Comparing Flight Simulation Platforms: An Experimental Study

Bernsland, A., Graichen, E, Jansson, A. Kourie, G., Lindqvist, B. Piscator, F.
Linköping University, Bachelor's Programme in Cognitive Science

emigr679@student.liu.se, fripi876@student.liu.se, bjoli469@student.liu.se

Abstract

This study compares and explicates the differences in three human-factor measurements, and one performance measure, between a virtual reality- and a monitor-based flight simulator. Quantitative questionaries were utilized to measure usability, workload, and situational awareness. Furthermore, semi-structured interviews were conducted to evaluate the usability of the system. A two-condition within-group experiment design was utilized where participants performed a flight mission in both Virtual Reality (VR) and on a conventional flight simulator. Participants reported their experienced workload, situational awareness, and the usability of the simulation platform. Furthermore, participants' objective performance in the task was measured. The results showed that the VR condition was associated with a higher workload, lower performance, and lower usability than the monitor condition

1. Introduction

Since the end of the second world war, man-controlled machinery has been molded to fit human intuition. Systems, especially the ones used under immense stress, need to work intuitively and seamlessly to minimize potential human errors. How can these systems be designed to minimize human errors? How should professionals be trained to be ready for situations where lives are at stake? Simulation could be one part of the answer. This study examines how different kinds of flight simulators affect participants' experience of the simulated system and situation.

Minimizing human error is especially important in safety-critical environments such as nuclear engineering, medicine, and aviation. In contemporary pilot training, virtual simulation training is included in addition to the usual theoretical and practical training phases. Virtual training enables a safe and controlled learning environment (Kneebone, 2003) where the prospective pilot can train – try and retry without risking their safety. There are several other reasons why it is more beneficial to use simulations over conventional training, for example, cost, environmental concerns, and accessibility. In addition, the system itself can present feedback on the pilots' strengths and possible improvements after a virtual mission has been accomplished. Alternatively, an instructor can give personal feedback on the flight that was conducted. Immensely stressful situations, found in the military domain, where weapons and complex flight maneuvers occur are better to simulate as it is difficult to accomplish in peacetime. In addition, testing new equipment and tactics is safer and more cost-effective to do in a simulated environment (Nählinder, 2009).

1.1 Purpose

The purpose of this study is to explore and understand the differences between a Virtual reality (VR) flight simulator platform and a conventional monitor-based platform. Previous studies have compared fidelity, the simulators perceived level of realism, and workload in VR flight simulators and conventional monitor platforms (Oberhauser et al., 2018). Oh's (2020) study supported the claims of Oberhauser et al. (2018) that flight simulation in VR might not replace conventional flight simulation completely. Oh (2020) expressed that there are some limitations to the maneuvering of the aircraft in VR compared to conventional simulators. However, the realism of the VR simulator was perceived as equal to or better than the conventional one.

To predict and measure the usefulness of a simulation platform, several, both subjective and objective, aspects need to be examined to develop a comprehensive understanding of which parts of a simulation platform need to be further developed. This study will examine how different simulation platforms perform in the following areas: workload, situational awareness, usability, and performance. Further, a thematic analysis of semi-structured interview data and observations explores participants' opinions on both systems. The results could help to verify the claims of Oh (2020) and Oberhauser et al. (2018) and further investigate which aspects of flight simulation are better or worse for both platforms.

1.2 Hypotheses

The hypotheses for this study were that there would be a difference in workload, situational awareness, and performance in both the VR-flight simulator and monitor-based flight simulator. The reasoning for this is that participants would be more likely to be familiar with monitor-based systems than VR-based ones, for example from playing video games. This would make monitor use less reliant on learning new things, causing the participants to be able to focus more on the task. Usability was hypothesized to specifically decrease in the VR condition compared to the monitor condition because of a lack of VR experience in the participants.

