphics conditions.

    a. With the corollary hypothesis that user experience could affect performance

1.3. Thesis Outline

Chapter 2 provides a review of the overarching theory of immersion (Forlizzi & Battarbee, 2004), upon which the present work was based. This is followed by a background discussion about the virtual reality experience, focused on contributing factors (both positive and negative), such as feelings immersion, presence, and side-effects from the hardware and software latency which can lead to motion sickness. These elements ultimately contribute to the user experience, considered both subjectively and objectively. Section 2.2 focuses on the cognitive load that pilots experience in flight training, and ways that this factor can be measured or estimated. Section 2.3 discusses flight performance metrics that would be important to validate VR as an alternative to a traditional flight training simulator.

Chapter 3 describes the methodology of the present user study, applying the concepts discussed in Chapter 2. Specifically, this is achieved by measuring various subjective and objective user experience metrics, cognitive load, and performance outcomes while participants flew a predefined flight path in each of the Broad Angle Display (BADS) and using a VR display. Chapter 4 presents the results of the user study. Chapter 5 revisits the three hypotheses that were posed at the onset of this work, discusses the implications of the results from the user study, and considers insight for using VR as an alternative flight

training approach. This work concludes by discussing its limitations offering recommendations for future related research.

2. Background

2.1. User Experience in Virtual Reality

The experience of the user is a critical issue in an interactive system. This 'user experience' has varying definitions. However for the purpose of this work, the Forlizzi and Battarbee's (2004) framework is used in which the user experience is viewed as an interaction of all the cognitive, emotional, physical and sensual experiences that the user has during their exposure to a product/interaction. Forlizzi and Battarbee suggest a framework for understanding the user experience from a multidisciplinary perspective. This framework guides the design and interpretation of the present research.

2.1.1. An Interaction-Centered Framework of Experience

Interaction focused frameworks consider the subjective experience and observable actions of an interaction (Basri, Noor, Adnan, Saman, & Baharin, 2016). Forlizzi and Battarbee's (2004) interaction centered framework describes both user-product interactions (i.e., fluent, cognitive, and expressive), and dimensions of experience (i.e., experience, an experience, and co-experience). When an individual interacts with a product, his or her experiences flow between fluent, cognitive and expressive interaction as they happen. Figure 2-1 provides a schematic of the dynamics of experience in interaction for individuals in social interactions.

According to this framework (Forlizzi & Battarbee, 2004), *fluent* user-product interactions are ones that do not compete for our attention, but rather are well learned and automatic. They allow us to focus on other things while engaging in them (e.g., consequences of our activities or other matters). Effortlessly driving a car on an empty highway, or riding a bicycle are two examples of *fluent* user-product interactions.

Cognitive user-product interactions are ones that center around the product being used and often result in either learning a skill or a solution, or frustration and errors if the product does not match anything in our past history of product use. Learning a new flying maneuver in a flight simulator, or solving a math problem are two examples of a *cognitive* user-product interaction.

Expressive user-product interactions are ones where the user forms a relationship with the product or some aspect of it. In these types of interactions, users may personalize or modify a product. This investment of effort into it often creates a better fit between the person and the product. For example, customizing a desktop dashboard or restoring a piece of furniture are both examples of *expressive* user-product interactions.

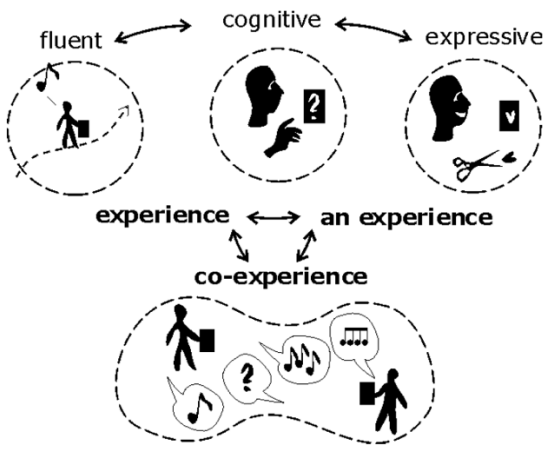


Figure 2-1: Summary of a framework of user experience as it relates to interactive systems (Forlizzi & Battarbee, 2004)

In Forlizzi and Battarbee's framework the three types of user-product interactions (i.e., fluent, cognitive, and expressive) unfold in a particular context, producing three types of experience: (1) experience, (2) an experience, and (3) co-experience.

Experience is the constant stream of "self-talk" that happens while we are conscious and interact with products. It occurs at any given time when we assess our goals relative to the environment, people, and products that surround us. For example, typical experiences are going for a walk or cleaning the house.

An experience is more defined, in that is has a beginning and an end, and often stimulates emotional and behavioural adaptations in the user. It can be a combination of several product interactions that combined, are schematized in one's memory with a sense of completion. For example, *an experience* could be flying a mission in a flight simulator or watching a television show.

Co-experience considers the user experience in a social context. It takes place as experiences are shared with others. Examples of *co-experiences* are playing an interactive video games with friends and collaborating on a group project at school.

Using the Framework

Design teams are encouraged by Forlizzi and Battarbee to use the framework to understand the interactions and experiences that novel systems offer. They suggest conducting research to gain a better understanding of the people, contexts, and activities to generate better solutions to design problems. They indicate that taking an objective perspective to evaluate a user's experience and interaction allows one to assess if a new product will be easily adaptable, learnable, and usable

To understand the *fluent* aspects of experience, they suggest capturing much of the user's interaction in real time without interrupting them. In the present work, measuring physiological indices of cognitive load enables capturing the experience of flying a flight simulator without disruption to the task.

To understand *cognitive* and *expressive* experiences, Forlizzi and Battarbee propose capturing the interactions both as they unfold and any expression after the fact. In the present work, evaluating

performance metrics during the experience and evaluating any negative effects from the immersion environment after the task help to achieve these objectives.

Effects of the VR environment contribute to how individuals perceive their experience in that environment. Some of the most cited include feelings of immersion and presence, side effects from the hardware or software latency, and feelings of unwell such as motion sickness, discussed in turn.

2.1.2. Immersion and Presence

Immersion is a function of the objective VR technology and its potential to engage users in the experience (Slater & Wilbur, 1997). It depends on the extensiveness of the sensory modalities, their congruency (e.g., visuals corresponding to user's head motion), the amount the user is enveloped by the experience (e.g., the field of view), the vividness (e.g., frame rate, resolution), the interactability (e.g., ability for the user to influence events), and the plot that forms (Slater & Wilbur, 1997). However, it is the concept of *presence* that describes how the user subjectively experiences the immersion.

Presence is an internal psychological state of the user, a sense of being somewhere although physically located elsewhere (Jerald, 2016d). The International Society of Presence Research defines it as a "subjective perception in which even though part or all of an individual's current experience is generated by and/or filtered through human-made technology, part or all of the individual's perception fails to accurately acknowledge the role of the technology in the experience" (2000). It is this feeling of presence that is commonly cited as a benefit to using VR for training as it gives users the ability to feel as if they are in this multi-sensory environment, although they are physically located elsewhere.

Presence and immersion are highly intertwined. Presence is limited by immersion, however immersion does not guarantee presence. In other words, if immersion is greater, there is potential for the user to feel greater presence, however the user can break the presence quite simply (Jerald, 2016d). The user can break presence by closing his or her eyes and imagining being elsewhere, by hearing noise from outside the VR environment (e.g., a phone ringing), through a loss of tracking, by unintentionally hitting an object in the real world, or due to system latency (discussed in Section 2.1.4.). Since a user cannot be controlled, maximizing immersion and minimizing breaks in presence are likely the simplest way to increase presence in VR.

While a VR user's sense of presence is most commonly evaluated by questionnaires, comparing a user's behavior in a real environment to that in a virtual environment can also provide information about their sense of presence (Baka, Stavroulia, Magnenat-Thalmann, & Lanitis, 2018). Similarly, so can comparing a user's physiological measures (heart rate and skin conductance) in the two environments (Meehan, Insko, Whitton, & Brooks, 2002).

2.1.3. Hardware

The VR hardware, known as the head mounted display (HMD), that the user wears contributes to the VR user experience. Users can experience physical fatigue and headaches from the weight of the HMD, from the weight distribution, or from pressure at contact points between the HMD and their head.

The weight of HMDs has decreased significantly since the 1990s when they weighed as much as 2kg (Costello, 1997). The Oculus Rift, which we employed for this study, weighs 470 grams (www.oculus.com). However, the location of the center of mass is still an important issue, because if the

HMD is located far from the head center of mass (particularly in the horizontal direction), it can cause strain on the neck from extra torque to offset gravity (Jerald, 2016b). This can lead to headaches and fatigue which negatively impact the user experience.

HMDs are typically designed to be adjustable to accommodate heads of various sizes. Nonetheless, users may experience discomfort at the contact points such as around their eye sockets, on their ears, nose, forehead, and/or back of their head. It can be especially difficulty for users with eyeglasses to fit the headset comfortably, although most HMDs claim to accommodate eyeglasses. Jerald (2016b) discusses options being investigated to reduce HMD discomfort, such as pump-inflatable liners, ergonomic inserts, and different types of foam. Geszten et al., (2015) suggest that technical equipment be light, comfortable, and wireless to improve the user experience in a 3D environment.

2.1.4. Software Latency

Latency refers the delay between a user making an action, and the system responding. In VR, as the user moves his or her head, the latency is the time it takes for the HMD to display the resulting changes (Raaen & Kjellmo, 2015). There is no universally accepted value of latency. However, researchers and developers agree it should be "low", or low enough so that users do not perceive scene motion (Jerald, 2016c).

Raaen and Kjellmo (2015) measured the delay from when the user moves their head until the screen of the Oculus Rift is updated, and found latency averaged between 4 – 14 ms, when vertical synchronization was turned off, or 41 – 63 ms when vertical synchronization was on. Jerald and Whitton (2009) found the system latency Just Noticeable Difference (JND) mean of 16.6 ms, with a minimum of 3.2 ms, suggesting that the Oculus Rift (without vertical synchronization) should have imperceptible latency. However, they also noted that latency thresholds differed significantly for different people, and their sample size was very limited (initially eight, but two were excluded due to noisy data). They suggest that developers design their VR HMDs with latency below that perceived by the most sensitive individuals, if possible.

Latency in VR can hamper the user experience as it can lead to both performance issues, breaks in the feeling of presence, and motion sickness. Latency was shown to result in negative training effects while using desktop displays (Cunningham, Billock, & Tsou, 2001), driving simulators with large screens (Cunningham, Chatziastros, Heyde, & Bulthoff, 2001), and an HMD with latency greater than 120 ms, while tasked to track a target (So & Griffin, 1995). This finding is particularly relevant to our study as the ultimate goal of using VR in lieu of a flight simulator is to improve training for pilots.

Latency can cause breaks in VR presence since the user moves his or her head and the scene does not move in a way consistent with the motion (Meehan, Insko, Whitton, & Brooks, 2002). This confusing scene motion brings the user to realize that they are in fact only in a simulation and not a real environment, thereby breaking their feeling of presence. This latency is also a major contributor to feelings of motion sickness in VR.

Motion sickness can be caused by latency since visual cues become out of phase with vestibular cues, creating a sensory conflict. This is discussed below in Section 2.1.5.

2.1.5. Motion Sickness

One of the greatest challenges of VR are reports of feeling unwell after exposure. Symptoms can include nausea, dizziness, disorientation, confusion, drowsiness, while observable signs include cold sweating and emesis (Lackner, 2014). Exposure to VR can result in a *visual-vestibular conflict*, which is a conflict between the patterns of sensations connected with real movement and the stimuli received from the images that are seen during immersion in VR, which has been shown to provoke feelings of motion sickness (Akiduki, et al., 2003).

For example, since the introduction of a $100 million VR ride "Mission: Space" at Walt Disney World in Florida, the park noticed the most hospital visits for a single ride (in an 8-month period): six visits for nausea and chest pain (Johnson, 2005). As a result, the "Mission: Space" ride now has sickness bags in the ride. We can speculate that if six individuals required hospitalization, many others experienced uncomfortable symptoms to a less severe extent. The undesirable side-effects of this ride degrade its user experience. Understanding these side-effects in VR, and ultimately how to eliminate them will contribute to an improved user experience and user acceptance of this technology.

2.1.5.1. Definitions

There are many terms used to refer to the unwell feelings that a user may experience from exposure to VR, including cyber sickness, motion sickness, and simulator sickness. According to Jerald (2016a, p. 160), cybersickness refers specifically to "*visually induced motion sickness resulting from immersion in a computer-generated virtual world*". Motion sickness is broader and according to Lawson (2014, p. 531), refers to "*adverse symptoms and readily observable signs that are associated with exposure to real (physical or visual) and/or apparent motion*. Simulator sickness has been defined as "*sickness that results from shortcomings of the simulation, but not from the actual situation that is being simulated*" (Pausch, Crea, & Conway, 1992, p. 344). In other words, simulator sickness implies that a user would feel sick during a simulated flight, but if they were to do the same flight in a real aircraft, they would not have the same feelings of sickness (however, they could still feel sick for other reasons associated with flying that were not present in the simulator version). For the purpose of this research, we will use the term cyber sickness to refer to any feelings of unwell after exposure to VR.

There is merit differentiating between simulator sickness and cyber sickness, although they are often used interchangeably. Simulator sickness results from small discrepancies between a user's normal, expected motion and the actual simulator motion, while cyber sickness is experienced in VR when users appear to be moving in the virtual scene while actually remaining stationary (Davis, Nesbitt, & Nalivaiko, 2014) (Stanney, Kennedy, & Drexler, 1997). The severity of cyber sickness has been reported to be three times that of simulator sickness, with oculomotor type symptoms often reported from simulator sickness, while disorientation is most common of cyber sickness (Stanney, Kennedy, & Drexler, 1997).

Cobb et al. (1999) coined the term virtual reality induced symptoms and effects (VRISE) after conducting a series of nine experiments with a total participant sample of 148 individuals, examining the effects they experienced during and after using various HMDs. They discovered that although the effects where similar to that for simulators or transport systems, the causes were sufficiently different that it justified the new term (i.e., VRISE).

2.1.5.2. Theories of Cybersickness

The exact cause of cybersickness and the underlying physiological responses are uncertain, however there are three prominent theories for its cause: the sensory-conflict theory, postural instability theory, and poison theory (Davis, Nesbitt, & Nalivaiko, 2014).

The *sensory conflict theory* is a popular and longstanding theory on motion sickness (Reason & Brand, 1975) that can be applied to cybersickness (Cobb, Nichols, Ramsey, & Wilson, 1999). This theory describes the conflicts between the visual and vestibular senses while engaged in VR. A sensory mismatch can occur when the vestibular system is telling the individual that they are remaining stationary, however the visual system senses that the body is moving due to the VR environment (Keshavarz, Hecht, & Lawson, 2014).

For example, consider a VR flight simulator. As the subject is flying, the optical flow patterns of the buildings, roads, clouds and trees move past the user's periphery providing a sense of vection. The visual system tells the user they are moving at a certain speed, in a certain direction, and accelerating at a certain rate depending on how hard they push on the throttle. In past real-world flying experience, the vestibular and visual senses both provided matching information, however in VR, since the user is not actually moving, there is a conflict between the two senses and they do not align with the experience the user expects. Cybersickness may occur.

The *postural instability theory* focuses around the idea that one of the primary goals of humans is to maintain postural stability in the environment (Riccio & Stoffregen, 1991). It suggests that cybersickness results from prolonged postural instability which can result from optically specified acceleration or rotations that are unrelated to the postural control strategies learned in the real-world environment. For example, it may be intuitive for a VR user to use muscular force to resist an angular acceleration that is visually induced because of his or her experience in the real world. However, since there is no physical tilt, the muscular contractions to maintain postural stability actually created an unintended divergence from a stable position, resulting in postural instability. This postural instability causes cybersickness.

The *poison theory* uses an evolutionary standpoint to provide an explanation for why cybersickness occurs (Treisman, 1977). This theory claims that cybersickness is based on a maladaptive process which originally helped the body get rid of toxic substances. Accordingly, the stimulation from VR can affect the vestibular and visual systems such that the body presumes it has ingested some type of poison, thus causing symptoms to rid the body of the poison (e.g., nausea, or an emetic response). However, the poison theory does not address the broader symptoms that can arise in VR (Davis, Nesbitt, & Nalivaiko, 2014).

All three theories have some valid arguments for the cause of cybersickness, yet one can find an example of a situation that could possibly refute each theory. However, none of the three theories explain why one individual experiences cybersickness while another does not in identical conditions.

2.1.5.3. Contributing Factors

There are a multitude of factors that contribute to feeling unwell after being exposed to VR. Nichols, Cobb, and Wilson (1997) proposed a framework of major factors that influence when VR symptoms arise, with four main factor groups: VR technical system, Virtual Environment design (i.e., the content of

the virtual world), circumstances of use, and individual participant characteristics. Similarly, Jerald (2016e) summarized the factors that contribute to cyber sickness and grouped them into system factors, individual user factors, and application design factors. Appendix A presents the full list, while the paragraphs below summarize some of the major contributors.

System Factors - Latency (described in Section 2.1.4) is the single system factor that has the greatest impact on cyber sickness. According to Holloway (1997) it has a greater effect than the remaining system factors combined. Accordingly, minimizing latency would be a critical factor for developing a VR flight training system with a favorable user experience.

Individual User Factors - Cyber sickness can potentially affect any individual with an intact and functioning vestibular system, however the range of vulnerability varies about 10 000 to 1 in the general population (Lackner, 2014). The individual variability is vast, with some users exhibiting no signs unless in the most extreme conditions, while others exhibit signs after very minimal VR exposure. Individual factors that affect one's susceptibility include gender, age, history of motion sickness, past experience with VR, and thinking about sickness.

Stanney, Kennedy, and Hale (2014) report that incidences of cyber sickness are three times more prevalent in females than in males, possibly due to hormonal differences, field of view differences (larger in females than in males, which is associated with increased cyber sickness), and biased self-report data (i.e., males may underreport cyber sickness symptoms).

Cyber sickness tends to increase with age (Brooks, et al., 2010) which is the reverse of what data has shown for motion sickness from physically induced environments (Paillard, et al., 2013; Reason & Brand, 1975). In other words, susceptibility to motion sickness decreases as one ages, while susceptibility to cyber sickness increases with age. While the exact reason for the difference is not known, it is speculated that factors such as experience, lack of the ability to accommodate, and balance could contribute to this difference.

The best predictor of one's susceptibility to cyber sickness is their history of motion sickness in another environment. Generally, the correlation between individual susceptibility to feelings of motion sickness in one environment and another is between 0.6 to 0.8 (Golding, 2006). Past experience with VR also contributes to one's individual susceptibility, with novice users almost always experiencing more symptoms than experienced ones (Lackner, 2014). This is likely due to adaptation that occurs with exposure (Hill & Howarth, 2000; Stanney K. M., Kennedy, Drexler, & Harm, 1999). Conversely, individuals with more experience in the real-world task, such as pilots (i.e., with more flight hours) are more susceptible to cyber sickness than less experienced pilots (Johnson, 2005). This likely occurs due to their higher sensitivity to how the real world should behave for the VR task being simulated.

While ethically it may seem like the best course of action, suggesting to individuals prior to exposure that they may get sick while in VR can in fact lead to increased symptoms than if they were not informed (Young, Adelstein, & Ellis, 2007). This finding presents some dilemma whether or not to conduct baseline sickness screening questionnaires as they may predispose participants to cyber sickness although this comes at the cost of not having baseline data. It is generally recommended not to administer the Kennedy Simulator Sickness Questionnaire (SSQ) (Kennedy, Lane, Berbaum, & Lilienthal, 1993), the standard VR sickness screening tool, for baseline measures. However, some researchers

suggest that casually mentioning possible side-effects, but not dwelling on them, allows baseline data collection with minimal to no bias (Jerald, 2016e).

Application Design Factors – The main application, or task factors, that affect VR sickness are the locus of control and the task duration (Davis, Nesbitt, & Nalivaiko, 2014). When a user has more control of the virtual environment, he or she can better predict future motion and experiences less cyber sickness symptoms. Several studies have shown that passive observers (e.g., the passenger in a car) experience a higher rate of cybersickness symptoms and greater severity of symptoms than the active participant (e.g., the driver or the pilot) (Sharples, Cobb, Moody, & Wilson, 2008; Stanney & Hash, 1998). The duration of VR exposure affects cyber sickness, where users are more susceptible to cyber sickness in tasks of longer duration (Stanney, Kennedy, & Drexler, 1997). Brief exposures to VR have therefore been suggested as a way to improve the speed of adaptation (Davis, Nesbitt, & Nalivaiko, 2014).