2. Theoretical Background

2.1 Workload

When studying flight decks and comparing different settings for evaluation, one measure commonly used is known as workload (Oberhauser et al., 2015). Workload can be defined as the physical and mental strain that a person is subject to when performing a task, though there are many different interpretations of the concept. The progress of technology has led to the tasks of operators in technical systems becoming increasingly demanding. However, any operator has a limited amount of attentional resources to use during a task. Therefore, the mental workload becomes an increasingly important matter to keep in check when designing a new system, especially one with abundant human-machine interaction, such as aviation. Mental workload may alter the outcome of a task performed in the system, with an excessive mental workload often leading to poor performance (Stanton et al., 2005). From a perspective of user experience mental workload may affect the subject's health and wellbeing, for example, a high workload when using flight simulators has been found to correlate with simulation sickness (Stein & Robinski, 2012).

There are several ways to measure mental workload. One well-established (Stanton et al., 2005) subjective rating technique is known as *The NASA Task Load Index (NASA-TLX)*. This method is based on the notion that researchers cannot get the required information about a user's perceived cognitive demand through observation alone. NASA-TLX, therefore, seeks to acquire a subjective rating on the workload of a user by letting them grade six sub-scales that eventually combine to produce the overall NASA-TLX score. Each of the sub-scales answers one of the following questions (Hart & Staveland, 1988):

- How mentally demanding was the task?
- How physically demanding was the task?
- How hurried was the pace of the task?
- How successful were you in accomplishing the goals of the task?
- How much effort did you have to put in to perform at this level?
- How frustrated were you?

The rating scores may be gathered either during a task, immediately upon the subject completing a task, or after a short delay. The timing might be important as the process can be intrusive to task performance if initiated before the subject is completely finished. However, if the process takes place after the task is done there is a risk of the subject forgetting workload aspects (Stanton et al., 2005). Moroney et al. (1992) found that ratings could be delayed up to 15 minutes without differing from immediate ones. To summarize, NASA-TLX is an easy-to-use and effective technique to estimate the workload of an operator (Stanton et al., 2005). It is thoroughly tested and has a high validity (Hart & Staveland, 1988). However, using it requires keeping a few downsides in mind, for example, that ratings can be affected by the performance of the subject in the task (Stanton et al., 2005).

2.2 Situational Awareness

Situational awareness (hereby SA) refers to a subject's level of awareness of the situation and environment its currently in. Endsley (1988) defines SA as: "*the perception of the elements in the environment within a volume of space and time, the comprehension of their meaning, and the projection of their status in the near future*". SA refers to the subject's ability to perceive and understand the situation and what it means for future situations. No universally endorsed model of SA has been developed, although attempts have been made. Models of SA can be divided into two overarching categories: individual approaches and distributed approaches. The distributed approaches take a holistic view of the situation and include artifacts, other actors, and the individual in their analysis of SA. Individual approaches focus on the perspective of the individual operator and will be the approach of this report (Stanton et al., 2005).

According to Salmon et al. (2009), there are six categories of techniques for measuring SA. Freeze-probe recall techniques, real-time probe techniques, self-rating techniques, observer-rating techniques, performance measures, and process indices. The techniques used in this study are a combination of freeze-probe recall techniques, self-rating techniques, and performance measures.

Freeze-probe recall techniques try to measure SA at its source. The procedure involves freezing the displays during the performance of a task and asking the subject about their understanding of the situation. A comparison between the state of the simulation and the operator's reported knowledge can then be made to calculate a SA score. The advantages of freeze-probe recall techniques are that they are objective and direct and avoid the shortcomings of asking participants post-trial, e.g., that subjects must recall their SA (Salmon et al., 2009).

Self-rating techniques are criticized for relying on subjects recalling their own situational awareness and their lack of sensitivity. They are however easy to use and widely applied to measure SA. (Salmon et al., 2009) One of the most common self-rating techniques, simply called the Situational Awareness Rating Technique (SART), was developed by Taylor (1990) and was used in this study. SART measures SA in ten dimensions, rated on a scale ranging from 1-to 7, which are later combined into a final subjective SA score. The dimensions combine into three categories: Information demand, information supply, and understanding. The total score is then calculated via this formula generating a value between 3 and 46:

$$SA = Understanding - (Information\ demand - Information\ supply)$$

2.3 Performance

Performance measures reflect an indirect way of measuring SA and workload where the measure highly varies depending on the task. They are based on how well an operator executes some aspects of a given assignment. Salmon et al. (2009) exemplify this by illustrating that a military exercise might have "kills" or "hits" as a performance measure which is dependent on the operator having good SA. In this study, performance could be measured by examining how much participants deviate from a given flight path, how well they perform takeoff and landing, or how many objects they could identify in a body of water.

Although performance reflects SA and workload indirectly, improving operator performance is the sole reason to study SA and workload. Airlines, aircraft manufacturers, and passengers mainly want their planes to be safe and perform well. The SA of the pilot has little intrinsic value and mainly represents the means to reach good performance. Therefore, sometimes it is preferred to measure performance directly and cut out secondary dependent variables, like SA or workload.

2.4 Usability

Usability is a comprehensive concept that is often associated with “ease-of-use" or “user-friendliness” by laymen. Formally, however, it’s defined as “the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” according to international standards (ISO 9241-11). Products or systems with low usability inconvenience users which can lead to irritation, confusion, delays, misleading, and a hampered ability to learn.

Although usability is defined by international standards, there is no obvious consensus between researchers regarding what aspects of a system affect the phenomenon (Hertzum, 2020). Brooke (1995) argues that part of the reason it is difficult to define usability is that it does not exist in any real or absolute sense; instead, the usability of any tool must be viewed in terms of the context in which it is used, and its appropriateness to this standard. There are, however, attempts to define general aspects important to the usability of a system.
Jakob Nielsen, a leading researcher in the usability field, provides one definition that is widely accepted and used in a wide variety of studies (Wilson, 2010). He divides usability into five quality components:

- Learnability
- Efficiency
- Memorability
- Error tolerance and prevention
- Satisfaction

Usability can be estimated either quantitively or qualitatively (Huang et al., 2016). There are pros and cons to both methods. Quantitative measurements provide a good estimate of the system's overall usability. However, these are seldom diagnostic, which makes it hard to identify exactly what users did or did not find usable with the system. Qualitative methods, however, are better at diagnosing what specific parts of the system users struggle with or appreciate due to their open-ended nature.
The system usability scale (SUS) is an industry standard when it comes to quantitatively measuring usability. The method was originally developed by John Brooke (1986) and has since been published in over 1200 studies to measure usability in a variety of different products and systems (Gallavin, 2014). SUS evaluates participants' subjective rating of their agreement with 10 questionnaire items. The items correspond to participants' willingness to use the system, the complexity of the system, and how effortless it was to use to name a few. The rating is then combined to a single number score from 0-to 100 that represents the overall usability of the system. A score of over 81 is generally considered excellent and a score below 51 is generally considered awful (Sasmito et al, 2019).

3. Method

A two-conditioned, ingroup experimental design was employed to examine the possible differences in performance, workload, situational awareness, and usability between a VR-based flight simulator and a monitor-based flight simulator. Participants were allowed to familiarize themselves with both simulators with the guidance of the test leader before each mission and ask questions regarding maneuvering, instruments, and controls.

The task was identical in both conditions. Take off from Kalmar Airport (KLR) and climb to an altitude of 1000 feet, followed by an eastward turn toward the bridge across Kalmar Strait. When the bridge was in sight the simulation was frozen and the participant answered the NASA-TLX questionnaire and the SART questionnaire. This was to measure the workload and Situational awareness of the participant without them having to recall their experience. This concluded the first phase. The participant was then instructed to search for ships in the water body south of the bridge, and orally report sightings of ships to the test leader. The participant had to fly southwards and not turn around but was allowed to fly in a search pattern if deemed necessary. The strait was to be searched until the participant found an oil rig, generated in the southern part of the Kalmar Strait. The mission was then to land the plane somewhere in the terrain. The NASA-TLX- and SART questionnaires were answered again together with the SUS questionnaire and five open-ended questions corresponding to Nielsen’s Usability dimensions.