2.1.5.4. Measuring Cyber Sickness

Questionnaires, such as the SSQ are most commonly used to measure cyber sickness. Kennedy et al. (1993) developed the SSQ from 1,119 pairs of pre and post flight simulator exposure data reported by United States Navy personnel, collected using a traditional Pensacola Motion Sickness Survey (Kellogg, Kennedy, & Graybiel, 1965). Through a series of factor analysis, Kennedy et al. (1993) identified twenty-seven commonly experienced symptoms, then eliminated ones with rare occurrence or ambiguity (e.g., boredom), resulting in a sixteen-symptom questionnaire known now as the SSQ (Davis, Nesbitt, & Nalivaiko, 2014). Participants rate each of the sixteen symptoms with either none, slight, moderate, or severe, and through some calculations, four representative scores are determined: (1) a Nausea-related subscore, (2) an Oculomotor-related subscore, (3) a Disorientation-related subscore, and (4) a total score. While it was initially intended for measuring simulator sickness, the SSQ is commonly applied for both simulator and cyber sickness measures.

Measuring physiological characteristics during VR exposure, provides an additional tool, that is much less frequently explored in VR research. The benefits of these measures are that they can provide uninterrupted, real-time data while the user is in the VR environment, unlike questionnaires that are administered at a specific point in time and require a pause from the study. Cowings et al. (1986) measured heart rate, respiration rate, finger pulse volume, and basal skin resistance on 127 subjects before, during, and after exposure to a nauseogenic rotating chair test. They observed significant changes in all autonomic responses across the tests, suggesting that these objective measures can be used as indicator of motion sickness. However, there may be inherent differences of physiological symptoms between motion sickness and cyber sickness.

Kim et al. (2005) exposed 61 participants to a virtual environment navigation task for about 10 minutes, while they measured their heart rate, blink rate, skin conductance level, and electroencephalogram (EEG) waves. They also administered several questionnaires for sickness susceptibility and immersive tendency prior to the study, followed by a simulator sickness questionnaire and a presence questionnaire after the study. They found that a significant physiological change occurred when participants experienced cyber sickness, measurable by heart rate (specifically, the heart period), eyeblink rate, upset stomach, and respiration rate.

Similarly, Ohyama, et al. (2007), examined heart rate variability and the development of subjective symptoms during motion sickness induced by VR of 10 young healthy volunteers. They found an

increase in sympathetic nervous activity, but no change in parasympathetic nervous activity during incidents of cyber sickness. They also noted no correlation of individual subjective symptoms with the individual results of power spectrum analysis.

Deducing that certain physiological changes (e.g., increased heart rate) during a VR task are solely a result of increased cyber sickness symptoms is challenging, as these measures also vary with the stress level the user is experiencing as a result of the task demands. Heart rate variability may provide some insight into cyber sickness, however for the purpose of this research, we leverage subjective questionnaires for measuring cyber and simulator sickness and focus on exploring heart rate and skin conductance as a measure of stress, or mental workload. These objective measures are another useful gauge for evaluating the user experience in VR compared to traditional training approaches. This topic is discussed successively in Section 2.2.

2.2. Cognitive Load

Cognitive load, or colloquially known as mental workload, refers to the effort being used in the working memory to solve a problem or complete a task (Sweller, 1988). Working memory resources are limited by two factors: its capacity and duration (Baddeley, 1992), meaning that it can only handle a limited number of items (arguably 4-7), known as "chunks" for a limited amount of time (Cowan, 2001). Cognitive load varies among individuals, even for the same task. It is due to an interaction that emerges from the circumstances under which the task is being performed; the skills, behaviours, and perceptions of the individual; and the requirements of the task (Hart & Staveland, 1988).

The Cognitive Load Theory suggests that cognitive load consists of at least two sources of load: intrinsic and extraneous[2] (Sweller, van Merrienboer, & Paas, 1998). Intrinsic cognitive load refers to the inherent task difficulty, whereas extraneous load refers to the complexity resulting from the manner the task material is presented. While intrinsic load cannot be altered by instructional interventions (as it is intrinsic to the material itself), extraneous cognitive load can. Situations that evoke excessive cognitive load can severely impede performance and learning because working memory may be substantially exceeded (Paas, Tuovinen, Tabbers, & Gerven, 2003). As such, instructional design needs to minimize the amount of extraneous cognitive load imposed, to ensure total cognitive load is within working memory limits (Sweller, van Merrienboer, & Paas, 1998).

For the purpose of this research, we consider the intervention of either VR or the traditional graphics display, to be the extraneous cognitive load, since the flying task itself (the intrinsic cognitive load) remains the same for both interventions. Accordingly, comparing cognitive load in these two environments is a necessary step in validating VR as a viable pilot training alternative to the traditional graphics condition. We discuss the impact of cognitive load on performance in Section 2.2.1 followed by methods used to estimate the cognitive load Section

2.2.1. Performance and Cognitive Load

The relationship between performance and cognitive load, known as the Yerkes-Dodson Law, was initially proposed by psychologists Robert Yerkes and John Dodson and has since been verified in many

[2] Extraneous cognitive load can be distinguished from germane cognitive load, however that level of detail is beyond the scope of the present paper.

experiments (Ahmadi & Alireza, 2007; Muse, Harris, & Field, 2003; Yerkes & Dodson, 1908). It explains that performance has an inverted-U shape when mapped against cognitive load. Performance increases with cognitive load up to an optimal point, after which it decreases. Complex situations in a flight environment can undoubtedly cause a pilot to experience high cognitive load however, it should not be overlooked that monotonous situations (e.g., long, trans-Atlantic flights in Autopilot) can also lead to a vigilance decrement, as performing the task requires a high effort to remain awake and alert (Paxion, Galy, & Berthelon, 2014).

Performance on a task can remain constant even with increasing workload conditions, until a point is reached where the performance level drops drastically (Skinner & Simpson, 2002). Task shedding of lower priority tasks may suggest that the pilot is approaching an overloaded condition, but often times this is only clear once failure occurs. In modern aircraft, a pilot's role is increasingly that of monitoring, information management, and decision making (Skinner & Simpson, 2002) while the manual tasks which may have been good indicators of workload (based on primary task performance), are now mostly automated. Monitoring performance alone may not be timely and sensitive enough to determine cognitive overload since one pilot may have remaining attentional resources (to respond to an unexpected event), while another does not yet they are both performing equally (Yeh & Wickens, 1988). This suggests that monitoring a pilot's cognitive load, in addition to performance, may have critical safety implications.

2.2.2. Measuring Cognitive Load

Literature on cognitive load discusses four approaches for estimating it: subjective ratings, task-performance based measures, physiological measures, and behavioural. For the purpose of this research, we discuss the first three types, including their advantages and disadvantages.

2.2.2.1. Subjective Measures

Subjective measures ask a user to rate the mental demand placed on them by using a pre-defined rating scale. Research has shown that people can assign numerical values to their imposed cognitive load or invested cognitive effort, as they are good at introspecting on their cognitive processes (Gopher & Braune, 1984). They are popular due to their ease of use, non-intrusiveness, low cost, high face validity, and known sensitivity to workload variations (Reid & Nygren, 1988).

There exist various standardized questionnaires for assessing cognitive load. In the context of piloting aircrafts, commonly used are the Cooper-Harper Scale (Cooper & Harper, 1969), the Subjective Workload Assessment Technique (SWAT) (Reid & Nygren, 1988), NASA Task Load Index (TLX) (Hart & Staveland, 1988), and the Instantaneous Self-Assessment (ISA) (Jordan & Brennen, 1992).

The Cooper-Harper Scale primarily assess a pilot's cognitive ability to handle an aircraft. The term "handling qualities" includes stability and control characteristics, but also the influence of the cockpit interface, the aircraft environment and stress (Cooper & Harper, 1969). After certain flight tasks, pilots assess the demands placed on them by the task using a unidimensional scale, ranging from 1 (*Excellent, highly desirable, pilot compensation not a factor for desired performance*) to 10 (*Major deficiencies, control will be lost during some portion of the required operation*). The Cooper-Harper scale has been criticized for lacking diagnostic power due to its unidimensional nature (Payne & Harris, 2000).

The SWAT consists of three component factors: Time Load, Mental Effort Load, and Psychological Stress Load to capture the multidimensional nature of mental workload (Reid & Nygren, 1988). Each dimension has three levels: low, medium and high. The SWAT requires users to conduct a card sorting pre-task procedure where they rank 27 SWAT cards, followed by a task scoring procedure. Although the SWAT appears to be more suitable than other mental workload techniques in terms of content validity and diagnostic capabilities, it is criticized as having low sensitivity for low mental workload, and it is time consuming due to the card-sorting pre-task and subsequent scoring (Luximon & Goonetilleke, 2001).

The NASA-TLX is another multidimensional assessment tool consisting of six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration (Hart & Staveland, 1988). For each task, users rate each of the six dimensions on a 100 point scale (in 5-point intervals). The second part of the TLX requires the users weigh the six subscales by doing a pairwise comparison based on their perceived importance, although some researchers skip this part and simply refer to the 'raw TLX' scores (Hart S. G., 2006). A workload score is obtained for each rated task by multiplying the weight by the individual dimension scale score, summing across scales, and dividing by 15 (the total number of paired comparisons). A possible draw-back to the NASA-TLX, similar to the SWAT, is the time intensiveness of preparing participants for it, administering it, and finally scoring it.

The ISA is a very simple workload rating scale, developed to provide immediate subjective workload ratings during performance of the primary task, and commonly used in air traffic control (Jordan & Brennen, 1992). Participants rate their perceived mental workload on a simple pre-defined Likert scale at pre-defined intervals during the task. Responses can either be oral, written manually, or entered into a pre-defined numerical keypad, typically on a scale of 1 (low) to 5 (high). It has the advantage that it is simple and easy to use, without any prior training. However, performance on the primary task has been found to decrease during the periods when ISA responses are required, regardless if they were manual or spoken (Tattersall & Foord, 1996).

Overall, subjective measures have common disadvantages. These questionnaires either have to interrupt task flow (and potentially increase cognitive load of already overloaded users) or be administered after the intervention (Chen, et al., 2012). If administered after, they rely on a user to rate their experience retrospectively. Users may forget important aspects depending on the time between the event and the time the questionnaire is administered. The events occurring closest to the time the questionnaire is administered can greatly impact the rating (Wilhelm & Grossman, 2010). Moreover, people tend to think about 'workload' differently; their impressions of workload can be influenced by the quality of their own performance and calibration exercises may be necessary to train participants (Casner & Gore, 2010).

2.2.2.2. Task-Performance-Based Measures

The task-performance-based measures are grounded on the assumption that any increase in task difficulty will lead to an increase in demands, which will decrease performance (Martins, 2016). However, due to speed-accuracy trade-offs, fatigue after-effects, and strategy adjustments, it is possible for performance to remain unaffected by increased task difficulty or cognitive load. For this reason, two performance measures are considered: the primary task measurement (e.g., flying the aircraft) which provides a direct indication of performance on the task of interest (see Section 2.3), and the secondary task, which provides a useful index of spare cognitive capacity (Paas, Tuovinen, Tabbers, & Gerven, 2003). Performance on the secondary task is influenced by the cognitive load of the primary task.

The subsidiary-task paradigm is when subjects are instructed to attend to the primary task and degradation is measured on the secondary task (Martins, 2016). For example, a primary task is flying an aircraft, and a secondary task is to respond to a visual or auditory stimulus. Subjects are presented with a sensory stimulus every few seconds and are asked to respond to it by pressing a button attached to their finger. Missed signals and reaction time to the stimuli are calculated to determine the cognitive load of the primary task, since as workload increases in the primary task, subjects will have less resources to attend to the secondary task, thereby the error rate increases and reaction time decreases.

This secondary task test is commonly referred to as a 'Peripheral Detection Task' (PDT), or as a 'Detection Response Task' (DRT). They are widely used in driving simulator studies to measure cognitive load (Bruyas & Dumont, 2013; Conti, Dlugosch, & Bengler, 2014; Martens & van Winsum, 2000) and to compare cognitive load between driving simulators and real driving (Riener, 2010). Riener (2010) found reaction time to be 13% better in the simulator than in real driving, likely due to the real risk associated with the real-world environment.

Although PDT tasks have been used successfully in aviation research to assess workload in the past (Kantowitz, Hart, & Bortolussi, 1983; Wickens, Hyman, Dellinger, Taylor, & Meador, 1986), they have been criticized by other researchers that they are foreign to the aviation operational environment (Lysaght, et al., 1989). They recommend using tasks that are usually performed during normal operations, but that can be separated from the primary task, such as aircraft radio communication.

A main disadvantage with PDT tasks are that ceiling effects have been shown with very high cognitive load. Also the method itself may affect performance on the primary task (Stojmenova & Sodnik, 2018). Nevertheless, the PDT is an easy to implement method that gives valuable estimates of the cognitive load of subjects performing a continuous task.

2.2.2.3. Physiological Measures

Physiological workload measurement techniques are based on the fact that the autonomic nervous system unconsciously regulates the bodily functions as workload changes (Mandrick, Peysakhovich, Rémy, & Lepron, 2016). They assume that changes in cognitive load are reflected in changes in human physiology, such as heart rate, skin conductance, brain activity, and eye activity, among others (Kramer, 1991). One of the biggest advantages of these measures are that they can be collected in real-time, however many require specialized equipment and trained technicians (de Waard, 1996).

Due to the obtrusiveness of some equipment required for physiological measurements (e.g., EEG headsets), some measures may not be best suited for a flight environment[3]. Heart rate and skin conductance (also known as galvanic skin response (GSR) or electrodermal activity) are nonobtrusive and non-invasive and thereby do not compromise flight safety. Thus these measures were used in the present research study.

Heart Rate – Heart rate is perhaps the most widely studied physiological measure to estimate cognitive load in pilots, dating back to 1917 when Gemelli (1917) used an electrocardiogram (ECG) to measure blood pressure, breathing rate, and pulse rate in experienced pilots during flight (Roscoe, 1992). He found "clear

[3] Research with EEG, eye tracking, and other obtrusive measures in flight is common, although for this study it was seen as a prohibitive factor.

signs" between increased flying stress and increased heart rate. ECG research in live flight steadily continued for many decades, and researchers began to examine heart rate as not just a result of physical and metabolic demands of flight, but also as a result of mental or psychological stress (Howitt, 1969; Lewis, 1967; Lidderdale, 1987; Roman, Older, & Jones, 1967; S'Jongere, Bertels, & Ego, 1977; Sekiguchi, et al., 1977; White, 1940). The studies mostly agreed that the landing phase caused the largest increase in heart rate for the pilot, followed by the takeoff phase. The research also eventually showed that the theoretically more challenging aspects of flight were correlated with higher heart rates (e.g., reduced visibility landing conditions, severe turbulence, or cross-winds) (Debijadji, Perovic, Nagulic, & Djurakic, 1973), and that heart rate was higher for the pilot flying than for the pilot not flying (Hart & Hauser, Inflight application of three pilot workload measurement techniques, 1988).

In the early 1970s, studies measuring pilots' heart rate in flight simulators began, which permitted increased task difficulty scenarios (e.g., low visibility, mechanical failures, adverse weather, etc.) that were previously not possible in real aircraft due to safety concerns (Roscoe & Goodman, An investigation of heart rate changes during a flight simulator approach and landing task, 1973). Roscoe (1992) found that experienced pilots' heart rates are typically lower during a simulated flight than during the same task in a real flight. Pilots without much experience in a particular flight simulator however, respond with heart rate increases with changing task demands. Likewise, the more sophisticated and realistic the simulator, the more likely a pilot is to have a significant increase in heart rate, similar to that expected in a real aircraft. Roscoe and Grieve (1988) recorded similar heart rates and workload ratings from pilots flying in the B767 airplane and flight simulator.

A study on cognitive load and heart rate in a Boeing 747-400 flight simulator showed that male pilots' peak heart rate occurred at landing, followed closely by take-off, and that these were directly related to their NASA-TLX scores (Lee & Liu, 2003). Allsop et al. (2016) measured heart rate in non-pilots to be directly related to self-reported state-anxiety (due to a high cognitive load condition) in a flight simulator and found that cognitive load negatively impacted flight performance.

Heart rate measurements taken during VR flight to estimate cognitive load in VR or compare heart rate in VR to a simulated or live flight, to our knowledge, have not yet been published in the research literature. Heart rate has shown promising results in VR environments in other contexts, such as detecting stress levels during arithmetic tasks (Cho, et al., 2017) and estimating cognitive load in VR beam-balancing tasks with various heights (Peterson, Furuichi, & Ferris, 2018).

The popularity of measuring heart rate is unsurprising as it does not compromise flight safety and is readily accepted by pilots. Electrocardiogram (ECG) remained the most common method to measure heart rate in the flight domain, however the last decade has shown interest in less obtrusive, less expensive, and less complex tools (Wang & Fu, 2016). Chest straps are a common alternative, which closely emulate a real ECG machine by measuring electrical pulse and sending that information to a wrist watch (Terbizan, Dolezal, & Albano, 2002). The recent ubiquity of wearable biosensors, such as smart watches and wristbands, enables the collection of heart rate data via a photoplethysmography (PPG) sensor which consists of LEDs (located at the back of the watch) that shine light onto the skin to detect changes in blood volume. The scattered light is sensed with a photodetector and run through an algorithm to compute heart rate.

Galvanic Skin Response – Galvanic Skin Response (GSR), also known as skin conductance response (SCR), or electrodermal activity (EDA), is a sensitive measure of changes in sympathetic arousal, which

can be measured to assess cognitive load (Critchley, 2002). It is based on the fact that the resistance of the skin will change with the degree of sweat production, which is innervated by the sympathetic system (Lysaght, et al., 1989). It is measured by applying a weak current through the skin and measuring the resistance. An individual's GSR is directly related to their stress level so as stress increases, so does their GSR (Shi, Ruiz, Taib, Choi, & Chen, 2007; Brunken, Plass, & Leutner, 2003). GSR consists of a tonic (slow) and phasic (fast) component, where increased activity in the phasic component is related to higher cognitive load (Reinhardt, Schmahl, Wust, & Bohus, 2012).

The earliest report of measuring GSR in a flight environment to assess cognitive workload was in 2002, when Wilson tested ten pilots in a 90-minute visual flight rules (VFR) and instrument flight rules (IFR) real-world flight (Wilson, 2002). On two separate occasions, each pilot repeated the same flight to test the reliability of the physiological measures. No statistically significant differences were found in GSR response between the two flights. Within each flight, GSR activity was greatest during takeoff and landing, indicating the high cognitive workload involved in these flight segments, however subjective workload ratings did not indicate the same results. Subjective workload ratings suggested that pilots found the two IFR tracking segments as the most difficult. This may be due to the unfamiliarity of those segments, versus takeoff and landing which are familiar, even though cognitive workload may be high. This same study found a high correlation between GSR and heart rate, with heart rate more sensitive to varying flight demands, indicating that perhaps it is sufficient to measure only heart rate.

Estimating cognitive load through GSR measurements in flight simulators has also shown promising results. Increasing the strength of turbulence in a flight simulator task (meant to increase task difficulty and cognitive load), showed corresponding increases in GSR frequency and heart rate for four student subjects (Boucsein, Haarmann, & Schaefer, 2008). Additionally, decreased GSR frequency and reduced sum of GSR amplitudes was shown in the progression of the flight missions (four total) indicating habituation. GSR frequency was then used as an indicator of cognitive workload in pilots for an adaptive automation task during simulated flight (Haarmann, Boucsein, & Schaefer, 2009). Haarmann et al. ran two experiments (n = 18 and n = 48) and found that combining GSR with heart rate variability was superior than GSR alone to adjust cognitive workload to a set-point through adaptive automation.

Other research measuring GSR and cognitive workload (through a modified ISA questionnaire) with six helicopter pilots in a full-flight simulator suggests that pilots show a direct relationship between the two measures, however the level of pilot expertise impacts the results (Gaetan, et al., 2015). They concluded that it is not possible to establish an accurate prediction of cognitive workload without considering individual specificities and cross analysing them with multiple physiological measures. The interpretation of GSR data can be rather ambiguous, even for an expert in the field. Often, additional information is required to determine if results are solely due to increased arousal, or if they are due to other factors (e.g., rise in temperature, physical exercise) (Bakker, Pechenizkiy, & Sidorova, 2011).

GSR measurements taken during VR flight to estimate cognitive load in VR or compare GSR in VR to a simulated or live flight, to our knowledge, have not yet been published in the research literature. GSR has however shown promising results in VR environments in other contexts. GSR was measured in VR exposure therapy (Wout, Spofford, Unger, Sevin, & Shea, 2017), during stress recovery and distraction from work duties using relaxing VR simulations (Ahmaniemi, Lindholm, Muller, & Taipalus, 2017), during a VR commercial cabin flight simulation, as a passenger in the window seat, and compared to a VR first-person driving task (Jang, et al., 2002), and as a measure of cognitive load comparing three different

training mediums (VR, technical manuals, and multimedia films) for complex tasks (Chao, Wu, Yau, Feng, & Tseng, 2017). Measuring GSR in a VR aviation environment would contribute to testing VR's viability as an alternative flight training medium to traditional flight simulator graphics.

2.3. Performance Metrics

Flight simulators have a long history of research showing that, in combination with real-world aircraft training, they improve flying performance more than aircraft training alone (Lintern, Roscoe, Koonce, & Segal, 1990; Taylor, Talleur, Emanuel, & Rantanen, 2005; Hayes, Jacobs, Prince, & Salas, 1992). Measuring flying performance in VR in order to validate it as a flight training graphics alternative is important to ensure the VR environment does not put the trainee at a disadvantage compared to training in a traditional flight simulator. Moreover, it is important that it does not put the trainee at a performance advantage (i.e., where flying is less difficult than in the real-world task) which would also be detrimental to training for live flight. Indeed, comparing performance metrics between the two conditions will help assess VR's viability as a flight training alternative to flight simulators.