This procedure was repeated for the second condition. Participants were randomly assigned into two groups; each group being assigned to simulator conditions in reversed order to counterbalance any learning effects. The test leaders made observations regarding maneuvering and other behaviors in both simulators which were later analyzed in the thematic analysis.

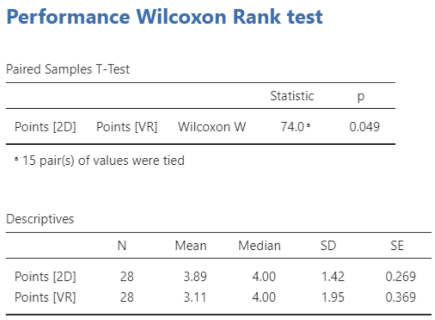
One performance point was assigned to every participant for finding a ship (4 possible points). One point was assigned for finding the oil rig. The maximum number of points was 5. These points reflected the participants’ performance in the task.

3.1 Participants

Twenty-eight (N=28) students between the ages of 21-27 years (SD = 1.85, mean = 23, median = 23) at Linköping University were recruited for this study. The recruitment letter consisted of an informative abstract and a short participation form, sent out to students via mail and Facebook messenger. The form surveyed their prior familiarity with virtual reality, flight simulation, and video gaming. A signed consent form was obtained from each participant prior to the start of the experiment. The assignment took place at the time of signing the consent form. The study was approved by supervisors at the University of Linköping.

4. Results

4.1 Performance points



Performance Wilcoxon Rank test

Paired Samples T-Test

			Statistic	p
Points [2D]	Points [VR]	Wilcoxon W	74.0 [a]	0.049

[a] 15 pair(s) of values were tied

Descriptives

	N	Mean	Median	SD	SE
Points [2D]	28	3.89	4.00	1.42	0.269
Points [VR]	28	3.11	4.00	1.95	0.369

Table 1: Paired samples Wilcoxon rank test of performance points in both the VR- and Monitor Condition.

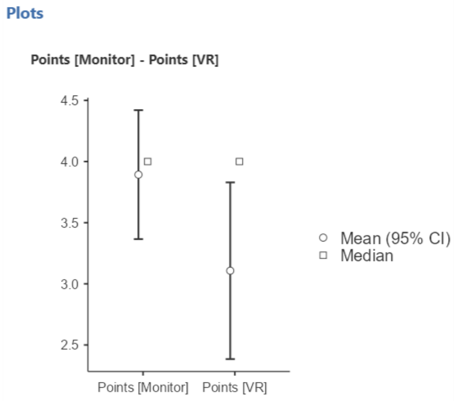
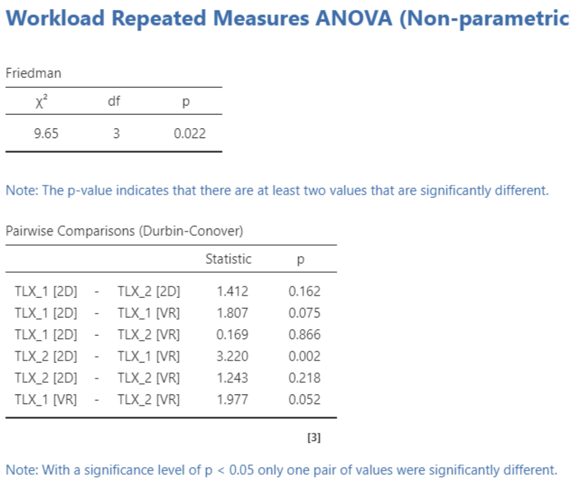


Figure 1: Graph of the performance Wilcoxon rank test.