Flight performance metrics, such as heading, altitude, and speed deviations from prescribed values are often used as a standard evaluation criterion for flying performance. The Federal Aviation Administration (FAA) flight test standards specify appropriate windows for performance during maneuvers of +/- 100 feet altitude deviations, and +/- 10 degree heading deviations (Federal Aviation Administration, 2002). Root mean square error (RMSE) is a commonly used objective measure of pilot performance (McClernon & Miller, 2011). It is the square root of the sum of mean error squared and standard deviation squared. In other words, it is a combined measure where the mean error and standard deviation are equally weighted (Hubbard, 1987).

Lysaght et al. (1989) differentiate between Type 1 and Type 2 performance measures. Type 1 measures of primary task performance are indices of both the operator's performance and the system. For example, changes in thrust would be a result of the operator activity and any associated system lag. Type 2 measures of primary task performance assess only the direct performance of the operator (Hart S. G., 1986). For example, the frequency of control inputs on the wheel, column, or throttle while flying in turbulence would be a Type 2 measure, while the actual glide slope error is a Type 1 measure. Type 1 measures typically provide an index of system performance and would be useful for the task of comparing two systems (e.g., in the present research, VR to BADS) while Type 2 are more sensitive measures of operator workload.

2.4. Present Research

The present research examined the user experience, the cognitive load, and the performance metrics in a traditional flight simulator graphics condition against the same flight path and tasks using a VR graphics environment. The objective of this work is to serve as a first attempt at understanding whether VR is a viable alternative to traditional flight simulator training.

The interaction-centered framework of experience, put forth by Forlizzi and Battarbee (2004), was leveraged in the study design as it is highly applicable to the aviation environment, and enabled the research to ensure that all three types of user-interactions were measured within the chosen experimental flight measures.

Specifically, to understand *expressive* interactions, any negative effects of the immersion environment were captured by the user experience measures and a subjective cognitive load rating. VR usability measures were gathered to provide more information on any adverse effects of this simulation environment. To understand the *fluent* interaction of flying, the study captured continuous physiological measures (heart rate and GSR) without disrupting the task, and measured the cognitive load through a secondary task. Lastly, the *cognitive* user interaction was captured by the objectively measured primary performance metrics, such as flight path and airspeed deviations.

To understand the viability of VR as a flight training alternative to traditional simulator training, the two flying experiences (VR and non-VR) were compared against the battery of metrics described above.

3. User Study Methodology

The study was conducted at the Advanced Cognitive Engineering (ACE) laboratory at Carleton University. Participants' user experience (i.e., via questionnaires probing motion sickness and other VR usability metrics), cognitive workload (i.e., heart rate, GSR, PDT, and subjective rating) and performance (i.e., deviations from optimal flight path and airspeed, and flying precision) where obtained while flying in a Cessna 172 flight simulator. Two different graphics conditions were compared: VR and a broad-angle display screen (BADS).

3.1. Participants

Forty-one participants (twenty-eight males) were recruited from the undergraduate (*n=34*) and graduate (*n=7*) population at Carleton University. Participants ranged in age from 18 to 32 years of age with a mean age of 21.32 (*SD = 3.72*) years. Two of the participants were licensed pilots, an additional one had experience flying an aircraft, while the remaining 38 participants had no prior flying experience. Eighteen participants indicated that they play video games at least once a week, while the remaining participants played rarely or never. All participants gave their informed consent to participate in the study. Participants were reimbursed for their participation with either refreshments or Psychology course-credit, for graduate and undergraduate students respectively.

While only 2.5 % of participants (*n=1*) identified as regular users of virtual reality products (e.g., the Oculus Rift), 45% (*n=18*) reported that they play video games at least once a week (see Figure 3-1). Most participants (90%, *n=36*) said they are not prone to motion sickness, however 12.5% (*n=5*) and 7.5% *(n=3)* reported sometimes feeling a little queasy while playing video games or virtual reality products, respectively (see Figure 3-2).

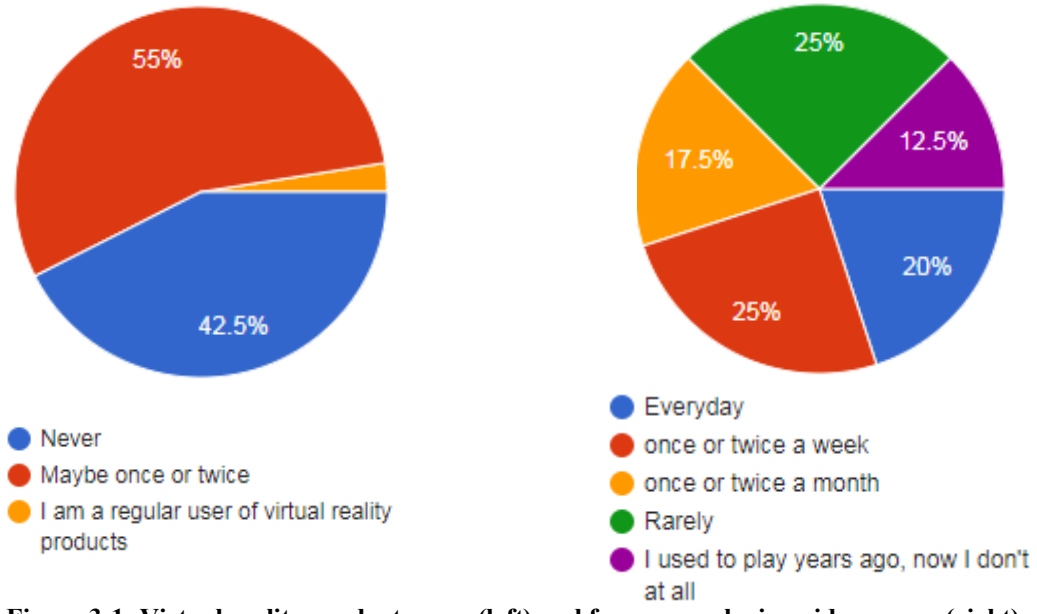


Figure 3-1: Virtual reality product usage (left) and frequency playing video games (right)

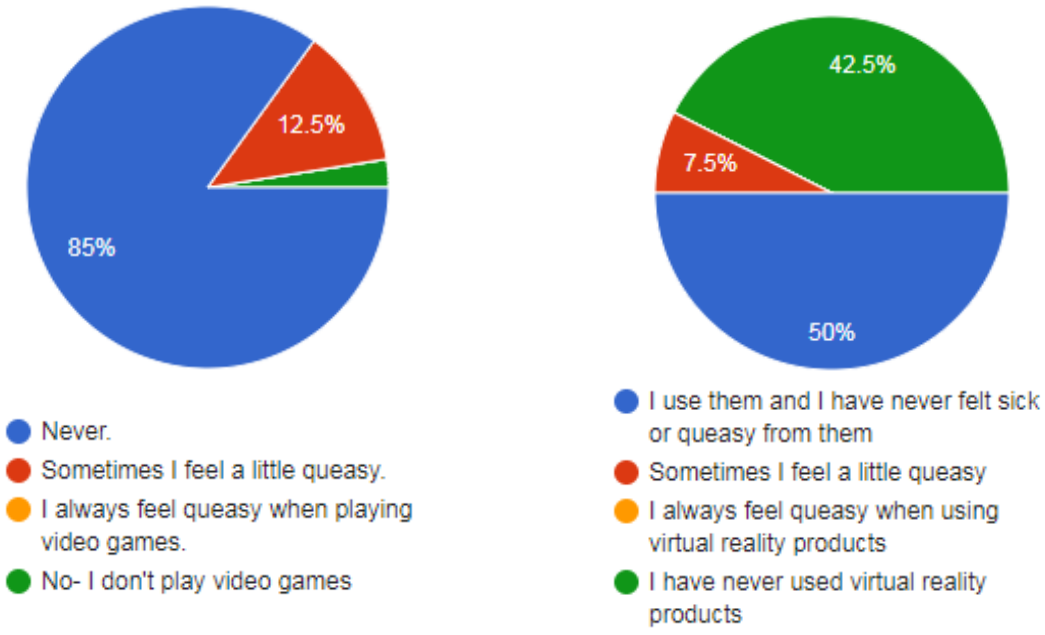


Figure 3-2: Queasiness from playing video games (left) or with VR products (right)

3.2. Technical Set-Up

Participants conducted a flight in each of two different graphics conditions: non-VR and VR. In the non-VR condition, visuals were presented using a Broad-Angle Display System (BADS), consisting of a large concave screen spanning 120 degrees, with eight projectors overhead projecting the simulated environment. The VR condition was presented through an Oculus Rift headset (Figure 3-3). The Oculus Rift has a resolution of 2160 x 1200, a 90 Hz refresh rate, and a 110 degree field of view. Figure 3-4 shows the graphics conditions for both the non-VR and VR environments respectively.

Figure 3-3: Oculus Rift headset, used for VR condition

Figure 3-4: Graphics for BADS/non-VR and VR condition (Left to Right)

Both conditions took place in the body of a 1966 Cessna 172 aircraft outfitted with working instruments and radio equipment, operating under a closed, simulated circuit. Participants interacted with physical instruments and controls (i.e., the yoke, throttle, and flaps), however could not see their hands during the VR condition. The remaining materials and set-up for both conditions were identical.

The flight scenario was rendered by the Prepar3D flight simulation software and consisted of an airfield named “Pendleton Airfield”, which was in reality a simulation of the Hong Kong airport. A computer recorded deviations from the prescribed flight path, including air-speed, heading, and altitude. The simulation included pre-recorded radio messages (typical aircraft radio chatter) that participants listened to through a headset. The headset had a microphone, which participants used to confirm they received instructions and to make radio calls, as instructed by the procedure.

Participants wore the Empatica E4 wristwatch biosensor on their left wrist during the flights, which collected their real-time heart rate and GSR at 1 and 4 Hz respectively (Garbarino, Lai, Bender, Picard, & Tognetti, 2014). This device uses a photoplethysmography (PPG) sensor to measure blood volume pulse (BVP) from which heart rate is derived, and a GSR sensor to measure sympathetic nervous system arousal. Both sensors are located on the back of the device, shown in Figure 3-5. The device records the data locally which is later transferred to a computer for processing. Participants also wore a thumb switch on their left thumb, which they were instructed to press against the yolk anytime they heard an auditory tone (i.e., ‘beep’) through the headset during their flights. The reaction time from this peripheral detection task (PDT) is automatically recorded by the computer, to use this information as a secondary workload measure. Figure 3-5 shows the PDT setup, and also the interior of the cockpit of the Cessna 172 aircraft.

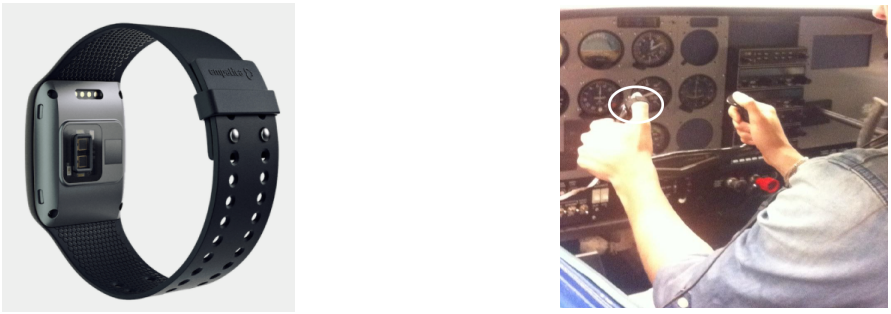


Figure 3-5: Empatica E4 wrist watch to measure heart rate and GSR, and thumb switch for PDT

3.3. Procedure

Participants were briefed about the experiment through a slide show presentation that explained the procedure and gave them an opportunity to ask questions. Participants filled out a demographics questionnaire, found in Appendix B that inquired about their flying experience, exposure to video games (both VR and non-VR), and their state of wellbeing (e.g., tiredness, queasiness, dizziness, etc.). The Empatica E4 was strapped onto the participant's wrist, and they proceeded into the Cessna 172 simulator where they were outfitted with the thumb switch, headset and microphone. Participants were taught to "fly" the Cessna 172 simulator using just the yoke, flaps, and throttle. To simplify the task, the rudders were set to 'auto" and not used by the participants. All participants completed three practice circuits of the flight scenario using the BADS graphics condition, while being coached by a flight instructor sitting in the co-pilot seat. Instructions progressed gradually in difficulty level and focused on familiarizing the participant to use the yoke, throttle, flaps, thumb switch, and finally make radio calls using a standard script. The script, "*Pendleton Airfield, this is Alpha-Brave-Charlie. Flying through [insert shape] hoop*" was posted on the dash of the aircraft so participants did not have to memorize it. It remained there for the experimental tasks, however it was out-of-sight during the VR condition due to the goggles. The Empatica E4 recorded the participants' heart rate and GSR during each of the flight conditions.

The experimental tasks, described in Section 3.4, were repeated in both the VR and BADS graphics conditions. Presentation order was counterbalanced to mitigate carryover effects (i.e., half of the participants performed the VR condition first and the other half performed the BADS condition first). Participants could take an optional break between the two conditions.

3.4. Experimental Flight Measures

Each participant flew three circuits of Pendleton Airfield for each graphics condition (i.e., BADS and VR) while seated in the pilot side of the Cessna 172 flight simulator. The following measures were collected during or after exposure (Figure 3-6).

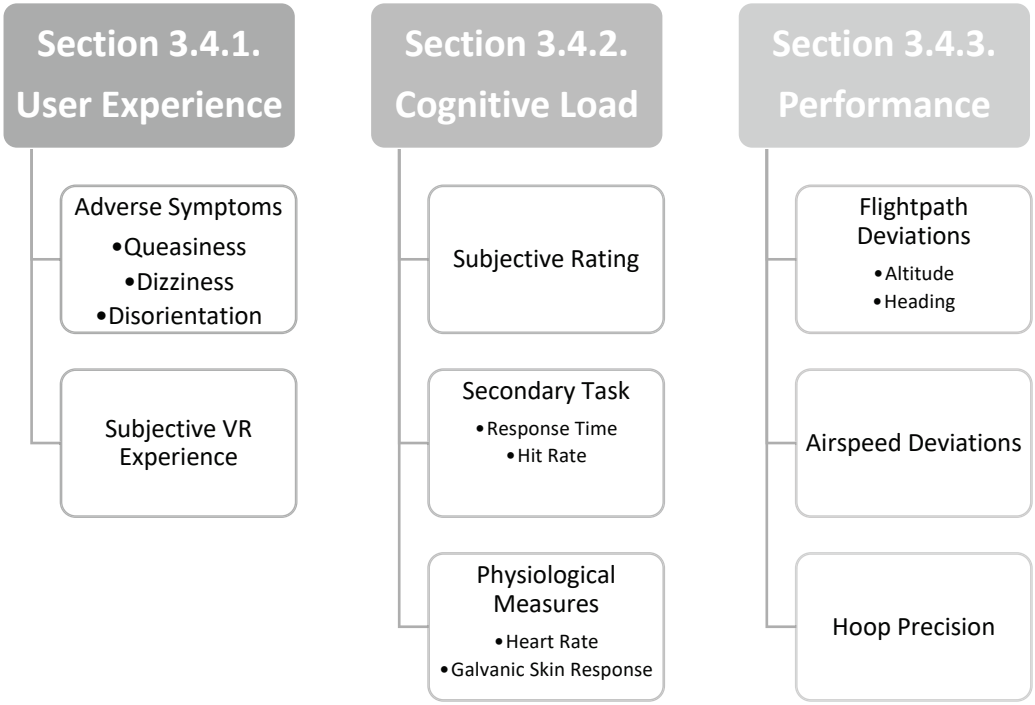


Figure 3-6: Overview of measures captured during/after each graphics condition

3.4.1. User Experience

Adverse Symptoms. Participants self-rated their level of queasiness, dizziness, and disorientation on a scale of 1 (*not at all*) to 7 (*very*) before beginning the assessment (i.e., baseline), and after each of the flights in the two graphics conditions. These questions were mostly adapted from the Simulator Sickness Questionnaire to assess any physiological side effects as a result of being immersed in a simulated environment (Kennedy et al., 1993), specifically rating queasiness, dizziness, and disorientation they may have experienced.

The results are presented as a difference between their score from the BADS/VR and their baseline score, where a negative result means the participant experienced less symptoms than during their baseline while a positive value indicates the participant experienced symptoms during the condition that are greater than their baseline.

Subjective VR Experience. Following the VR flight, participants answered questions about their experience with the Oculus Rift. These questions inquired about the comfort of the goggles, their perceived weight, the crispness and realness of the visual images, and any other physical symptoms they felt during the simulation. See Questions 26-34 in Appendix A for complete list of questions.

3.4.2. Cognitive Load

Subjective Rating. After each condition, participants rated their perceived mental workload on a Likert scale, rated from 1 (*not at all difficult*) to 7 (*very high workload*).

Secondary Task (PDT). Participants heard auditory tones in the headset played randomly every 5 to 10 seconds throughout their flights and were instructed to press the thumb switch against the yoke

anytime they heard one. Their response rate and reaction time was automatically logged. This task assumes that as workload demands of the primary task (i.e., piloting the aircraft) increase, the participant's capacity devoted to the secondary task (i.e., the PDT) decreases (Newman & Greeley, 2016). As a result, the workload demands of the primary task are inferred based on the ***response time*** and ***hit rate*** to the tone (i.e., as response time increases and hit rate decreases, we infer that the cognitive workload of the primary task also increases).

Physiological Measures. The Empatica E4 wristwatch biosensor recorded participants' heart rate and GSR continuously throughout all flights, using the PPG and GSR sensors located at the rear of the device. The Empatica E4 stores up to 60 hours of data, however we transferred the recordings from the device to the Empatica Connect dashboard (which allows access to the data) after no more than five consecutive participants (i.e., less than three hours of data) to maintain data control. Although the device recorded continuously, we averaged the heart rate and GSR outputs for each of the three circuits for each graphics condition to conduct the analysis.

i. Heart Rate: The PPG sensor illuminates a green and red light emitting diode (LED) onto the ventral portion of the participant's wrist and combines these signals in a way to optimize the detection of the pulse wave (i.e., estimation of the heart rate) (Empatica, 2016). A portion of the light is reflected back and measured by a light receiver. Figure 3-7 depicts this process, where it can be noted that the time occurrences of the valleys in the measured light during green exposure are used to estimate the pulse wave (i.e., heart rate). As the blood is more oxygenated, more light is absorbed, indicating a lower heart rate, and the reverse is also true. The time exposure during the red light is used to cancel out motion artefacts, as it contains a reference light level. An algorithm converts the two observed light signals into a BVP reading at 64Hz, which then converts to a beats per minute (bpm) heart rate reading, averaged for every second of data.

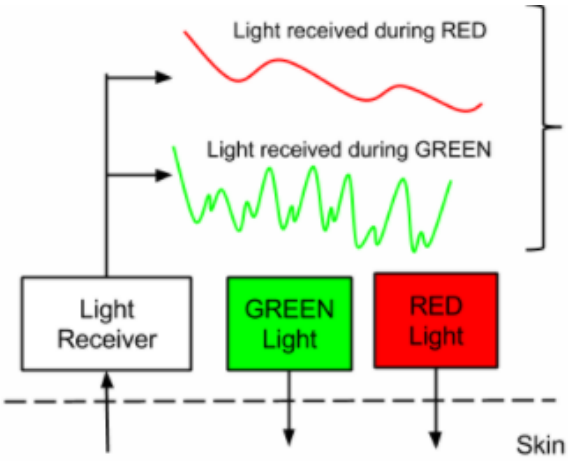


Figure 3-7: Empatica E4 PPG sensor (Empatica, 2016)

ii. Galvanic Skin Response: A GSR sensor measures electrical skin conductance by passing a small amount of current across two electrodes in contact with the ventral portion of the wrist (Empatica, 2016). This measurement output is in microSiemens (µS), sampled at 4 Hz. We preprocessed the raw GSR signal before conducting any analysis. First, it was truncated from

four recordings per second to one by only keeping every fourth data-point. Since GSR consists of a tonic and phasic component (sometimes referred to as the Skin Conductance Level (SCL) and Skin Conductance Response (SCR), respectively), where only the phasic/SCR component is related to emotional stimuli/attention processes (Henriques, Paiva, & Antunes, 2013), we split the two components using a deconvolution method (Benedek, 2010). We focus our analysis on the phasic/SCR component. For each flight condition for each participant, we extracted the number of SCR peaks and the average amplitude-sum of the peaks.

3.4.3. Performance Metrics

Participants used the yoke, throttle, and flaps to control the aircraft and were instructed to maintain 100 knots airspeed during the downwind leg and maneuver the aircraft through the centre of green holographic hoops that were placed along the flight path. Figure 3-8 illustrates the circuit, showing the location of the nine hoops.

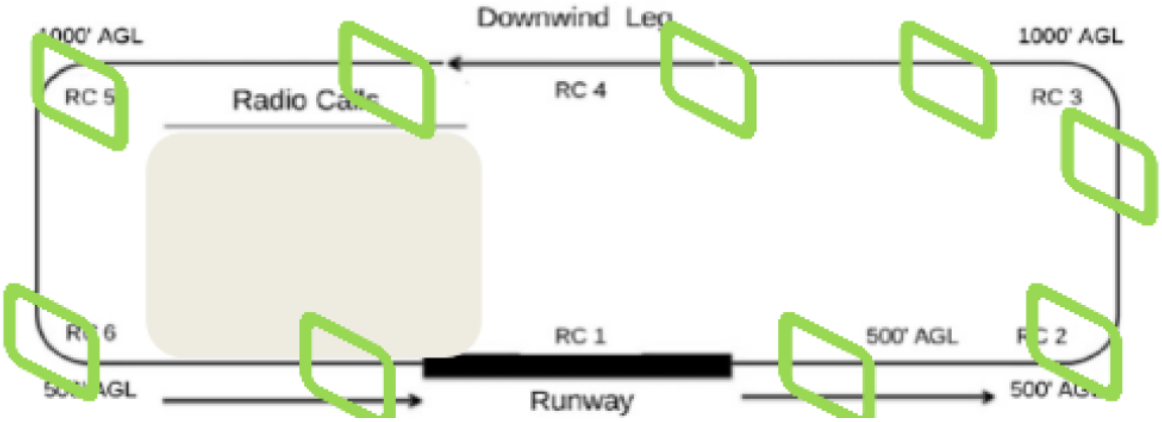


Figure 3-8: Pendleton Airfield flight circuit with hoops

Each hoop had a holographic shape above it, either a question mark, downward facing arrow, or an arrow within a circle. These shapes were referred to as "cues" as to when the participant was to make a radio call using the standard script. At the beginning of each circuit, participants were advised what the cue shape was for that circuit, indicating that they were required to make a radio call while flying through the *subsequent* hoop. For example, if the cue was a question mark, the participant was expected to make a radio call whenever they encountered (or flew through) the hoop *after* the hoop with a question mark. The cue shape changed for each of the three circuits within each graphics condition, and the order was reversed for the next condition. Participants also listened to other aircraft chatter in their headset and were instructed to remember the call signs of as many aircraft as possible. While both the radio call information (testing prospective memory) and aircraft call sign information (testing situational awareness) were not used for analysis, they contributed to increasing the overall cognitive demand of the flying task.

Meanwhile, the simulation software recorded the position and orientation of the plane at every second, permitting calculations of deviations from the optimal altitude, heading, and airspeed to objectively rate the participants' flying performance. Additionally, the software recorded the location where the participant flew through each hoop to rate the precision. The software recorded if the plane went through the hoop near its center, near the periphery, or missed the hoop entirely.

4. Results

The analysis compares the user experience, the cognitive load, and the flight performance in two graphics conditions: standard BADS graphics and graphics provided solely via a VR platform.

Because hypotheses in this work postulated a null effect, Bayesian statistical analysis was conducted alongside the more commonly used Frequentist statistical tests (which are not designed to test the truth of a null hypothesis) (Rouder, Speckman, Sun, & Morey, 2009). Bayesian statistics provide estimates of the evidence regarding the probabilities of the null and/or alternative hypothesis, thus permitting an analysis of the probability of there being no difference between the two graphics conditions for various parameters (e.g., mental workload, performance, etc.) compared to the alternative hypothesis that there was an effect of graphics condition. Understanding how VR does, or does not, impact pilot performance and experience in a simulator is an important preliminary step in the development of VR based training.

We used the JASP statistical package (jasp-stats.org) to perform both the frequentist and Bayesian[4] analyses. When an effect of graphics condition was found, those results were further explored to determine if the effects were moderated by individual factors, such as gender or video game experience.

The first part of the analysis focuses on comparing the Human Factors aspects of the piloting task in the two graphics conditions. These are broken down into User Experience metrics (i.e., Adverse Symptoms, and VR Experience Questionnaire results) and the Cognitive Load (i.e., Subjective Rating, Secondary Task, and Physiological Measures). The second part of the analysis presents the Performance Metrics, as measured by the flight path deviations (i.e., heading and altitude), airspeed deviation, and hoop precision while flying in the two graphics conditions. Figure 3-6 outlines the structure of the results section of this report.

4.1. High Level Overview

Paired samples t-tests (both Frequentist and Bayesian) were conducted to compare the measures within the three categories (i.e., User Experience, Cognitive Load, and Performance Metrics) between the two graphics conditions (i.e., BADS and VR). Independent samples t-tests (both Frequentist and Bayesian) were used when looking at only the first or second exposure condition. Table 4-1 shows a summary of all the findings.

There were no differences in User Experience (on first flying condition) between the two graphics conditions, however there was an increase in queasiness after VR as a second condition. The remaining VR experience metrics are presented in Section 4.2.2, but not in Table 4-1 since they are not being compared to the BADS condition.

The VR graphics conditions resulted in a higher subjective rating of cognitive load, heart rate, and GSR (measured by SCR peaks). The Performance Metrics suggest that the BADS graphics condition had a greater airspeed RMSE than the VR (i.e., participants were able to keep their speed closer to the

[4] Given that research regarding the effects of VR with naïve subjects in a Cessna simulator has not been performed before, we defaulted to using Bayesian tests with a non-informative Cauchy prior of 0.707.

recommended 100 knots during the downwind section in the VR graphics condition than in the BADS). There were no significant differences for the remaining measures (i.e., $p > 0.05$, $BF_{10} < 3$). **Error! Reference source not found.**

Table 4-1: High level overview of results

Measure			Mean	SD	P-Value	Bayes Factor (BF_{10})
User Experience	**Queasiness after 2nd exposure**	BADS	1.32	0.58	**0.05***	**3.32**
		VR	1.67	1.23		
	Dizziness after 2nd exposure	BADS	1.53	0.70	0.35	0.32
		VR	2.00	1.41		
	Disorientation after 2nd exposure	BADS	1.53	0.70	0.23	0.92
		VR	1.91	1.09		
Cognitive Load	**Subjective Rating** (subjective score, 7 = high mental effort)	BADS	4.88	1.23	**0.012***	**3.888**
		VR	5.37	1.22		
	Secondary Task - Response Time (ms)	BADS	0.998	0.39	0.198	0.395
		VR	1.031	0.39		
	Secondary Task - Hit Rate (%)	BADS	87.43	16.26	0.164	0.462
		VR	84.1	21.1		
	Heart Rate (bmp)	BADS	80.64	9.20	**0.027***	**2.017**
		VR	82.65	8.85		
	GSR (SCR Peaks)	BADS	40.60	30.21	**0.001***	**35.23**
		VR	54.10	32.42		
	GSR (Peak Amplitude, µS)	BADS	0.07	0.09	0.062	1.024
		VR	0.11	0.14		
Performance Metrics	Altitude (RMSE)	BADS	82.36	32.84	0.169	0.596
		VR	96.78	44.45		
	Heading (RMSE)	BADS	2.36	0.91	0.142	0.685
		VR	2.82	0.33		
	Airspeed (RMSE)	BADS	10.91	3.71	**0.009***	**6.536**
		VR	8.82	1.92		
	Center of Hoop (n)	BADS	24.24	6.37	0.341	0.351
		VR	24.94	1.92		
	Hoop Periphery (n)	BADS	27.47	7.17	0.299	0.392
		VR	28.47	2.00		

**** p < .05***

4.2. User Experience

4.2.1. Adverse Symptoms after First Exposure

The analysis considers only the first condition the participant is exposed to (i.e., either BADS or VR) when analyzing the results for subjective physiological symptoms as feelings of queasiness, dizziness, and disorientation could have a cumulative effect and ultimately be rated higher in the second condition regardless of the actual graphics condition being presented.

Independent samples t-tests showed that there were no differences in queasiness, dizziness, and disorientation ratings after first exposure, t(39) = -0.38, *p = 0.35*; t(39) = 0.82, *p = 0.21*; and t(39) = 0.21, *p = 0.42* respectively. Bayesian Statistics further suggest there is no difference, with BF_{10} = 0.41, BF_{10} = 0.612, and BF_{10} = 0.36, respectively.

Next, the analysis considered only the second condition the participant is exposed to (i.e., either BADS or VR) to determine if there is a physiological effect that differs between the two graphics conditions, if it is primed by the other condition. No significant results were found in queasiness, dizziness, and disorientation after second exposure, between the VR and BADS conditions, discussed in Section 4.2.1.1.

4.2.1.1. Adverse Symptoms after Second Exposure

The mean queasiness rating appeared greater for participants who flew in the VR graphics condition on second exposure (mean = 1.68, SD = 1.2) than for those who flew in the BADS graphics condition on second exposure (mean = 1.32, SD = 0.58), and a repeated measures ANOVA suggests that the difference is statistically significant, F(1, 38) = 4.05, *p = 0.05*. There were no differences in dizziness and disorientation scores on second exposure, between the BADS and VR: F(1, 38) = 0.91, *p = 0.35*, and F(1, 38) = 1.51, *p = 0.23.* A Bayesian repeated measures ANOVA, where BF_{10} = 3.32, suggests that it is 3.32 times more likely that the VR graphics condition on second exposure resulted in more queasiness than the BADS graphics condition on second exposure. There were no differences in dizziness and disorientation after second exposure, where BF_{10} = 0.32 and BF_{10} = 0.92, respectively.

These results suggest that although participants who experienced the BADS graphics condition on first exposure (BADS first, then VR) had no queasiness symptoms, after experiencing the VR graphics condition second, they felt queasier than the participants who experienced the two graphics conditions in the reverse order (i.e., VR first, then BADS).

4.2.2. Subjective VR Experience

Participants were asked to rate how comfortable they found wearing the Oculus Rift at the onset of and towards the end of their flying experience. Figure 4-1 presents the results for both times, where *1 = Very Uncomfortable* and *7 = Very Comfortable*. While none of the participants reported being *Very Uncomfortable* at either point in time, participants generally found it more uncomfortable at the onset (mean = 4.45, SD = 1.28), and gradually considered it more comfortable near the end (mean = 5.3, SD = 1.45). A paired samples t-test shows the increase in comfort was significant, t(39) = -3.48, *p = 0.001*.

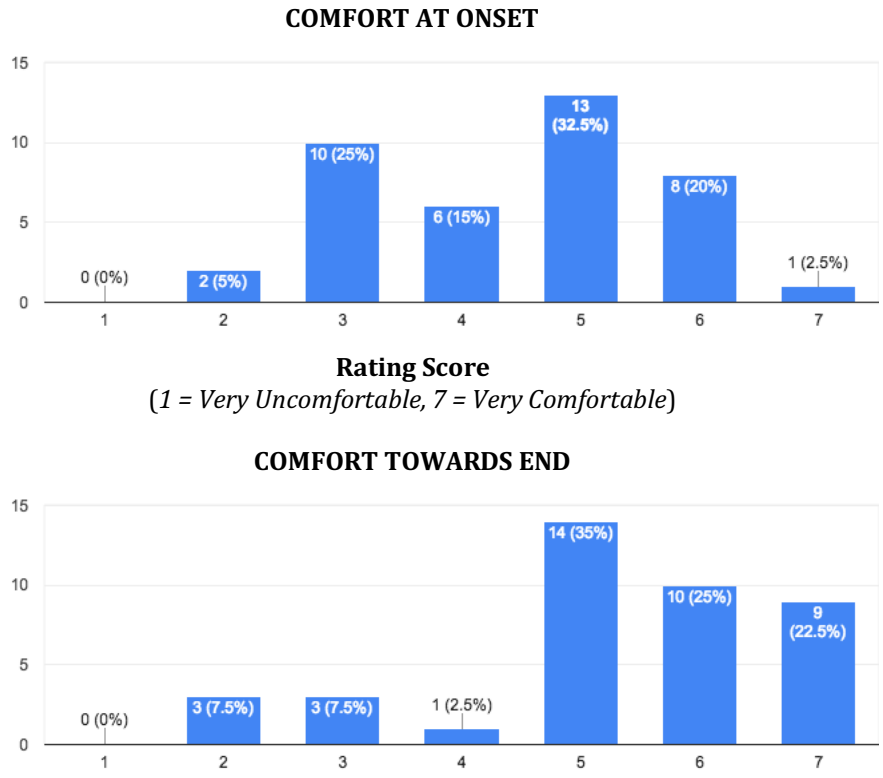


Figure 4-1: Comfort rating Oculus Rift at onset and towards end of flying experience

Participants rated the perceived 'heaviness' of wearing the Oculus Rift. Figure 4-2 presents the results, where *1 = Very Light* to *7 = Very Heavy*. None of the participants perceived it as *Very Heavy*. Participants generally followed a bell-curve distribution with their heaviness ratings, with an average rating of 4.6 (SD = 1.26), indicated results are slightly skewed towards the *'heavier'* end.

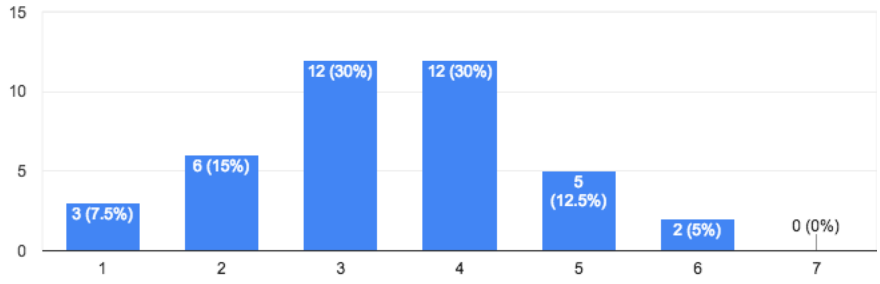
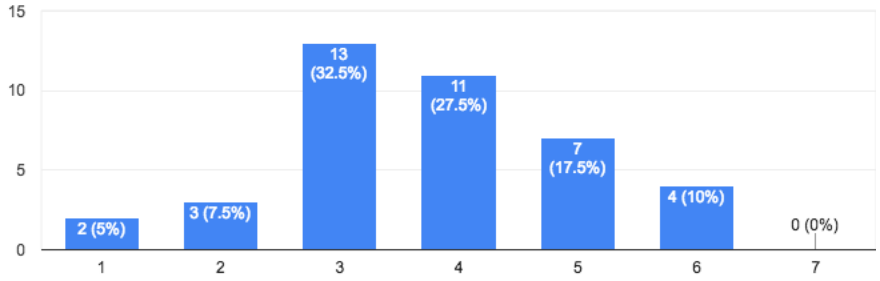


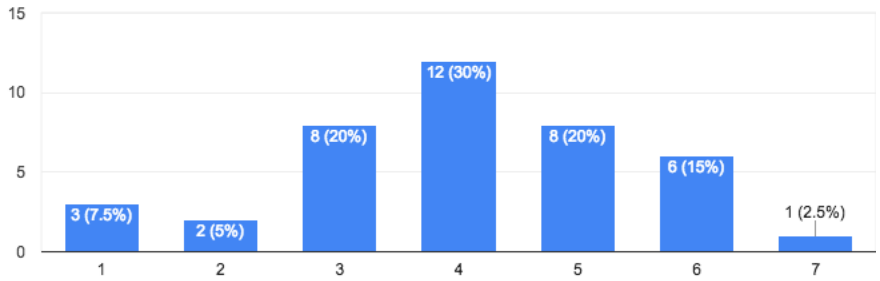
Figure 4-2: Perceived heaviness of Oculus Rift

Participants rated the visual scene in the Oculus Rift, in terms of its crispness and the realness. Figure 4-3 and Figure 4-4, respectively, show their ratings, where *1 = Not at all Crisp* to *7 = Extremely Crisp;* and *1 = Not at all Real* to *7 = Very Real;* respectively. The average rating for crispness was 3.75 (SD = 1.28).

The average rating for realness was 4.05 (SD = 1.47). Both ratings fall near the middle of the scale, although very slightly skewed towards *Very Crisp* and *Very Real* ends, respectively (versus *Not at all Crisp*, and *Not at all Real*, respectively).



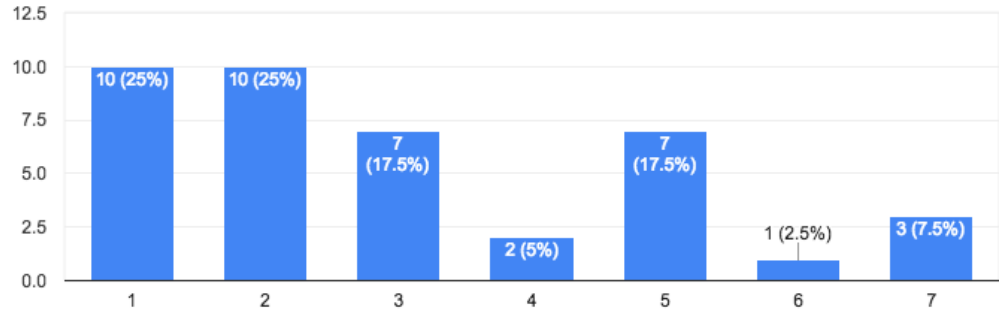
Rating Score
(*1 = Not at all Crisp, 7 = Extremely Crisp*)
Figure 4-3: Perceived crispness of visuals in Oculus Rift



Rating Score
(*1 = Not at all Real, 7 = Extremely Real*)
Figure 4-4: Perceived realness of visuals in Oculus Rift

To further explore the user experience with the Oculus Rift, participants were asked about the fact that they could not see their own hands on the yoke, or when reaching for the throttle and flaps. Figure 4-5 presents the results for the question "*how would you describe the fact that you could not see your own hands on the yoke in the Oculus Rift?*" where *1 = I barely noticed, and this did not affect my flying* and *7 = This really bothered me, and affected my flying*. Similarly, Figure 4-6 presents the results to the same question, but regarding the throttle and flaps (instead of the yoke), using the same rating scale.

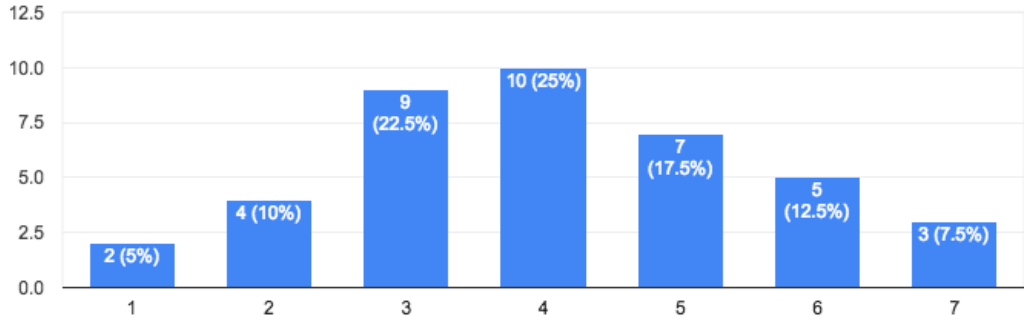
Figure 4-5 illustrates that fifty percent of participants (*n = 20*) were content not being able to see their hands on the yoke (i.e., rating of 1 or 2), with a mean rating of 3.03 (SD = 1.87). Figure 4-6 illustrates that participants had a normal distribution when rating how they felt about not being able to see their hands when using the flap/throttle (mean = 4.08, SD = 1.58). A paired samples t-test suggests this difference is significant, t(39) = -3.13, *p = 0.002,* where participants were more bothered by not being able to see their hands reaching for the throttle/flaps than not being able to see their hands on the yoke. This can likely be attributed to the fact that participants had to reach for the throttle/flaps many times throughout the flight, whereas at least one hand generally stayed put on the yoke. The yoke was also larger therefore likely easier to locate than the throttle/flaps.



Rating Score

(1 = I barely noticed, 7 = this really bothered me and affected my flying)

Figure 4-5: "How would you describe the fact that you could not see your hands on the yoke?"



Rating Score

(1 = I barely noticed, 7 = this really bothered me and affected my flying)

Figure 4-6: "How would you describe the fact that you could not see your hands on the throttle or flaps?"

4.3. Cognitive Load

4.3.1. Subjective Rating

The average mental workload rating was higher in the VR graphics condition (mean = 5.4, SD = 1.2) than in the BADS (mean = 4.9, SD = 1.3), t(40) = -2.36, *p = 0.012*. Using a Bayesian paired samples t-test, the Bayes Factor (BF_{10} = 3.9) suggests that the data were about 3.9 times as likely to occur in favor of the alternative hypothesis, providing moderate evidence that a participant's self-rated mental workload is higher when flying using the VR graphics condition than the BADS. Figure 4-7 illustrates this relationship.

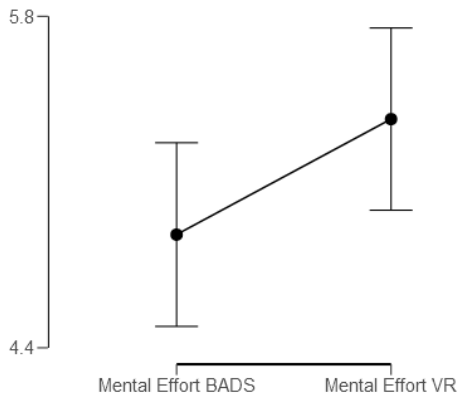


Figure 4-7: Self-reported mental effort in BADS vs VR. Error bars indicate 95% credible intervals[5].