Participants' performance scores were compared between the monitor, and VR conditions. On average, participants had fewer points in the VR (Mean = 3.11) condition than in the monitor condition (Mean = 3.89). A two-tailed Wilcoxon signed-rank test indicated that this difference was statistically significant, W = 74.0, p = .049.

4.2 Workload



Workload Repeated Measures ANOVA (Non-parametric)

Friedman

χ^2	df	p
9.65	3	0.022

Note: The p-value indicates that there are at least two values that are significantly different.

Pairwise Comparisons (Durbin-Conover)

			Statistic	p
TLX_1 [2D]	-	TLX_2 [2D]	1.412	0.162
TLX_1 [2D]	-	TLX_1 [VR]	1.807	0.075
TLX_1 [2D]	-	TLX_2 [VR]	0.169	0.866
TLX_2 [2D]	-	TLX_1 [VR]	3.220	0.002
TLX_2 [2D]	-	TLX_2 [VR]	1.243	0.218
TLX_1 [VR]	-	TLX_2 [VR]	1.977	0.052

[3]

Note: With a significance level of $p < 0.05$ only one pair of values were significantly different.

Table 2: A Friedman test of workload, measured twice in both conditions.

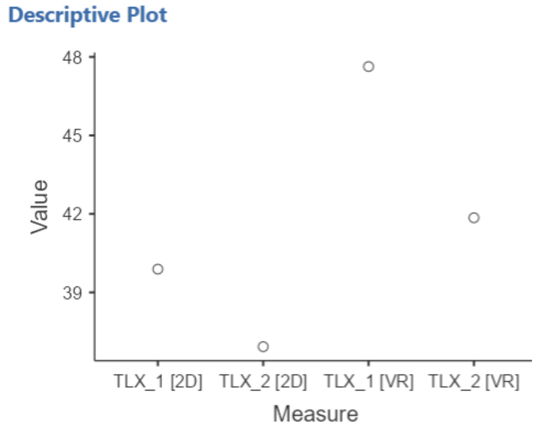
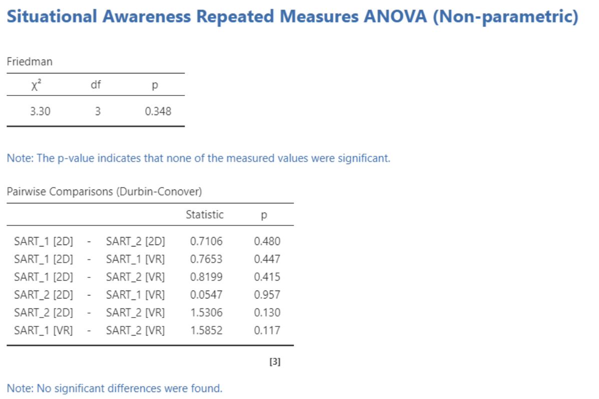


Figure 2: A graph of the Friedman test measuring workload. The points represent means of the measured values.

A Friedman test was conducted to determine whether there was a difference in workload between VR use versus on a monitor. This resulted in a significant difference, χ^2 (3) = 9.65, p = 0.022. The p-value indicated that there were at least two values that were significantly different. In this case these values were trial 2 using monitors and trial 1 using VR, Statistic = 3.222 p = 0.002. This means that workload was significantly lower during trial 1 using monitors than in trial 2 using VR.

4.3 Situational Awareness



Situational Awareness Repeated Measures ANOVA (Non-parametric)

Friedman

χ^2	df	p
3.30	3	0.348

Note: The p-value indicates that none of the measured values were significant.