We investigated whether individual factors might moderate the effects of graphics condition on mental effort. Table 4-2 shows that neither age, gamer experience, or gender was significantly associated with participants rating of mental workload for both graphics conditions. As expected, mental workload rating in BADS was associated with the rating in the VR condition. No order effects were noted.

Table 4-2: Pearson Correlations to investigate moderating effects of graphics condition on mental effort

	Mental Effort BADS	Mental Effort VR	Age	Gender	Gamer	CB
Mental Effort BADS	—					
Mental Effort VR	0.414**	—				
Age	−0.112	−0.032	—			
Gender	0.241	0.011	−0.216	—		
Gamer	−0.194	0.006	0.009	−0.198	—	
CB	−0.269	−0.120	0.146	−0.108	−0.055	—

* p < .05, ** p < .01, *** p < .001

4.3.2. Secondary Task - PDT

The PDT suggests that the mental workload, as measured by response time and hit rate on a secondary task, was essentially the same in the BADS and VR graphics conditions.

A. Response Time

Response time to the PDT was measured as the time between the audible tone and the time when the participant pressed the thumb switch. Therefore, a higher response time indicates the participant took longer to detect the sound, providing evidence for a higher cognitive demand in the primary task (i.e., flying the aircraft). There was no difference between the two graphics conditions, t(36) = -0.86, *p = 0.198*. Similarly, a Bayesian paired samples t-test (BF_{10} = 0.395) suggests that there is no difference in the response times between the two graphics conditions.

[5] A credible interval is the interval in which an (unobserved) parameter has a given probability. It is the Bayesian equivalent of the confidence interval.

B. Hit Rate

The Hit Rate indicates the percentage of beeps that the participant detected, therefore a lower hit rate suggests higher workload in the primary task (i.e., flying the airplane). The average PDT hit rate, as a percentage of beeps detected, was lower in the VR graphics condition (mean = 84.1%, SD = 21.1) than in the BADS (mean = 87.43%, SD = 16.26), however a paired samples t-test showed that this difference was not significant, t(36) = -0.93, p = 0.164. A Bayesian paired samples t-test (BF_{10} = 0.462) suggests that there is no difference in hit rate between the two graphics conditions.

4.3.3. Physiological Measures

The objective physiological data is measured by the heart rate and GSR responses, described respectively.

4.3.3.1. Heart Rate

The average heart rate was higher in the VR graphics condition (mean = 82.65, SD = 8.85) than in the BADS (mean = 80.64, SD = 9.2), t(35) = -1.99, *p = 0.027*. Using a Bayesian paired samples t-test, the Bayes Factor (BF_{10} = 2.017) suggests that the data were about twice as likely to occur in favor of the alternative hypothesis, providing anecdotal evidence that a participant's heart rate is higher when flying using the VR graphics condition than BADS. Figure 4-8 illustrates this difference, where the error bars indicate the 95% credible interval.

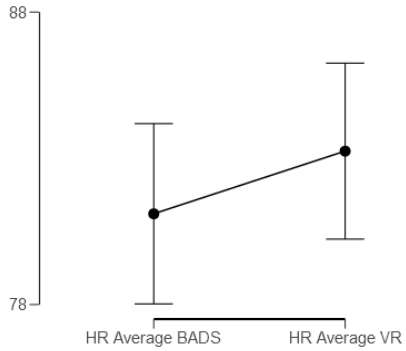


Figure 4-8: Mean heart rate in BADS vs VR. Error bars indicate 95% Credible Intervals.

We investigated the potential for individual factors to moderate the effects of graphics condition on heart rate. Table 4-3 shows that neither age, gender, nor gamer experience was significantly associated with heart rate in either graphics condition. No order effects were noted (i.e., CB factor).

Table 4-3: Pearson Correlations to investigate moderating effects of graphics condition on heart rate[6]

	HR BADS	HR VR	Age	Gender	Gamer	CB
HR BADS	—					
HR VR	0.743***	—				
Age	−0.019	0.218	—			
Gender	0.318	0.220	−0.216	—		
Gamer	0.314	0.231	0.009	−0.198	—	
CB	−0.201	0.105	0.146	−0.108	−0.055	—

* p < .05, ** p < .01, *** p < .001

4.3.3.2. Galvanic Skin Response

The average number of SCR peaks was higher in the VR graphics condition (mean = 54.09, SD = 32.42) than in the BADS (mean = 40.60, SD = 30.21), t(31) = -3.37, *p = 0.001*. Using a Bayesian paired samples t-test, the Bayes Factor (BF_{10} = 35.23) suggests that the data were 35.23 more likely to occur in favor of the alternative hypothesis, providing very strong evidence that GSR arousal is higher in the VR graphics condition than in the BADS.

The average SCR amplitude was marginally higher in the VR graphics condition (mean = 0.105 µS, SD = 0.139) than in the BADS (mean = 0.068 µS, SD = 0.094), t(35) = -1.57, *p = 0.062*. Using a Bayesian paired samples t-test, the Bayes Factor (BF_{10} = 1.024) suggests that the data were only 1.024 times more likely to occur in favor of the alternative hypothesis, providing essentially no evidence to suggest that the VR graphics condition elicited a higher GSR arousal response than the BADS, as measured by average SCR amplitude.

These findings may suggest that participants experienced more physiological arousal in the VR graphics condition than in the BADS, as measured by the average number of SCR peaks. However, the results of the GSR amplitude sum show do not mirror these findings, suggesting that perhaps the two GSR measures have a different sensitivity. Figure 4-9 illustrates these two differences, where the error bars indicate a 95% (Bayesian) Credible Interval.

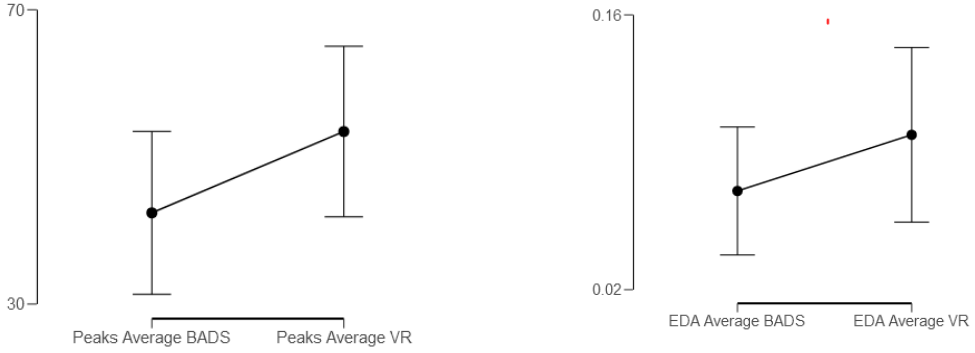


Figure 4-9: Average # of SCR Peaks between BADS and VR graphics condition (left) and average amplitude sum of SCR between BADS and VR (right). Error bars indicate 95% Credible Intervals.

[6] CB = counterbalance order, where 1 = BADS first and 2 = VR first

We investigated the potential for individual factors to moderate the effects of graphics condition on either SCR peaks or average SCR amplitude. Table 4-4 shows that gender may have a moderating effect on SCR peaks in the VR graphics condition. However, neither age nor gamer experience was significantly associated with SCR peaks or SCR amplitude in either graphics condition, and no order effects were noted.

Table 4-4: Pearson Correlations to investigate moderating effects of graphics condition on SCR amplitude and SCR peaks

	SCR Amplitude BADS	SCR Amplitude VR	SCR Peaks BADS	SCR Peaks VR	Age	Gender	Gamer	CB
SCR Amplitude BADS	—							
SCR Amplitude VR	0.578***	—						
SCR Peaks BADS	0.781***	0.585***	—					
SCR Peaks VR	0.638***	0.803***	0.827***	—				
Age	0.063	−0.095	0.032	0.051	—			
Gender	−0.111	−0.038	−0.221	−0.340*	−0.216	—		
Gamer	0.216	0.134	0.094	0.066	0.009	−0.198	—	
CB	0.047	0.280	−0.167	0.118	0.146	−0.108	−0.055	—

* p < .05, ** p < .01, *** p < .001

4.3.3.2.1. Moderating Effects of Gender on Number of SCR Peaks

To further investigate the differences on the number of SCR peaks in the two graphics conditions, we ran a two-way mixed ANOVA to determine if gender has a moderating effect. Mauchly's test of sphericity indicated that the assumption of sphericity was met therefore no correction was required. There was no statistically significant interaction between the graphics condition and gender on the number of SCR peaks, $F(1, 30) = 0.74$, $p = 0.398$, partial $\eta 2 = 0.024$. A Bayesian repeated measures ANOVA supported the findings, suggesting no moderating effects of gender on the number of SCR peaks, $BF_{01} = 2.11$.

There was however, a main effect of graphics and gender. The main effect of graphics condition showed a statistically significant difference in the number of SCR peaks in the BADS or VR condition, $F(1, 30) = 8.55$, $p = 0.007$, partial $\eta^2 = 0.222$, where participants had a greater number of SCR peaks in the VR than the BADS condition. This main effect was also found with Bayesian statistics, where $BF_{10} = 15.97$, providing strong evidence that there is a difference between the two graphics conditions.

The main effect of gender showed that there was a near statistically significant difference in average number of SCR peaks between males and females $F(1, 30) = 3.65$, $p = 0.066$, partial $\eta^2 = 0.109$, where males had a greater number of SCR peaks than females, as illustrated in Figure 4-10. A Bayesian repeated measures ANOVA found only weak evidence to suggest that there is a main effect of gender, $BF_{10} = 1.506$.

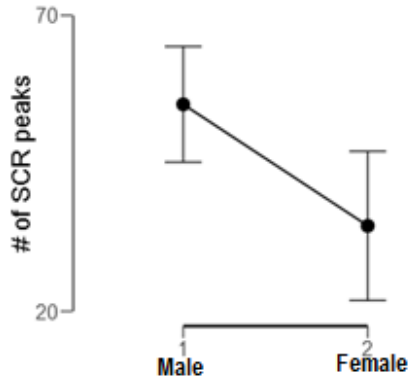


Figure 4-10: Main effect of gender on # of SCR peaks. Error bars indicate a 95% confidence interval.

4.4. Performance Metrics

To mitigate against any learning effects that may have occurred from the first condition and carried over into the second condition, only performance metrics from the first condition are considered. Performance between the BADS and VR graphics condition was compared by examining the root-mean-squared-error (RMSE) of data logged on the downwind segment of each circuit, by averaging the squared differences between optimal and observed performance. Flight path deviations (i.e., heading, altitude) are considered, followed by airspeed deviations, where a lower RMSE indicates a better overall performance (i.e., it is more in line with the desired path or speed, respectively). Finally, the flying precision through the green holographic hoops is considered, and whether participants flew near the center of the hoop (as instructed) or towards the periphery.

Two participants were unable to complete the flying task successfully, consequently their outlier data was removed from the set. Technical issues resulted in lost data from six participants in the hoop flying task.

4.4.1. Flight Path Deviations

The average altitude RMSE appeared to be greater in the VR graphics condition (mean = 96.789, SD = 44.45) than in the BADS (mean = 82.363, SD = 32.845), however, a paired samples t-test suggests there is no difference between the two graphics conditions, t(18) = -0.99, p = 0.169. A Bayesian samples t-test supported this finding, where BF_{10} = 0.596, indicating support for the null hypothesis that there is no difference in the altitude RMSE between the two graphics conditions.

The average heading RMSE appeared greater in the VR graphics condition (mean = 2.822, SD = 0.327) than in the BADS (mean = 2.356, SD = 0.911), where a paired samples t-test found this difference not statistically significant, t(18) = -1.11, p = 0.142. To substantiate these results, a Bayesian paired samples t-test was conducted, where the Bayes Factor (BF_{10} = 0.685) provides essentially no evidence to support that the heading RMSE was higher in the VR condition than the BADS condition. Figure 4-11 illustrates both the altitude and heading RMSE.

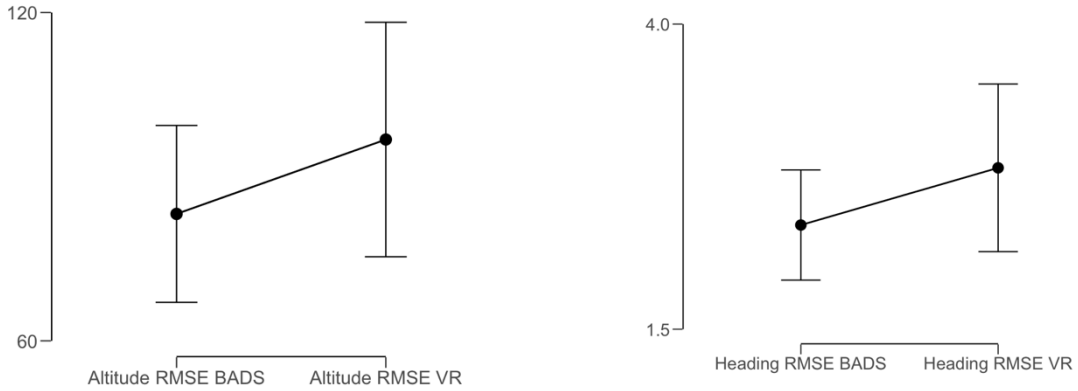


Figure 4-11: Altitude RMSE and heading RMSE in the BADS versus VR. Error bars indicate 95% credible interval.

4.4.2. Airspeed Deviation

The average airspeed RMSE was greater in the BADS graphics condition (mean = 10.911, SD = 3.714) than in the VR (mean = 8.825, SD = 1.925), where a paired samples t-test found this difference to be statistically significant, t(18) = 2.62, p = 0.009. To substantiate these results, a Bayesian paired samples t-test was conducted, where the Bayes Factor (BF_{10} = 6.536) provides evidence to support that the airspeed RMSE was greater in the BADS condition than the VR condition. This suggests that participants were able to keep their speed closer to the recommended 100 knots during the downwind section in the VR graphics condition than in the BADS. Figure 4-12 illustrates the airspeed RMSE in the BADS and VR conditions, where the error bars indicate the 95% credible interval.

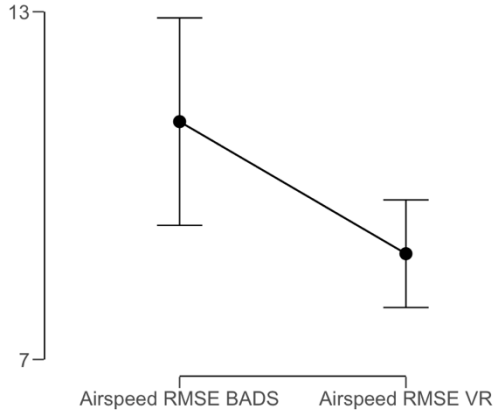


Figure 4-12: Airspeed RMSE in BADS and VR condition. Error bars indicate 95% credible interval.

We investigated the potential for individual factors to moderate the effects of graphics condition on the airspeed RMSE. Table 4-5 shows that neither gender nor gamer experience was significantly associated with airspeed RMSE in either graphics condition, and no order effects were noted.

Table 4-5: Pearson Correlations to investigate moderating effects of graphics condition on airspeed RMSE

	Airspeed RMSE BADS	Airspeed RMSE VR	Age	Gender	Gamer	CB
Airspeed RMSE BADS	—					
Airspeed RMSE VR	0.687***	—				
Age	−0.403*	−0.302	—			
Gender	0.208	0.256	−0.234	—		
Gamer	0.071	0.013	0.012	−0.141	—	
CB	−0.196	−0.197	0.153	−0.041	−0.105	—

* p < .05, ** p < .01, *** p < .001

4.4.3. Hoop Precision

Paired samples t-tests found no difference in the average number of hoops that participants maneuvered through near the center of the hoop, t(16) = -0.42, p = 0.341 or near the periphery t(16) = -0.54, p = 0.299, during their first flying task exposure[7]. To substantiate these results, Bayesian paired samples t-tests were conducted, where the Bayes Factors (BF_{10} = 0.351, BF_{10} = 0.392) provide essentially no evidence to support that there is a difference in flying precision between the VR condition and the BADS condition. Figure 4-13 and illustrate these two scenarios, respectively.

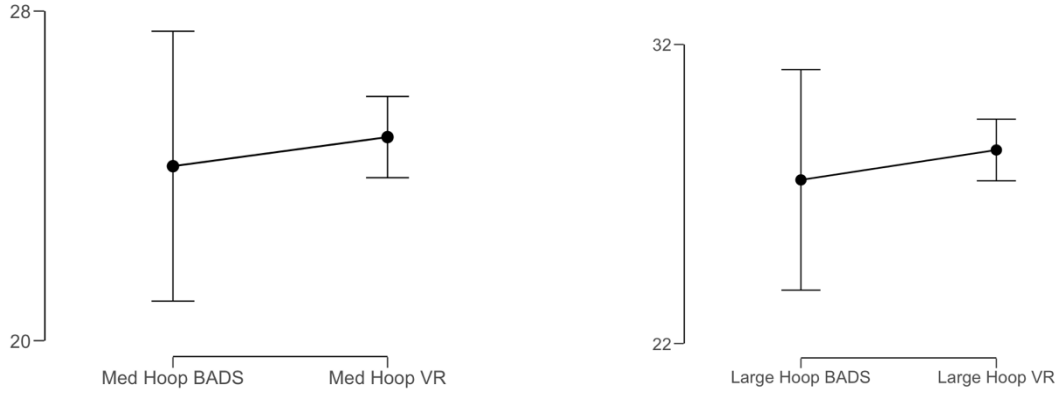


Figure 4-13: Flying precision as indicated by the number of hoops flown through the center (left), and near the periphery (right) in BADS and VR condition. Error bars indicate 95% credible interval.

4.4.4. Impact of Adverse Symptoms in VR on Flight Performance

A series of one-way ANOVAs were conducted to test the hypothesis that negative symptoms associated with the VR environment could impact performance in VR. Participants who scored a 3 out 7 or greater for each of queasiness, dizziness, or disorientation symptoms after their first flying experience in VR were grouped as symptomatic for each symptom, while the remainder were grouped as having no adverse symptoms from VR.

There was a statistically significant difference between queasiness groups on heading RMSE performance, as determined by a one-way ANOVA (F(1,17) = 7.32, *p = 0.015*), where post hoc tests revealed that the group experiencing queasiness symptoms in VR had a larger heading RMSE in VR.

[7] Data from six participants removed due to technical errors with software.

Similarly, there was a statistically significant difference between dizziness groups on heading RMSE performance, as determined by a one-way ANOVA ($F(1,17) = 10.46$, $p = 0.018$), where post hoc tests revealed that the group experiencing dizziness symptoms in VR had a larger heading RMSE in VR.

The remainder of the groups (queasiness/no queasiness, dizziness/no dizziness, disorientation/no disorientation) had no significant effects on performance measures in VR (heading, altitude, airspeed). There were no significant effects on performance measures in the BADS groups.

5. Discussion and Conclusion

5.1. Overview

The objective of the present research was to determine whether a VR is a viable alternative to the traditional graphics displays used in flight simulators. The user experience, cognitive load, and performance metrics of flying a simulator were compared across two display screen conditions: a traditional broad angle (BADS) display versus a VR display.

5.2. Research Questions

The present research leveraged the interaction-centered framework of experience, put forth by Forlizzi and Battarbee (2004), and subsequently took an objective perspective to evaluate a user's experience and interaction. To understand *cognitive* and *expressive* experiences, the present research captured the interactions both as they unfolded and any expression after the fact. This was achieved by evaluating performance metrics during the experience and evaluating any negative effects from the immersion environment after the task. The present work measured physiological indices of cognitive load enabling capturing the experience of flying a flight simulator without disruption to the task to understand the *fluent* aspects of the VR flying experience.

5.2.1. User Experience

Q1. Are there differences in the user experience between the two graphics conditions?

Overall, the user experience in VR for this particular flying task was suitable, and comparable to that in the BADS. The adverse effects between the two graphics conditions were compared (i.e., queasiness, dizziness, and disorientation), while the remaining user experience metrics were gathered from a subjective questionnaire about the VR HMD experience.

No differences in BADS and VR in terms of queasiness, dizziness, or disorientation ratings between baseline and first exposure. Queasiness increases in VR after second exposure.

The first condition the participant is exposed to is considered (i.e., either BADS or VR) when comparing the queasiness, dizziness, and disorientation as these symptoms could have a cumulative effect and ultimately be rated higher in the second condition regardless of the actual graphics condition being presented. There were no differences between the two graphics conditions for all three measures (queasiness, dizziness, and disorientation). In every single instance, participants rated their queasiness, dizziness, and disorientation symptoms identically at baseline as after their first exposure. This means that not only were there no differences between conditions, there were also no instances of any symptoms after either condition.

These findings are somewhat unexpected due to many previous reports of adverse effects from VR. There are several factors that could explain these findings, or lack thereof. The fairly homogenous group of participants in terms of predisposition to motion sickness, age, gender (to a lesser extent), and prior experience with VR could account for these findings. The best predictor of one's susceptibility to cyber sickness is their history of motion sickness in another environment (Golding, 2006). In the present

research study, only 10% (n = 4) of participants indicated that they are prone to motion sickness therefore the predisposition to cyber sickness in this sample was very low.

Age is also a predictor of cyber sickness vulnerability, with older individuals being more susceptible to it than younger individuals (Brooks, et al., 2010). The participants in our study ranged in age from 18 to 32 with a mean age of 21.32 (*SD = 3.72*) years age of age. The low age of participants may have contributed to a lack of any cyber sickness symptoms. It is possible that cyber sickness may have been measured had the age of participants been greater, with more older participants than the current sample.

Additionally, gender has been shown to affect one's susceptibility to cyber sickness in VR, with females being three times more susceptible than males (Stanney, Kennedy, & Hale, 2014). Our study had 67.5% males and 32.5% females which may have decreased the likelihood of detecting any cyber sickness symptoms.

However, novice users of VR almost always experience more symptoms than experienced ones (Lackner, 2014), and almost every participant self-identified as a novice user (55% user VR once or twice, 42.5% never used VR before) while only one participant self-identified as a regular user of VR. Surprisingly, there were no incidents of cyber sickness after first exposure to VR, even though almost all participants were novice VR users. Task factors, such as duration of exposure may have mitigated against the symptoms of cyber sickness.

The time of VR exposure was quite low, 15 minutes, and the duration of exposure is directly related to cyber sickness susceptibility (Stanney, Kennedy, & Drexler, 1997). The lack of symptoms may have been a direct measure of the short exposure time. Unsurprisingly, when considering only the second condition the participant is exposed to (i.e., either BADS or VR, after a first condition in the opposite one), there were significant results between queasiness ratings after second exposure, between the VR and BADS. Participants who were exposed to VR second experienced more queasiness. This may indicate that being exposed to VR after BADS could exacerbate symptoms, as the total exposure time is about 30 minutes by the end of the second condition. On the other hand, participants who flew in the BADS condition second did not experience an increase in symptoms the way participants in VR second did, suggesting that BADS may have less of an effect on cyber sickness than VR with longer exposure time.

Comfort of HMD increased by end of flying experience. HMD considered relatively light.

Participants generally found the Oculus Rift HMD to be comfortable, and gradually considered it more comfortable near the end of their VR flying experience. Participants rated the perceived 'heaviness' of wearing the Oculus Rift and considered it relatively light, which is unsurprising since the HMD we used for this study, the Oculus Rift, weighs only 470 grams (www.oculus.com). These findings contribute to a more positive user experience, as uncomfortable, improper fitting, and heavy HMDs can lead to headaches and slippage, which causes scene motion, especially prevalent in those wearing eyeglasses (Jerald, 2016a).

Visual scene in VR rated crisp and real.

Participants rated the visual scene in the Oculus Rift in terms of its crispness and the realness, with responses slightly skewed towards the *Very Crisp* and *Very Real* ends, respectively (versus *Not at all Crisp*, and *Not at all Real*, respectively). While we did not measure presence in the VR environment directly, participants commonly cited feeling 'immersed' in the VR environment either during or after

their flight. Specifically, they noted how immersed they felt in the simulated environment when they could turn their head around and still see the aircraft. Some participants indicated they felt 'enveloped' in the environment. This is in contrast to the BADS, where there was a clear distinction between the real-world and the simulated world, when participants looked out the aircraft windows.

This is in line with other research about immersiveness and presence. Geszten et al. (2015) studied the user experience in a 3D virtual environment, and from 117 interviews, qualitatively analysed the data to determine factors that contributed to and weakened the participants' feelings of presence. They found that getting involved in the task, solving the task, and communicating were the three factors that increased presence. In our research study, participants were directly involved in the task (i.e., they were the pilot of the aircraft), they generally 'solved' the task (i.e., flew through the holographic hoops), and communicated throughout the task (i.e., by making radio calls throughout their flight and listening to other aircraft chatter). Similarly, Sutcliffe and Alrayes (2012) found that feeling involved, sensing movement, and a large field of view increased the feelings of presence, whereas interactions that felt unnatural, and situations where they had minimal involvement tended to decrease presence.

Moreover, technical problems and an unrealistic environment can weaken presence (Geszten, et al., 2015). The airfield in our present research study, while named after a fictitious place (i.e., Pendleton Airfield) was actually a simulation of the Hong Kong airspace and highly realistic. Additionality, we encountered minimal technical problems that would have caused a break a presence. Most technical problems were mitigated before beginning the VR condition, such as calibrating the HMD visuals, to ensure they were crisp for the user.

Most participants were content not being able to see their hands on the yoke, however participants were slightly bothered by not being able to see their hands reaching for the throttle/flaps. This can likely be attributed to the fact that participants had to reach for the throttle/flaps many times throughout the flight, whereas at least one hand generally stayed on the yoke throughout the flight. The yoke was also larger therefore likely easier to locate than the throttle/flaps. These factors of the simulation may have caused some frustration and perhaps a break in presence if the user had difficulty locating the position of the throttle/flaps. In the BADS condition, this was not an issue as the user could see their hands as they reached for the throttle and flaps. Although participants subjectively expressed this as a concern (i.e., in the questionnaire), participants first completed a VR practice circuit, where we encouraged them to practice reaching for the throttle and flaps. By the end of the practice circuit, most participants located the position of the throttle/flats almost immediately when reaching for it. It is possible that they attributed this practice round as part of the experimental study when answering the questionnaire.

While in future similar studies it would be useful to measure presence (e.g., with a dedicated questionnaire), it was out of scope of the current work due to the high amount of other data being gathered. Ultimately, the User Experience in VR was positive and similar to the BADS condition, although some adverse effects may have been mitigated due to the homogenous participant group. Longer flight times and a more heterogenous group of participants would be the next step to validate this aspect of VR as a viable flight training alternative to BADS.

5.2.2. Cognitive Load

Q2. Are there differences in the cognitive load between the two graphics conditions?

Overall, we found that the cognitive load may have been higher in the VR graphics condition than in the BADS. We estimated the cognitive load in the two flight conditions by measuring the participants' subjective mental workload rating, performance on a secondary task, and physiologically through heart rate and GSR response. We find mixed results, with the subjective rating, heart rate, and GSR peaks indicating a higher workload in the VR condition, while the GSR amplitude sum and secondary task performance suggest there are no differences between the two conditions.

Higher subjective rating of mental workload in the VR graphics condition.

Participants described their average mental workload as significantly higher in the VR graphics condition than in the BADS, and these results were not moderated by age, gamer experience, gender, nor were any order effects noted. These findings are in line with the increased heart rate and number of SCR peaks recorded, but in contrast to the secondary task performance measure and SCR peak amplitude.

It is possible that participants indicated they were more mentally taxed in VR than in BADS due to the greater sense of presence in VR. Peng (2008) suggests that replacing physical reality with VR can create a more lifelike experience and incite more realistic reactions from participants. The participant's senses can be flooded in VR, and although BADS can also captivate a participant's attention, it does not create a barrier between the physical and virtual world or eliminate surrounding stimuli. These factors could lead participants to rate their mental effort in the VR environment as higher than in the BADS graphics condition even while their performance remained the same.

Discrepancies between the subjective workload rating, the GSR amplitude sum and secondary task results could also be a result of the participant's own perception of the meaning of the term 'mental effort' that was used in the questionnaire. Their impressions of 'mental effort' can be influenced by the quality of their own performance (Casner & Gore, 2010). Although we did not note significant performance differences between the two graphics conditions, the novelty of VR and the immersiveness of the visual condition (360 degrees) compared to the BADS, may have contributed to a perceived mental effort that was higher in VR. Additionally, although we didn't register any performance decrements in VR compared to BADS, heart rate and GSR peaks (i.e., physiological indicators of increased cognitive load) were significantly higher in the VR condition, suggesting that participants were being more cognitively loaded.

We did not use one of the standard aviation cognitive load questionnaires in this study, such as the Cooper-Harper scale (Cooper & Harper, 1969), the SWAT (Reid & Nygren, 1988), or the NASA-TLX (Hart & Staveland, 1988) for various reasons. The Cooper-Harper Scale has a large focus on aircraft handling and since the participants were not trained pilots (only two had a pilot's license), its terminology appeared overly specific to the aviation domain to gather the data we sought. The SWAT and NASA-TLX, while diagnostically adequate, are time consuming in preparatory tasks (e.g., card sorting task in SWAT, training participants on the scale for NASA-TLX). As such, we leveraged the Instantaneous Self-Assessment (ISA) (Jordan & Brennen, 1992) but in a modified form where it was only administered after each flying condition. Due to the numerous tasks involved in each flying condition (i.e., PDT, making radio calls, listening to other aircraft chatter, flying through holographic hoops), administering the ISA in real-time would have likely led to performance decrements on the primary task, which has previously been shown (Tattersall & Foord, 1996). Ultimately, the ISA after each flying condition was likely the least disruptive approach.

However, a disadvantage of administering the questionnaire after the task is that we relied on the user to rate their experience retrospectively. Users may have forgotten important aspects of the flying tasks, each of which were about 15 minutes in length. The most recent events to the time the questionnaire was administered can greatly impact the rating (Wilhelm & Grossman, 2010). Although the subjective rating was taken post-hoc (i.e., after the experience) and it was static (i.e., at a single point in time), its results can be corroborated with the heart rate and GSR measures which were recorded continuously, in real-time to provide a full picture of the user's cognitive load.

No difference in secondary task performance.

Both the hit rate and response time from the PDT results suggest no difference in the cognitive load, indicating that although participants were overall more stimulated in the VR graphics condition (as measured by the subjective rating, heart rate and GSR), it was insufficient to necessitate task shedding. This is in line with the primary task performance metrics that also show no performance differences (e.g., in heading RMSE, altitude RMSE, and precision flying through the hoops).

Higher heart rate **response in the VR graphics condition.**

Participants experienced a significantly higher heart rate response during the VR graphics condition than the BADS, suggesting increased cognitive load in VR. Neither age, gender, nor gamer experience was significantly associated with heart rate in either graphics condition and no order effects were noted.

This finding supports past research that found that the more sophisticated and realistic the flight simulator, the more likely a pilot is to have a significant increase in heart rate (Roscoe, 1992). In this sense, a VR environment could be considered more sophisticated and realistic than the BADS, contributing to the higher heart rate response. The increased field of view, stimuli, and feelings of immersiveness could all contribute to the higher heart rate, as there were no primary or secondary task performance differences. Heart rate measurements may represent indices of the underlying processes involved in responses, unlike performance outcomes alone.

Our findings also agree with Allsop et al. (2016) who measured heart rate in non-pilots to be directly related to self-reported state-anxiety (due to a high cognitive load condition) in a flight simulator. Heart rate and subjective ratings were directly related in our current research, where both indicated an increased workload in the VR versus BADS condition.

Since heart rate in a flight simulator and VR flight condition has not been compared prior to this current study (to our knowledge), we cannot directly comment on our findings relative to such similar work. However, heart rate has been measured in VR and compared to real-life tasks. For example, Kothgassner et al. (2016) measured heart rate (amongst other physiological markers) in a public speaking task with a real-live audience and with a VR audience. They found heart rate response was similar in both conditions. Perhaps our results would have been different if we compared flying in VR with live flight, however our results in the BADS simulator are not in agreement with their findings indicating that there may also be differences with heart rate response in live and simulated flights.

Some researchers suggest that heart rate (or other continuously measured physiological indicators) allows for a sensitive detection of changes in cognitive load even when no task performance deterioration occurs (Ferreira, et al., 2014). Such was the case in our research study, as we measured an increased heart rate in the VR graphics conditions, without any secondary task performance (i.e., PDT

response time and hit rate) or primary task performance decrements. Using heart rate to measure cognitive load, (or other physiological methods to continuously measured cognitive load), is supported for tasks that leverage major cognitive processes such as perception, memory, or reasoning (Beatty, 1982).

Higher number of SCR peaks in the VR graphics condition, no difference in amplitude sum.

The VR graphics condition elicited a greater amount of GSR peaks than the BADS, however there were no differences in the peak GSR amplitude sum between the two graphics conditions. The number of GSR peaks and the average GSR peak amplitude are both directly related to an individual's arousal level (iMotions, 2016), therefore it is surprising that the two methods did not produce consistent results.

A possible reason for the inconsistent GSR results is that within a technique, sensitivity may vary with the selected measure (Lysaght, et al., 1989). Kroese and Siddle (1983) found that the GSR amplitude tends to decline, or habituate, with repeated presentation of stimuli. They measured the GSR response of participants in a visual monitoring task, where stimuli were presented along with task irrelevant tones. They carried out a second experiment where the pitch of the task irrelevant tones changed, and again found the GSR amplitude habituated, although at a slower rate than in the first experiment.

However, Boucsein, Haarmann, and Schaefer (2008) found that both the SCR amplitude and the number of peaks decreased in the progression of four flight simulator missions (per participant) indicating habituation. In a future study, they decided to use GSR frequency (and not amplitude) as an indicator of cognitive workload in pilots for an adaptive automation task during simulated flight (Haarmann, Boucsein, & Schaefer, 2009), perhaps as they consider it more sensitive than amplitude for detecting changes in cognitive load.

In our research study, these results suggest that participants became habituated to the task demands required while flying in both the VR or BADS, however a greater number of SCR peaks in VR suggests that the VR condition evoked a higher cognitive load for the duration of the flying task. This result supports those obtained through measuring heart rate and the subjective workload rating, suggesting that although there were no differences in secondary task performance, cognitive load was in fact higher in the VR condition than the BADS.

The Empatica E4 as a valid measure of heart rate **and GSR.**

To substantiate the physiological measures, we gathered in our current research, we must also consider the validity of the device used to gather the data. The Empatica E4, although a medical grade device, is still in its infancy in terms of being validated. However, due to limitations in the flight environment (i.e., hands on the yoke/throttle/flaps, PDT device on fingertip), there was merit in using a non-obtrusive, non-invasive device, with minimal setup. The wristwatch design of the Empatica E4 made it ideal for this environment.

Other studies that measure GSR have used skin conductance sensors placed on the distal phalanges of two fingers by having the user rest his/her hand over top of a device with built in sensors (Zhai & Barreto, 2006); by attaching two electrodes to the fingertips (Kroese & Siddle, 1983; Villarejo, Zapirain, & Zorrilla, 2012); by attaching two electrodes to the inside of the right foot (Reinhardt, Schmahl, Wust, & Bohus, 2012); by placing electrodes on the wrist (Hernandez, Morris, & Picard, 2011); or by placing sensors into an armband that is worn on the back of the upper arm (Perala & Sterling, 2007). Although

some of these approaches may have been feasible in the simulator environment, the Empatica E4 was chosen for its ease of use, as it only required strapping on a wristwatch.

Ollander et al. (2016) compared the heart rate and GSR signals from the Empatica E4 with a stationary ECG and finger skin conductivity electrodes of high sampling rates, during a classical laboratory stress protocol (i.e., the Trier Social Stress Test) with seven participants. Compared to the ECG, they found that the time-domain features, such as the mean heart rate and standard deviation were well estimated, with good stress discrimination power, however the E4 had a significant loss of detected interbeat intervals. This suggests that using the E4 to compare mean heart rates (as in our current research) is supported, however it would not be recommended to use the E4 for heart rate variability tests.

In terms of skin conductance, they found that although the signals measured at the two different locations (i.e., wrist and fingers) had no visual resemblance, the signal from the Empatica E4 yielded higher stress discrimination power than the ones measured at the fingers. This provides a promising result for using the Empatica E4 for measuring heart rate and GSR as an indicator of cognitive load.

5.2.3. Performance Metrics

Q3. Are there differences in performance between the two graphics conditions?

No difference in flight path deviations RMSE, and flying precision (through hoops). Greater airspeed RMSE in BADS than VR.

We measured no differences in performance, as determined by RMSE in flight path deviations (i.e., altitude and heading) between the BADS and VR graphics conditions. There were also no differences in flying precision, as measured by the hoop flying task. We did however measure a difference in the airspeed RMSE, where participants showed a greater deviation from the expected measure in the BADS condition versus the VR.

Possibly reasons for this unexpected result are that there was an emphasis placed on the speed indicator in the VR condition while calibrating the visuals. Participants were asked if they could clearly see the speedometer in the VR condition and reminded that they should fly at 100 knots during the downwind segment. In the BADS condition, there was no emphasis placed on the speedometer as the visuals did not need to be calibrated (it was an analog indicator). This extra emphasis in the VR condition may have biased participants into being more careful and flying at the recommended speed, compared to the BADS condition.

The fact that we did not observe any flight performance differences, although there was a higher cognitive load in the VR condition (as indicated through heart rate, GSR, and subjective ratings), we must consider possible reasons. One such explanation is the types of performance measures we assessed. RMSE measures in this context are a Type 1 primary performance measure, therefore they represent a combination of the participant and system performance (Lysaght, et al., 1989). The fact that there were no differences does not allow the conclusion that the level of effort required by the participant was identical. In fact, Type 2 primary task performance measures, such as the amount of correction required on the yoke, throttle, and flaps to achieve this level of (Type 1) performance would provide an indication of the workload (Hart S. G., 1986). While we did not collect Type 2 primary task performance measures, the increased heart rate, GSR, and subjective mental workload ratings indicate that perhaps if we did, we may have observed lower performance in VR.

The research literature, at this time, has no study comparing the performance metrics between a flight simulator and a VR flying task. However, some studies have compared the use of VR with traditional simulators in other contexts, or VR against live training.

Stevens and Kincaid (2015) compared aerial door gunnery in either a VR graphics condition (using an HMD) or a flat screen display simulator. They noted that participants had better performance (and presence) in the more immersive visual display (i.e., the VR environment). The discrepancy between our findings and theirs could be a result of the technology used – their flat screen simulator was far less immersive than the BADS and they attribute the increased performance as a result of the increased immersiveness in the VR condition.

A team of Australian Army parachute jumpers received advanced jump training either through a VR simulator, or in a traditional classroom setting, where training length was similar in both conditions (Butavicius, Vozzo, Braithwaite, & Galanis, 2012). They found no differences in performance in subsequent live jumps in terms of jump accuracy, jump safety, or confidence. Similarly, a team of researchers studied the use of VR for surgical simulation training (Seymour, et al., 2002). In comparison to the control group that received traditional non-VR training only, the VR group significantly improved in performance during laparoscopic cholecystectomy (i.e., surgically removing the gallbladder).

These three studies taken together (Butavicius, Vozzo, Braithwaite, & Galanis, 2012; Seymour, et al., 2002; Stevens & Kincaid, 2015) suggest that at worst, VR training should match the performance that traditional training methods generate, which is in line with the findings of our current study, where we noted no performance differences between the two graphics conditions. This result is promising when validating VR as an alternative to traditional simulator training.

Negative symptoms in the VR environment could impact performance in VR.

By grouping participants as either those who had adverse symptoms in VR (in terms of queasiness, dizziness, or disorientation) and those who did not, several one-way ANOVAs were conducted to determine if symptoms in VR could impact performance in VR.

There was a statistically significant difference between queasiness groups on heading RMSE performance, and between dizziness groups on heading RMSE performance. Participants who experienced queasiness or dizziness had greater heading RMSE (therefore worse performance) than those participants who did not report experiencing queasiness or dizziness, respectively. These findings suggest that there is a need to develop screening methods to identify individuals susceptible to cybersickness in VR, without inducing them to experience sickness. Certain individuals may not be suitable candidates for VR flight training as it may impact their performance thereby defeating the actual benefit of training.

5.3. Takeaways and Recommendations

This research study took steps toward validating VR as a pilot training alternative to traditional flight training in simulators. The present research leveraged the interaction-centered framework of experience, put forth by Forlizzi and Battarbee (2004), and subsequently focused on measuring the user

experience and the cognitive load on users, and not just outcomes (e.g., performance) by taking an objective perspective to evaluate a user's experience and interaction.

To understand the *fluent* aspects of the VR flying experience, the present work measured physiological indices of cognitive load enabling capturing the experience of flying a flight simulator without disruption to the task. To understand *cognitive* and *expressive* experiences, the interactions were captured both as they unfolded and any expression after the fact. This was achieved by evaluating performance metrics during the experience and evaluating any negative effects from the immersion environment after the task.

One might measure only performance and assume that if there are no differences, one modality can be substituted for the other and be just as effective as a training tool. In fact, the main takeaway of this research suggests that although the user experience and performance metrics were comparable (apart from differences in maintaining airspeed), the VR experience likely causes an increased cognitive load on users compared to the BADS. Furthermore, adverse symptoms in VR were associated with decreased performance in VR, suggesting that individual susceptibility to VR may preclude certain individuals from leveraging it as a flight training alternative. Thereby, completely replacing traditional flight training in a BADS environment with VR could have undesirable side-effects.

At this time, it is suggested that VR could be slowly incorporated as an additional tool in flight training, although it is premature to conclude it as a viable replacement. The many advantages of VR (e.g., low cost, easy upgrades, environmentally friendly, smaller space requirements, etc.) outweigh its disadvantages (e.g., possible cyber sickness, possible increased cognitive load), suggesting those interested in flight training could consider it a viable, at home alternative. Users should be cognizant of possible side effects and monitor their own well-being to determine if they should continue or discontinue use. Users should be particularly aware of possible queasiness from prolonged exposure or increased cognitive load. Screening tools to predict cybersickness symptoms in VR are critical, as those who have symptoms may have decreased performance.