Pairwise Comparisons (Durbin-Conover)

			Statistic	p
SART_1 [2D]	-	SART_2 [2D]	0.7106	0.480
SART_1 [2D]	-	SART_1 [VR]	0.7653	0.447
SART_1 [2D]	-	SART_2 [VR]	0.8199	0.415
SART_2 [2D]	-	SART_1 [VR]	0.0547	0.957
SART_2 [2D]	-	SART_2 [VR]	1.5306	0.130
SART_1 [VR]	-	SART_2 [VR]	1.5852	0.117

[3]

Note: No significant differences were found.

Table 3: A Friedman test of situational awareness, measured twice in both conditions.

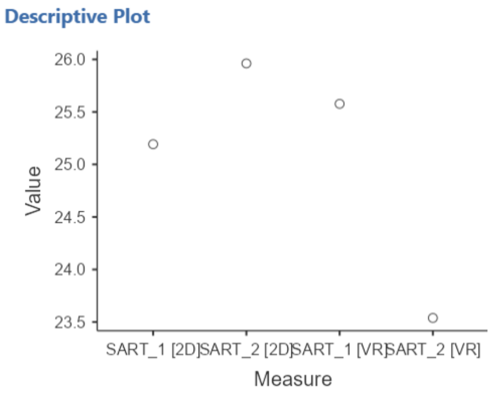
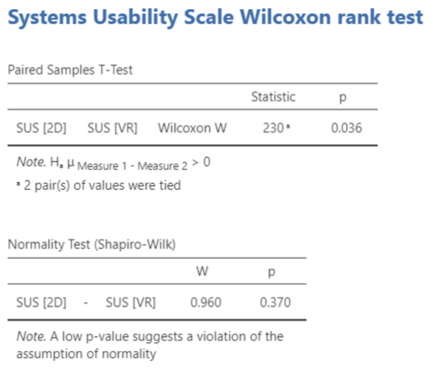


Figure 3: A graph of the Friedman test measuring situational awareness. The points represent the means of the measured values.

A Friedman test was conducted to determine whether there was a difference in situational awareness between VR versus on a monitor. This did not result in a significant difference, χ^2 (3) = 3.30, p = 0.348. The p-value indicated that none of the measured differences in values were significant.

4.4 Usability



Systems Usability Scale Wilcoxon rank test

Paired Samples T-Test

			Statistic	p
SUS [2D]	SUS [VR]	Wilcoxon W	230 ᵃ	0.036

Note. Hₐ μ $_{\text{Measure 1 - Measure 2}}$ > 0

ᵃ 2 pair(s) of values were tied

Normality Test (Shapiro-Wilk)

			W	p
SUS [2D]	-	SUS [VR]	0.960	0.370

Note. A low p-value suggests a violation of the assumption of normality

Table 4: A Wilcoxon rank test of differences in reported Usability. Measured once after phase two of each condition.

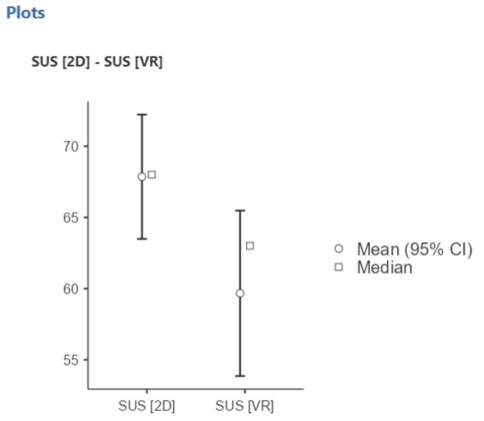


Figure 4: A graph of the Usability Wilcoxon rank test.

A one-tailed Wilcoxon signed-rank test indicated that there was a statistically significant difference, W = 230, p = 0.036. This means that participants generally found monitor use to be more useable (*Mean* = 67.9) than VR use (*Mean* = 59.7).

4.5 Thematic Analysis

Interviews were conducted after the second phase. The themes that were found were *Control complications in virtual reality*, *Monitor learning advantage*, *Natural movements in Virtual Reality, and graphical issues in Virtual Reality*.