5.4.Limitations and Future Work

The current research has many promising results and is the first research to compare the user experience, cognitive load, and performance metrics between BADS and VR graphics conditions. However, results should be interpreted with the limitations in mind.

The current research did not compare VR to a full-flight simulator with a motion platform. In fact, the research is limited to comparing the visuals in the two conditions only, since participants remained in the cockpit of the Cessna aircraft for both conditions. This could have important implications for using VR without the tactile feedback that our current participants enjoyed (e.g., hands on a responsive yoke, throttle, and flaps). VR flight training without this tactile feedback is a worthwhile area of future research. Creating this tactile feedback for at home VR use is another viable alternative.

The fact that some of the results are contradictory, suggests further analysis and clarification is required. For example, certain measures of cognitive load suggested cognitive load was higher in VR (i.e., subjective rating, heart rate, and number of SCR peaks) while others suggested there was no difference (i.e., SCR amplitude sum, PDT). A larger and more heterogenous participant sample (e.g., age, gender, experience) could provide more insight into these findings. Similarly, these various user groups could

also provide more context to the lack of cyber sickness noted after first exposure, which is in contradiction to many studies.

In our current work, the concept of *presence* in VR was inferred, but not measured. Future work should consider presence, as it may be an important contributor to the user experience. Lastly, the current work did not measure the actual training benefit of VR, as we did not compare the transfer of training into the real flying. This is a subsequent area worth exploring.

6. Bibliography

Ahmadi, K., & Alireza, K. (2007). Stress and job satisfaction among Air Force military pilots. *Journal of Social Sciences, 3*(3), 159-163.

Ahmaniemi, T., Lindholm, H., Muller, K., & Taipalus, T. (2017). Virtual reality experience as a stress recovery solution in workplace. *2017 IEEE Life Sciences Conference* (pp. 206-209). Sydney, Australia: IEEE.

Akiduki, H., Nishiike, S., Watanabe, H., Matsuoka, K., Kubo, T., & Takeda, N. (2003). Visual-vestibular conflict induced by virtual reality. *Neuroscience Letters*, 197-200.

Allsop, J., Gray, R., Bulthoff, H., & Chuang, L. (2016). Effects of anxiety and cognitive load on instrument scanning behavior in a flight simulation. *2016 IEEE Second Workshop on Eye Tracking and Visualization (ETVIS)* (pp. 55-59). Baltimore, MD: IEEE.

Baddeley, A. D. (1992). Working memory. *Science*, 556-559.

Baka, E., Stavroulia, K. E., Magnenat-Thalmann, N., & Lanitis, A. (2018). An EEG-based evaluation for comparing the sense of presence between virtual and physical environments. *CGI 2018: Proceedings of Computer Graphics International 2018* (pp. 107-116). Bintan Island, Indonesia: ACM.

Bakker, J., Pechenizkiy, M., & Sidorova, N. (2011). What's your current stress level? Detection of stress patterns from GSR sensor data. *11th IEEE International Conference on Data Mining Workshops* (pp. 573-580). IEEE.

Barfield, W., Zeltzer, D., Sheridan, T., & Slater, M. (1995). Presence and performance within virtual environments. In W. Barfield, & R. A. Furness, *Virtual environments and advanced interface Design.* New York City: Oxford University Press.

Basri, N. H., Noor, N. L., Adnan, W. A., Saman, F. M., & Baharin, A. H. (2016). Conceptualizing and understanding user experience. *4th International Conference on User Science and Engineering* (pp. 81-84). IEEE.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychology Bulletin*, 276-292.

Benedek, M. &. (2010). Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology*, 647-658.

Boeing. (2015). *Statistical Summary of Commercial Jet Airplane Accidents - Worldwide Operations 1959-2015.* Boeing.

Boucsein, W., Haarmann, A., & Schaefer, F. (2008). The usability of cardiovascular and electrodermal measures for adaptive automation during a simulated IFR flight mission. In J. Westerink, M. Ouwerkerk, F. Pasveer, & B. d. Ruyter, *Probing Experience: From Assessment of User Emotions and Behaviour to Development of Products* (pp. 235-243). Dordrecht: Springer.

Brooks, J. O., Goodenough, R. R., Crisler, M. C., Klein, N. D., Alley, R. L., Koon, B. L., . . . Wills, R. F. (2010). Simulator sickness during driving simulation studies. *Accident Analysis and Prevention*, 788-796.

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 53-61.

Bruyas, M.-P., & Dumont, L. (2013). Sensitivity of detection response task (DRT) to the driving demand and task difficulty. *Proceedingss of the Seventh International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, (pp. 64-70). Bolton Landing, New York.

Butavicius, M. A., Vozzo, A., Braithwaite, H., & Galanis, G. (2012). Evaluation of a virtual reality parachute training simulator: assessing learning in an off-course augemented feedback training schedule. *The International Journal of Aviation Psychology, 22*(3), 282-298.

Casner, S. M., & Gore, B. F. (2010). *Measuring and evaluating workload: a primer.* Moffett Field, California: NASA.

Causse, M., Dehais, F., Arexis, M., & Pastor, J. (2011). Cognitive aging and flight performances in general aviation pilots. *Aging, Neuropsychology, and Cognition, 18*(5), 544-561.

Chao, C.-J., Wu, S.-Y., Yau, Y.-J., Feng, W.-Y., & Tseng, F.-Y. (2017). Effects of three-dimensional virtual reality and traditional training methods on mental workload and training performance. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 187-196.

Chen, F., Ruiz, N., Choi, E., Epps, J., Khawaja, M. A., Taib, R., . . . Wang, Y. (2012). Multimodal behavior and interaction as indicators of cognitive load. *ACM Transactions on Interactive Intelligent Systems* (pp. 22:1-22:35). New York City: ACM.

Cho, D., Ham, J., Oh, J., Park, J., Kim, S., Lee, N.-K., & Lee, B. (2017). Detection of stress levels from biosignals measured in virtual reality environments using a kernel-based extreme learning machine. *Sensors*, 1-18.

Chung, K. H., Shewchuk, J. P., & Williges, R. C. (2002). An analysis framework for applying virtual environment technology to manufacturing tasks. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 335-348.

Cobb, S. V., Nichols, S. C., Ramsey, A. D., & Wilson, J. R. (1999). Virtual reality induced symptoms and effects. *Presence: Teleoperators and Virtual Environments*, 169-186.

Conti, A. S., Dlugosch, C., & Bengler, K. (2014). The effect of task set instruction on detection response task performance. *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2013 Conference* (pp. 107-117). HFES.

Cooper, G. E., & Harper, R. J. (1969). *The use of pilot rating in the evaluation of airscraft handling qualities.* Washington: National Aeronautics and Space Administration (NASA).

Costello, P. (1997). *Health and safety issues associated with virtual reality: a review of current literature .* Loughborough University: JISC Advisory Group on Computer Graphics (AGOCG) Technical Report Series No. 37.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, 114-185.

Cowings, P. S., Suter, S., Toscano, W. B., & Kamiya, J. (1986). General autonomic components of motion sickness. *Psychophysiology*, 542-551.

Critchley, E. (2002). Electrodermal responses: what happens in the brain. *Neuroscientist*, 132-142.

Cunningham, D. W., Billock, V. A., & Tsou, B. H. (2001). Sensorimotor adaptation to violations in temporal contiguity. *Psychological Science*, 532-535.

Cunningham, D. W., Chatziastros, A., Heyde, M. V., & Bulthoff, H. H. (2001). Driving in the future: temporal visuomotor adaptation and generalization. *Journal of Vision*, 88-98.

Davis, S., Nesbitt, K., & Nalivaiko, E. (2014). A systematic review of cybersickness. *Proceedings of the 2014 Conference on Interactive Entertainment* (pp. 1-9). Newcastle, Australia : ACM.

de Waard, D. (1996). *The measurement of drivers' mental workload.* PhD thesis, University of Groningen.

Debijadji, R., Perovic, L., Nagulic, S., & Djurakic, D. (1973). Psychological reactions of pilots in super-sonic aircraft in flight. *Revue de Medecine Aeronautique et Spatiale*, 367-371.

Duarte, E., Rebelo, F., & Wogalter, M. S. (2010). Virtual reality and its potential for evaluating warning compliance. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 526-537.

Empatica. (2016, March 31). *Utilizing the PPG.BVP Signal*. Retrieved from Empatica : https://support.empatica.com/hc/en-us/articles/204954639-Utilizing-the-PPG-BVP-signal

Farra, S. L., Miller, E. T., & Hodgson, E. (2015). Virtual reality disaster training: translation to practice. *Nurse Education in Practice, 15*(1), 53-57.

Federal Aviation Administration. (2002). *Private pilot practical test standards for airplane (FAA Pub. No. FAA-S-8081-14A).* FAA.

Ferreira, E., Ferreira, D., Kim, S., Siirtola, P., Roning, J., Forlizzi, J. F., & Dey, A. K. (2014). Assessing real-time cognitive load based on psychol-physiological measures for younger and older adults. *2014 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain* (pp. 39-48). IEEE.

Forlizzi, J., & Battarbee, K. (2004). Understanding experience in interactive systems. *DIS '04: Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques* (pp. 261-268). Cambridge, MA: ACM.

Fuhua, L., Duffy, V. G., & Su, C. J. (2002). Developing virtual environments for industrial training. *Information Sciences*, 153-170.

Gaetan, S., Dousset, E., Marqueste, T., Bringoux, L., Bourdin, C., Vercher, J.-L., & Besson, P. (2015). Cognitive workload and psychophysiological parameters during multitask activity in helicopter pilots. *Aerospace, Medicine, and Human Performance*, 1-6.

Garbarino, M., Lai, M., Bender, D., Picard, R. W., & Tognetti, S. (2014). Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. *EAI 4th International Conference on Wireless Mobile Communication and Healthcare (Mobihealth)*, (pp. 39-42). Cambridge, MA.

Gemelli, A. (1917). *On the application of psycho-physical methods of examining candidates as military pilots.* Milan: Vita ePensiero.

Geszten, D., Hamornik, B. P., Komlodi, A., Hercegfi, K., Szabo, B., & Young, A. (2015). Qualitative analysis of user experience in a 3D virtual environment. *ASIST '15: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community* (pp. 1-4). St. Louis, MO: American Society for Information Science.

Golding, J. F. (2006). Motion sickness susceptibility . *Autonomic Neuroscience*, 67-76.

Gopher, D., & Braune, R. (1984). On the psychophysics of workload: Why bother with subjective measures? *Human Factors*, 519-532.

Grantcharov, T. P., Kristiansen, V. B., Bendix, J., Bardram, L., Rosenberg, J., & Funch-Jensen, P. (2004). Randomized clinical trial of virtual reality simulation for laparoscopic skills training. *British Journal of Surgery*, 146-150.

Haarmann, A., Boucsein, W., & Schaefer, F. (2009). Combining electrodermal responses and cardiovascular measures for probing adaptive automation during simulated flight. *Applied Ergonomics*, 1026-1040.

Harm, D. L. (2002). Motion Sickness Neurophysiology, Physiology Correlates, and Treatment. In K. M. Stanney, *Handbook of Virtual Environments* (pp. 637-661). N. J. : Lawrence Erlbaum Associates.

Hart, S. G. (1986). Theory and measurement of human workload. In J. Zeldner, *Human Productivity Enhancement* (pp. 398-455). New York City: Praeger.

Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX): 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, (pp. 904-908).

Hart, S. G., & Hauser, J. R. (1988). Inflight application of three pilot workload measurement techniques. *Aviation, Space, and Environmental Medicine*, 511-516.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, 139-183.

Hayes, R. T., Jacobs, J. W., Prince, C., & Salas, E. (1992). Flight simulator training effectiveness: A meta-analysis. *Military Psychology*, 63-74.

Henriques, R., Paiva, A., & Antunes, C. (2013). Accessing emotion patterns from affective interactions using electrodermal activity. *Humaine Assocation Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 43-48). IEEE.

Hernandez, J., Morris, R. R., & Picard, R. W. (2011). Call center stress recognition with person-specific models. *Affective Computing and Intelligent Interaction*, 125-134.

Hill, K. J., & Howarth, P. A. (2000). Habituation to the side effects of immersion in a virtual environment. *Displays*, 25-30.

Holloway, R. (1997). Registration error analysis for augmented reality. *Presence: Teleoperators and Virtual Environments*, 413-432.

Howitt, J. (1969). Flight deck workload studies in civil air transport aircraft. *Conference Proceedings (No. 56) Measurement of Aircrew Performance.* Paris: AGARD.

Hubbard, D. (1987). Inadequacy of root mean square error as a performance measure. *Proceedings of the 4th International Symposium on Aviation Psychology* (pp. 698-704). Columbus, OH: Ohio State University.

ICAO. (2009). *Doc 9625, Manual of Criteria for the Qualification of Flight Sumulation Training Devices - Volume I - Aeroplanes.* Montreal : International Civil Aviation Organization.

iMotions. (2016). *GSR Pocket Guide.* iMotions Biometric Research Platform.

International Society for Presence Research. (2000). *The Concept of Presence: Explication Statement.* Retrieved from https://smcsites.com/ispr/

Jang, D. P., Kim, I. Y., Nam, S. W., Wiederhold, B. K., Wiederhold, M. D., & Kim, S. I. (2002). Analysis of physiological response to two virtual environments: driving and flying simulation. *Cyber Psychology & Behavior*, 11-18.

Jerald, J. (2016a). Adverse Health Effects. In J. Jerald, *The VR Book - Human-Centered Design for Virtual Reality* (p. 159). New York City: ACM Books.

Jerald, J. (2016b). Hardware Challenges. In J. Jerald, *The VR Book: Human-Centered Design for Virtual Reality* (pp. 177-181). New York City: ACM Books.

Jerald, J. (2016c). Latency. In J. Jerald, *The VR Book: Human Centered Design for Virtual Reality* (pp. 183-194). New York City: ACM Books.

Jerald, J. (2016d). Presence. In J. Jerald, *The VR Book: Human-Centered Design for Virtual Reality* (p. 46). New York City: ACM Books.

Jerald, J. (2016e). Summary of Factors that Contribute to Adverse Effects. In J. Jerald, *The VR Book: Human-Centered Design for Virtual Reality.* New York City: AMC Books.

Jerald, J., & Whitton, M. (2009). Relating scene-motion thresholds to latency thresholds for head-mounted displays. *IEEE Virtual Reality* (pp. 211-218). Lafayette: IEEE.

Johnson, D. M. (2005). *Introduction to and review of simulator sickness research.* Arlington, Virginia: US Army Research Institute.

Jordan, C. S., & Brennen, S. D. (1992). *Instantaneous self-assessment of workload technique.* Portsmouth: Defence Research Agency.

Kantowitz, B. H., Hart, S. G., & Bortolussi, M. R. (1983). Measuring pilot workload in a moving-base simulator: I. Asynchronous secondary choice-reaction task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 319-322). HFES.

Kellogg, R. S., Kennedy, R. S., & Graybiel, A. (1965). Motion sickness symptomatology of labyrinthine defective and normal subjects during zero gravity maneuvers. *Aerospace Medicine Research Labs*, 315-318.

Kennedy, R. S., & Fowlkes, J. E. (1992). Simulator sickness is polygenic and polysymptomatic: implications for research. *The International Journal of Aviation Psychology*, 23-38.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *The International Journal of Aviation Psychology*, 203-220.

Keshavarz, B., Hecht, H., & Lawson, B. D. (2014). Visually induced motion sickness: Causes, characteristics, and countermeasures. In *Handbook of Virtual Environment: Design, Implementation, and Applications, 2nd ed.* (pp. 648-681). Boca Raton, FL: CRC Press.

Kikhia, B., Stavropoulos, T. G., Andreadis, S., Karvonen, N., Kompatsiaris, I., Sävenstedt, S., . . . Melander, C. (2016). Utilizing a wristband sensor to measure the stress level for people with dementia. *Sensors*, 1-17.

Kim, Y. Y., Kim, H. J., Kim, E. N., Ko, H. D., & Kim, H. T. (2005). Characteristic changes in the physiological components of cybersickness. *Psychophysiology*, 616-625.

Kinciad, J. P., & Westerlund, K. K. (2009). Simulation in Education and Training. *Proceedings of the 2009 Winter Simulation Conference* (pp. 273-280). IEEE.

Knight, M. M., & Arns, L. L. (2006). The relationship among age and other factors on incidence of cybersickness in immersive environment users. *APGV '06: Proceedings of the 3rd symposium on applied perception in graphics and visualization* (p. 162). Boston: ACM.

Kothgassner, O. D., Felnhofer, A., Hlavacs, H., Beutl, L., Palme, R., Kryspin-Exner, I., & Glenk, L. M. (2016). Salivary cortisol and cardiovascular reactivity to a public speaking task in a virtual and real-life environment. *Computers in Human Behavior, 62*(C), 124-135.

Kramer, A. F. (1991). Physiological metrics of mental workload: a review of recent progress. In D. L. Ed, *Multiple Task Performance* (pp. 279-328). London: Taylor and Francis.

Kroese, B. S., & Siddle, D. A. (1983). Effects of an attention demanding task on the amplitude and habituation of the electrodermal orienting response. *Psychophysiology, 20*, 128-135.

Lackner, J. R. (2014). Motion sickness: more than nausea and vomiting. *Experimental Brain Research*, 2493-2510.

Lawson, B. D. (2014). Motion Sickness Symptomatology and Origins. In K. Hale, & K. Stanney, *Handbook of Virtual Environments (2nd ed.)* (pp. 531-600). Boca Raton: CRC Press.

Lee, Y.-H., & Liu, B.-S. (2003). Inflight workload assessment: Comparison of subjective and physiological measurements. *Aviation Space and Environmental Medicine*, 1078-1084.

Lehmann, K. S., Ritz, J. P., Maass, H., Cakmak, H. K., Kuhnapfel, U. G., Germer, C. T., . . . Buhr, H. J. (2005). A prospective randomized study to test the transfer of basic psychomotor skills from virtual reality to physical reality in a comparable training setting. *Annals of Surgery*, 442-449.

Leirer, V. O., Yesavage, J. A., & Morrow, D. G. (1989). Marijuana, aging, and task difficulty effects on pilot performance. *Aviation, Space, and Environmental Medicine, 60*(12), 1145-1152.

Lewis, C. (1967). Flight research programme IX: Medical monitoring of carrier pilots in combat II. *Aerospace Medicine*, 581-592.

Lidderdale, I. G. (1987). Measurement of aircrew workload low-level flight. In A. H. Roscoe, *The Practical Assessment of Pilot Workload.* Paris: AGARD.

Lintern, G., Roscoe, S. N., Koonce, J. M., & Segal, L. D. (1990). Transfer of landing skills in beginning flight training. *Human Factors*, 319-327.

Luximon, A., & Goonetilleke, R. S. (2001). Simplified subjective workload assessment technique. *Ergonomics*, 229-243.

Lysaght, R. J., Hill, S. G., Dick, A. O., Plamondon, B. D., Linton, P. M., Wierwille, W. W., . . . Wherry, R. J. (1989). *Operator workload: comprehensive review and evaluation of operator workload methodologies.* Alexandra, VA: US Army Research Institute for the Behavioral and Social Sciences.

Mandrick, K., Peysakhovich, V., Rémy, F., & Lepron, E. (2016). Neural and psychophysiological correlates of human performance under stress and high mental workload. *Biological Psychology*, 62-73.

Martens, M. H., & van Winsum, W. (2000). *Measuring distraction: the Peripheral Detection Task.* Soesterberg, Netherlands: TNO Human Factors.

Martins, A. P. (2016). A review of important cognitive concepts in aviation. *Aviation*, 65-83.

McClernon, C. K., & Miller, J. C. (2011). Variance as a measure of performance in an aviation context. *The International Journal of Aviation Psychology*, 397-412.

Meehan, M., Insko, B., Whitton, M., & Brooks, F. P. (2002). Physiological measures of presence in stressful virtual environments. *ACM Transactions on Graphics (TOG)* (pp. 645-652). San Antonio: ACM.

Moroney, W. F., & Moroney, B. W. (2010). Flight Simulation. In J. A. Wise, D. Hopkin, & D. J. Garland, *Handbook of Aviation Human Factors (2nd ed)* (pp. 19-4 to 19-6). Boca Ratton: CRC Press.

Muse, L., Harris, S., & Field, H. (2003). Has the inverted-U theory of stress and job performance had a fair test? *Human Performance, 16*(4), 349-364.

Nathanael, D., Mosialos, S., Vosniakos, G. C., & Tsagkas, V. (2016). Development and evaluation of a virtual reality training system based on cognitive task analysis: The case of CNC tool length offsetting. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 52-67.

Newman, R. L., & Greeley, K. W. (2016). Chapter 7: Situation Awareness - Techniques for Measuring Workload. In R. L. Newman, & K. W. Greeley, *Cockpit Displays: Test and Evaluation.* New York: Routledge.