The first theme represents the participants' opinions about the difficulty in maneuvering the controls in virtual reality. Their lack of feedback was a recurring account, as well as their sensitivity. The controls were described as non-intuitive in a way that they felt far away from reality, as it was not an actual joystick that was used in VR, but a portable control that was held horizontally. This resulted in the participants having too much freedom in moving them around and together with their sensitivity made the aircraft often spin uncontrollably. The next theme, *monitor learning advantage*, showed a profit in learning gained by the participants who started their test in the monitor-based simulator. These assumptions are supported both by the estimated time given by the participants, as well as interpretations of descriptions from the simulator learning experience. Simply, the participants often experienced it easier to perform in VR if their first flight had been in the monitor-based simulator. The third theme, *Natural movements in Virtual Reality*, describes a close connection to reality as looking around in VR is represented by the same action as turning one's head to identify an object in everyday life. However, the graphics were quite bad, which made this feature mainly brought up for its positive and true-to-life influence on experience, rather than something crucial for a successful flight. The graphical issue represents the fourth theme. Almost every participant expressed that the VR display had poor graphics, which they expressed to be one of the main reasons for their errors. Some further thought the low resolution prolonged the time it took to learn the VR system.

5. Discussion

5.1 Quantitative findings

The statistical differences showed that the first phase in VR on average caused a higher workload to be experienced by the participants than the second monitor phase. This along with the fact that each monitor phase had a lower mean value than its respective VR phase (Table 2) indicates that VR use causes a greater workload than monitor use. The results are reasonable for several reasons but chiefly because of the implications of the learning process. As the thematic analysis showed, the monitor phases offered, in general, a faster learning experience than the VR phases. The resulting higher amount of workload in VR would go along with Stanton's (2005) statement that a higher workload leads to poorer performance since participants scored significantly higher in the monitor phases (Table 1). It would also match platform differences in the notes taken by the testing group during the tests. It was for example noted that at least ten participants expressed signs of simulator sickness such as nausea, headaches, and sweating during the VR testing while no such signs were seen during monitor testing. This is in line with the findings of Stein & Robinski (2012) concerning a high workload causing simulator sickness. One relationship found was between perceived workload and previous video game experience which indicates that familiarity with video games seems to reduce the perceived physical and mental strain, Workload, that the simulator task had on participants.

Situational awareness' results were not significant (Table 3), but it may still be of importance to investigate why that is. It could be argued that the SA task was considered too easy to get a varied outcome – hence the insignificant result. In phase two, where the participant only identifies boats and follows the strait, it is not too demanding on the attentional resources. This could potentially resolve the problem.

Usability tests showed that there was a significant difference between VR-mode and monitor-mode. The mean difference according to the SUS scores was 58.8 in the VR condition and 67.9 in the monitor condition which means that the latter had a higher-rated usability score. This may be because of the high complexity of VR and the relative simplicity of the monitor setup. This difference occurs because of the *mental models* (Preece, et al., 2020) that the participants have when using monitor screens and tactile levers and buttons over virtual ones. Another difference that might explain this difference is the user interface variety that the two modes have.

The results in performance showed that participants on average scored significantly higher during monitor testing than during VR testing (Table 1). This is a logical consequence of the higher workload during the VR testing and, again, it goes along with Stanton's (2005) statement that a high workload leads to poor performance. The higher grade of usability in the monitor phases (Table 4) is also likely to influence how well the participant could navigate in order to complete the task. Also, in agreement with the thematic analysis, the participants' dissatisfaction with the controls and graphical issues in VR is likely to have affected their performance in the mission, thereby having an impact on their score. Unsurprisingly, previous flight simulator experience had a significant positive effect on the performance points gathered in both the monitor condition, (r(26) = 0.391, p = 0.040), and the VR condition, (r(26) = 0.542, p = 0.003). Participants with video game experience also scored significantly better in the monitor condition (r(26 = 0.412, p = 0,029). No significant improvement was found for video gamers in the VR condition.