Nichols, S. (1999). Physical ergonomics of virtual environment use. *Applied Ergonomics*, 79-90.

Nichols, S. C., Cobb, S. V., & Wilson, J. R. (1997). Health and safety implications of virtual environments: measurement issues. *Presence: Teleoperators and Virtual Environments, 6*(6), 667-675.

Ohyama, S., Nishiike, S., Watanabe, H., Matsuoka, K., Akizukia, H., Takeda, N., & Harada, T. (2007). Autonomic responses during motion sickness induced by virtual reality. *Auris Nasus Larynx*, 303-306.

Ollander, S., Godin, C., Campagne, A., & Charbonnier, S. (2016). A comparison of wearable and stationary sensors for stress detection. *2016 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 4362-4366). Budapest: IEEE.

Paas, F., Tuovinen, J. E., Tabbers, H., & Gerven, P. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychology*, 63-71.

Paillard, A. C., Quarck, G., Paolino, F., Denise, P., Paolino, M., Golding, J. F., & Ghulyan-Bedikian, V. (2013). Motion sickness susceptibility in healthy subjects and vestibular patients: effects of gender, age and trait-anxiety. *Journal of Vestibular Reseach: Equilibrium and Orienation*, 203-209.

Pausch, R., Crea, T., & Conway, M. J. (1992). A literature survey for virtual environments: military flight simulator visual systems and simulator sickness. *PRESENCE: Teleoperators and Virtual Environments*, 344-363.

Paxion, J., Galy, E., & Berthelon, C. (2014). Mental workload and driving. *Frontiers in Psychology*, 1-11.

Payne, K., & Harris, D. (2000). The development of a multi-dimensional aircraft handling qualities rating scale. *International Journal of Aviation Psychology*, 343-362.

Peng, W. (2008). The mediational role of identification in the relationship between experience mode and self-efficacy: enactive role-playing versus passive observation. *Cyber Psychology & Behavior*, 649-652.

Perala, C. H., & Sterling, B. S. (2007). *Galvanic skin response as a measure of soldier stress.* Adelphi, MD: Army Research Laboratory.

Peterson, S. M., Furuichi, E., & Ferris, D. P. (2018). Effects of virtual reality high heights exposure during beam-walking on physiological stress and cognitive loading. *PloS One*, 1-17.

Raaen, K., & Kjellmo, I. (2015). Measuring latency in Virtual Reality systems. *International Federation for Information Processing*, 457-462.

Reason, J. T., & Brand, J. J. (1975). *Motion Sickness.* London: Academic Press.

Regan, E. C., & Ramsey, A. D. (1994). *Some side-effects of immersion virtual-reality: the results of four immersions.* Farnborough, Hampshire: Army Personal Research Establishment.

Reid, G. B., & Nygren, T. E. (1988). The Subjective Workload Assessment Technique: A scaling procedure for measuring mental workload. *Advances in Psychology*, 185-218.

Reinhardt, T., Schmahl, C., Wust, S., & Bohus, M. (2012). Salivary cortisol, heart rate, electrodermal activity and subjective stress responses to the Mannheim Multicomponent Stress Test (MMST). *Psychiatry Research*, 106-111.

Riccio, G. E., & Stoffregen, T. A. (1991). An ecological theory of motion sickness and postural instability. *Ecological Psychology, 3*(3), 195-240.

Riener, A. (2010). Simulating on-the-road behavior using a driving simulator. *ACHI' 10: Third International Conference on Advances in Computer-Human Interactions, 2010* (pp. 25-31). IEEE.

Roman, J. A., Older, H., & Jones, W. L. (1967). Flight research programme VII: Medical monitoring of navy carrier pilots in combat. *Aerospace Medicine*, 133-139.

Roscoe, A. H. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration? *Biological Psychology*, 259-287.

Roscoe, A. H., & Goodman, E. A. (1973). *An investigation of heart rate changes during a flight simulator approach and landing task.* RAE Technical Memorandum Avionics.

Roscoe, A. H., & Grieve, B. S. (1988). *Assessment of pilot workload during Boeing 767 normal and abnormal operating conditions.* Warrendale, PA: Society of Automotive Engineers.

Rouder, J. N., Speckman, P. L., Sun, D., & Morey, R. D. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 225-237.

Salas, E. (2017). Using Simulation for Training. In F. Jentsch, M. Curtis, & E. Salas, *Simulation in Aviation Training* (p. 540). Routledge.

Sekiguchi, C., Hunda, Y., Gotoh, M., Kurihara, Y., Nagasawa, Y., & Kuroda, I. (1977). Continuous ECG monitoring on civil air crews during flight operations. *Aviation, Space, and Environmental Medicine* , 872-876.

Seymour, N., Gallagher, A. G., Roman, S. A., O'Brien, M. K., Bansal, V. K., Andersen, D. K., & Satava, R. M. (2002). Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Annals of Surgery*, 458-464.

Sharples, S., Cobb, S., Moody, A., & Wilson, J. R. (2008). Virtual reality induced symptoms and effects (VRISE): Comparison of head mounted display (HMD), desktop and projection display systems. *Displays, 29*, 58-69.

Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007). Galvanic Skin Response (GSR) as an index of cognitive load. *CHI* (pp. 2651-2656). San Jose: ACM.

S'Jongere, J. J., Bertels, A. M., & Ego, T. (1977). Some psychosocial and physiologic aspects of Belgian glider pilots. *Bruxelles Medicine*, 309-320.

Skinner, M. J., & Simpson, P. A. (2002). Workload Issues in Military Tactical Airlift. *The International Journal of Aviation Psychology*, 79-93.

Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculation on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 603-616.

So, R. H., & Griffin, M. J. (1995). Effects of lags on human operator transfer functions with head-coupled systems. *Aviation, Space and Environmental Medicine*, 550-556.

Stanney, K. M., & Hash, P. A. (1998). Locus of user-initiated control in virtual environments: influences on cybersickness. *Presence: Teleoperators and Virtual Environments, 7*(5), 447-459.

Stanney, K. M., Kennedy, R. S., & Drexler, J. M. (1997). Cybersickness is not simulator sickness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, (pp. 1138-1142).

Stanney, K. M., Kennedy, R. S., & Hale, K. S. (2014). Virtual environment usage protocols. In K. S. Hale, & K. M. Stanney, *Handbook of Virtual Environments* (pp. 797-809). Boca Raton: CRC Press.

Stanney, K. M., Kennedy, R. S., Drexler, J. M., & Harm, D. L. (1999). Motion sickness and proprioceptive affereffects following virtual environment exposure. *Applied Ergonomics*, 27-38.

Stanney, K. M., Mourant, R. R., & Kennedy, R. S. (1998). Human factors issues in virtual environments: A review of the literature. *Presence*, 327-351.

Stevens, J. A., & Kincaid, J. P. (2015). The relationship between presence and performance in virtual simulation training. *Open Journal of Modelling and Simulation*, 41-48.

Stojmenova, K., & Sodnik, J. (2018). Detection response task - uses and limitations. *Sensors*, 1-17.

Sutcliffe, A., & Alrayes, A. (2012). Investigating user experience in Second Life for collaborative learning. *International Journal of Human Computer Studies*, 508-525.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 257-285.

Sweller, J., van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 251-296.

Tattersall, A. J., & Foord, P. S. (1996). An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, 740-748.

Taylor, H. L., Talleur, D. A., Emanuel, T. W., & Rantanen, E. M. (2005). Incremental transfer of training effectiveness of a flight training device (FTD). *Proceedings of the 13th International Symposium on Aviation Psychology* (pp. 1-4). Dayton, OH: Wright State University.

Terbizan, D. J., Dolezal, B. A., & Albano, C. (2002). Validity of Secen Commercially Available Heart Rate Monitors. *Measurement in Physical Education and Exercise Science*, 243-247.

Treisman, M. (1977). Motion sickness: an evolutionary hypothesis. *Science*, 493-495.

Van Benthem, K., & Herdman, C. M. (2016). Cognitive factors mediate the relation between age and flight path maintenance in general aviation. *Aviation Psychology and Applied Human Factors, 6*, 81-90.

Villarejo, M. V., Zapirain, B. G., & Zorrilla, A. M. (2012). A stress sensor based on Galvanic Skin Response (GSR) controlled by ZigBee. *Sensors, 12*, 6075-6101.

Wang, Z., & Fu, S. (2016). Evaluation of a strapless heart rate monitor during simulated flight tasks. *Journal of Occupational and Environmental Hygiene*, 185-192.

White, M. (1940). The effect of anoxia in high altitude flight on the electrocardiogram. *Journal of Aviation Medicine*, 166-180.

Wickens, C. D., Hyman, F., Dellinger, J., Taylor, H., & Meador, M. (1986). The Sternberg memory search task as an index of pilot workload. *Ergonomics*, 1371-1383.

Wilhelm, F. H., & Grossman, P. (2010). Emotions beyond the laboratory: theoretical fundamentals, study design, and analytic strategies for advanced ambulatory assessment. *Biological Psychology*, 552-569.

Wilson, G. F. (2002). An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures. *The International Journal of Aviation Psychology*, 3-18.

Wout, M. v., Spofford, C. M., Unger, W. S., Sevin, E. B., & Shea, M. T. (2017). Skin conductance reactivity to standardized virtual reality combat scenes in veterans with PSTD. *Applied psychophysiology and biofeedback*, 209-221.

Wu, X., Mu, G., Yang, Z. J., & Gu, C. (2012). Design and implementation of interactive virtual maintenance training system for tank gun. *Computer Science and Electronics Engineering* , 383-387.

Yeh, Y., & Wickens, C. (1988). Dissociation of performance and subjective measures of workload. *Human Factors*, 111-120.

Yerkes, R. M., & Dodson, J. (1908). The relationship of strength stimulus to rapidity of habitat formation. *Journal of Comparative Neurology and Psychology*, 459-482.

Young, S. D., Adelstein, B. D., & Ellis, S. R. (2007). Deman characteristics in assessing motion sickness in a virtual environment: or does taking a motion sickenss questionnaire make you sick? *IEEE Transactions on Visualization and Computer Graphics*, 422-428.

Zhai, J., & Barreto, A. (2006). Stress detection in computer users based on digital signal processing of noninvasive physiological variables. *Conference proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 1355-1358). IEEE Engineering in Medicine and Biology Society.

Appendix A – Factors that Contribute to Adverse Effects in VR

The following table summarizes the system, individual, and application factors that Jerald (2016e) describes as contributing to motion sickness in VR or causing other symptoms that may be related indirectly (e.g., headaches, eye strain, etc.). Each column is ordered from factors that have the greatest effect to ones that have the least.

System Factors	Individual User Factors	Application Design Factors
Latency – high latency is the greatest cause of motion sickness in VR	**History** – motion sickness by other physically or visually induced ways (e.g., from wavy waters) predict motion sickness in VR	**Frame Rate** – slow frame rates cause latency, which causes motion sickness
Calibration – bad calibration causes scene motion/scene instability when the user moves his/her head	**Health** – degraded health, such as ear infection, head cold, upset stomach, dehydration, hangover, etc. can increase motion sickness	**Locus of Control** – being actively in control reduces chances of motion sickness (e.g., a driver would experience less symptoms than passenger)
Tracking Accuracy – low accuracy head trackers cause motion sickness, but error in tracking the hands does not	**VR Experience** – less experience in VR can predict incidents of motion sickness (due to adaptation)	**Visual Acceleration** – increases chances of motion sickness, so should be avoided
Tracking Precision – low precision causes jitter which appears as shaking of the world	**Thinking about Sickness** – telling someone they may feel sick in VR may increase their susceptibility to it	**Physical Head Motion** – motion sickness is reduced when the user can keep their head still
Lack of Position Tracking – results in the world moving with the user when the user translates (e.g., the floor moves when the user leans down to pick up something)	**Age** – sickness in VR increases with age	**Duration** – Motion sickness increases with exposure time and breaks between sessions can help
Field of View – a wider field of view causes motion sickness due to scene motion	**Mental Model** – a person's expectations of the scene motion can cause sickness if they do not match (e.g., first person shooter games have different navigation than VR so this past experience can shape a mental model)	**Vection** – an illusion of self-motion when the person is not actually moving can cause motion sickness
Refresh Rate – low refresh rates cause latency, judder, and flicker, which lead to motion sickness	**Interpupillary Distance** – if the HMD is not calibrated to match the distance between the specific user's eyes, can cause feelings of sickness	**Binocular-Occlusion Conflict** – cues that do not match up can lead to motion sickness from eye strain and confusion

Judder – jerky or unsmooth motions can cause motion sickness	**Not Knowing What Looks Correct** – educating a user how to properly align an HMD can reduce chances of motion sickness, since most new users do not know how to set it up properly	**Virtual Rotation** – rotating the user's viewpoint can lead to feelings of motion sickness
Display Response Time and Persistence – trade-offs with judder, motion smear/blur, flicker and latency	**Sense of Balance** – Motion sickness in VR is high for those with postural instability	**Gorilla Arm** – extended use of the user's arm above their waist can cause arm fatigue
Flicker – can be distracting and cause eye fatigue, and even seizures	**Flicker-Fusion Frequency Threshold** – the threshold where flicker become noticeable is affected by gender, age, intelligence and others and this flicker can cause feelings of unwell	**Rest Frames** – stationary applications reduce tendency of motion sickness since they are consistent with the vestibular system
Vergence/Accommodation Conflict – visuals placed too close to the eye for long periods of time can cause motion sickness	**Real-World Task Experience** – having higher expectations of how the task is done in the real world can lead to motion sickness due to a mismatch	**Standing/Walking versus Sitting** – sitting during VR causes less motion sickness than standing
Binocular Images – two different images for each eye in an HMD can result in eye strain	**Migraine History** – correlated with motion sickness in VR	**Height Above the Ground** – a discrepancy between being visually at a different height than real world physical height can cause motion sickness
Eye Separation – conflict between inter-image distance, inter-lens distance, and interpupillary distance can conflict and cause symptoms		**Excessive Binocular Disparity** – motion sickness can result from seeing double when objects are too close to the eyes and cannot be fused from the left and right eye images
Real-World Peripheral Vision – unless an HMD is perfectly calibrated, seeing the real world from the periphery (in HMDs that are partially open) could cause symptoms,		**VR Entrance and Exit** – a blank screen during putting on and taking off the HMD can reduce feelings of motion sickness
Headset fit – uncomfortable and improper fitting HMDs, especially for those who wear glasses, can cause headaches, and result in HMD slippage which causes scene motion		**Luminance** – dark conditions can result in less motion sickness for displays with little persistence and low refresh rates (e.g., flicker)

Weight and Center of Mass – heavy HMDs and ones with a center of mass that is offset can modify the way distance and self-motion are perceived and lead to headaches		
Motion Platforms – can reduce motion sickness if well implemented otherwise can cause motion sickness		
Hygiene – unpleasant smells from an improperly cleaned HMD can cause nausea		
Temperature – increases in room temperature can cause feelings of sickness		
Dirty Screens – can result in eye strain due to not being able to see the image clearly		

Appendix B – Pre and Post-Test Questionnaire

Simulator Sensation Parts One - Three

Please complete each question below.

* Required

1. **Code** *

_____

2. **On a scale of 1 to 7 how alert are you right now?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Not at all alert	◯	◯	◯	◯	◯	◯	◯	Very alert.

3. **1.2 On a scale of 1 to 7 how fatigued are you right now?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Not at all fatigued	◯	◯	◯	◯	◯	◯	◯	Very fatigued.

4. **1.3 When did you last eat?** *
Mark only one oval.

- ◯ Within the last hour
- ◯ Within the last 2 to 5 hours
- ◯ More than 5 hours ago

5. **1.4 On a scale of 1 to 7 how queasy are you feeling right now?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Not at all queasy	◯	◯	◯	◯	◯	◯	◯	Very queasy

6. **1.5 On a scale of 1 to 7 how dizzy are you feeling right now?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Not dizzy at all.	◯	◯	◯	◯	◯	◯	◯	Very dizzy.

7. **1.6 On a scale of 1 to 7 how disoriented are you feeling right now?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Not at all disoriented.	◯	◯	◯	◯	◯	◯	◯	Very disoriented.

8. **1.7 Are you prone to motion sickness?** *
Mark only one oval.

- ◯ Yes
- ◯ No

9. **1.8 Have video games ever made you feel sick or queasy in the past?** *
Mark only one oval.

- ◯ Never.
- ◯ Sometimes I feel a little queasy.
- ◯ I always feel queasy when playing video games.
- ◯ No- I don't play video games

10. **1.9 Have you ever used Virtual Reality products before? E.g. Oculus Rift goggles** *
Mark only one oval.

- ◯ Never
- ◯ Maybe once or twice
- ◯ I am a regular user of virtual reality products

11. **1.10 Have Virtual Reality products ever made you feel sick or queasy in the past?** *
Mark only one oval.

- ◯ I use them and I have never felt sick or queasy from them
- ◯ Sometimes I feel a little queasy
- ◯ I always feel queasy when using virtual reality products
- ◯ I have never used virtual reality products

12. **1.11 How often do you play video games?** *
Mark only one oval.

- ◯ Everyday
- ◯ once or twice a week
- ◯ once or twice a month
- ◯ Rarely
- ◯ I used to play years ago, now I don't at all

13. **1.12 Do you have any piloting experience?** *
Mark only one oval.

- ◯ I have no piloting experience
- ◯ Yes I am a licensed pilot
- ◯ Yes, I used to be a licensed pilot
- ◯ I am in pilot training

14. **1.13 Gender** *

Mark only one oval.

- ◯ Male
- ◯ Female
- ◯ Other
- ◯ Prefer not to say

15. **1.14 Age** *

_____

16. **1.15 Language** *

Mark only one oval.

- ◯ English is my first language
- ◯ English is my second language

17. **1.16 Handedness** *

Mark only one oval.

- ◯ I am right-handed
- ◯ I am left-handed

18. **STOP HERE** *

Mark only one oval.

- ◯ OK

19. **2.1 On a scale of 1 to 7 how queasy are you feeling right now?** *

Mark only one oval.

	1	2	3	4	5	6	7	
Not at all queasy	◯	◯	◯	◯	◯	◯	◯	Very queasy

20. **2.2 On a scale of 1 to 7 how dizzy are you feeling right now?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Not dizzy at all.	◯	◯	◯	◯	◯	◯	◯	Very dizzy.

21. **2.3 On a scale of 1 to 7 how disoriented are you feeling right now?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Not at all disoriented.	◯	◯	◯	◯	◯	◯	◯	Very disoriented.

22. **STOP HERE** *
Mark only one oval.

◯ OK

23. **3.1 On a scale of 1 to 7 how queasy are you feeling right now?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Not at all queasy	◯	◯	◯	◯	◯	◯	◯	Very queasy

24. **3.2 On a scale of 1 to 7 how dizzy are you feeling right now?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Not dizzy at all.	◯	◯	◯	◯	◯	◯	◯	Very dizzy.

25. **3.3 On a scale of 1 to 7 how disoriented are you feeling right now?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Not at all disoriented.	◯	◯	◯	◯	◯	◯	◯	Very disoriented.

26. **3.4 On a scale of 1 to 7 how comfortable was the Oculus Rift at first?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Very uncomfortable	◯	◯	◯	◯	◯	◯	◯	Very comfortable

27. **3.5 On a scale of 1 to 7 how comfortable was the Oculus Rift by the end of your flight?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Very uncomfortable	◯	◯	◯	◯	◯	◯	◯	Very comfortable

28. **3.6 On a scale of 1 to 7 how would you rate the heaviness of the Oculus Rift?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Very light- barely felt it	◯	◯	◯	◯	◯	◯	◯	Very heavy

29. **3.7 On a scale of 1 to 7 how would you rate the crispness of the visual scene in the Oculus Rift?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Not crisp at all e.g. very blurry	◯	◯	◯	◯	◯	◯	◯	Extremely crisp e.g. no blurriness

30. **3.8 On a scale of 1 to 7 how would you rate the reality of the look of the visual scene in the Oculus Rift?** *

Mark only one oval.

	1	2	3	4	5	6	7	
Not real looking at all	◯	◯	◯	◯	◯	◯	◯	Extremely real looking

31. **3.9 On a scale of 1 to 7 how would you describe the fact that you could not see your own hands on the yoke (steering wheel) in the Oculus Rift?** *

Mark only one oval.

	1	2	3	4	5	6	7	
I barely noticed, and this did not affect my flying	◯	◯	◯	◯	◯	◯	◯	This really bothered me, and this affected my flying

32. **3.10 On a scale of 1 to 7 how would you describe the fact that you could not see your own hands when reaching for the throttle or flaps in the Oculus Rift?** *

Mark only one oval.

	1	2	3	4	5	6	7	
I barely noticed, and this did not affect my flying	◯	◯	◯	◯	◯	◯	◯	This really bothered me and affected my flying

33. **3.11 List any physical symptoms you felt, not already mentioned above, while flying with the Oculus Rift?** *

_____

34. **3.12 Indicate when these physical symptoms started while flying with the Oculus Rift?** *

Mark only one oval.

- ◯ Right away
- ◯ After about 10 minutes
- ◯ Near the end
- ◯ NA- no symptoms
- ◯ Other: _____