5.2 Qualitative findings

The graphical issues and the control complications in virtual reality were two complexities that were considered to be contributory to impaired learning and completion of the given tasks. For example, to achieve an ideal takeoff the participants would have to use the two controls, the altitude tracker and compass at the same time. However, as the graphics were poor and the controls were not that intuitive, it seemed hard for some participants to implement all three gadgets at the same time. Which also were expressed by a few. These complexities often made the aircraft spin around in an uncontrolled manner and could make the participant lose track of space. Instead, a tactile system that implies physical feedback when handling its controls is highlighted as beneficial according to thematic analysis. Accordingly, descriptions from the Wilcoxon signed-rank test generated from performance measured in VR respective monitor-based simulator also act in accordance with what is just discussed. Together, this illustrates a convincing result of how the monitor-based simulator was better designed to accomplish the given tasks.

6. Conclusions

Through the analysis of the collected data and the information from the theoretical background, a summary conclusion can be drawn. Participants found monitor-based flight simulation to be more usable and sometimes less workload-inducing. They also scored higher on the associated tests. Concerning situational awareness, no significant difference was found. While this speaks to the advantage of monitor-based flight simulators, the thematic analysis discovered certain aspects of VR that the participants appreciated, such as the ability to look around more freely. Many of the downsides of VR could be traced to factors such as poor graphics or a lack of experience and mental models, both of which were less prevalent when using a monitor. The conclusion is therefore that a combination of both platforms (for example, a VR headset with physical controls accurate to the visuals) would make an interesting study object for future experiments. A longer learning period and better hardware could perhaps counterbalance the lacking graphics and mental models that held the medium back in this study. Furthermore, an altered mission difficulty might be necessary to make the participants more sensitive to changes in situational awareness.

References

Brooke, J. (1996). *"SUS-A quick and dirty usability scale." Usability evaluation in industry*. CRC Press.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In: Proceedings of the Human Factors Society 32nd Annual Meeting, pp. 97-101. Human Factors Society, Santa Monica, CA.

Gallavin, G., 2022. *System Usability Scale (SUS): Improving Products Since 1986*. [online] Digital.gov. Available at: <https://digital.gov/2014/08/29/system-usability-scale-improving-products-since-1986/> [Accessed 25 May 2022].

Hart, S. G., & Staveland, L. E. (1988). *Advances in Psychology, 52*, 139-183. https://doi.org/10.1016/S0166-4115(08)62386-9

Huang, D., Bevilacqua, V., Premaratne, P. and Jo, K., 2016. *Intelligent computing theories and application*. 12th ed. Cham: Springer International Publishing.

Kneebone, R. (2003). Simulation in surgical training: educational issues and practical training. Medical Education. Vol. 37(3), pp. 267-277. https://doi.org/10.1046/j.1365-2923.2003.01440.x

Moroney, W. F., Biers, D. W., Eggemeier, F. T., & Mitchell, J. A. (1992). A comparison of two scoring procedures with the NASA task load index in a simulated flight task. *Proceedings of the IEEE 1992 National Aerospace and Electronics Conference*. 10.1109/NAECON.1992.220513

Nielsen Norman Group. 2012. *Usability 101: Introduction to Usability*. [online] Available at: <https://www.nngroup.com/articles/usability-101-introduction-to-usability/> [Accessed 25 May 2022].
Nählinder, S. (2009). *Flight Simulator Training: Assessing the Potential.* (Rapport 1250). Department of Management and Engineering. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1038.3067&rep=rep1&type=pdf

Oberhauser, M., Dreyer, D., Mamessier, S., Convard, T., Bandow, D., & Hillebrand, A. (2015). Bridging the Gap Between Desktop Research and Full Flight Simulators for Human Factors Research. *EPCE 2015: Engineering Psychology and Cognitive Ergonomics*, 460-471. *Lecture Notes in Computer Science*, *9174*. https://doi.org/10.1007/978-3-319-20373-7_44

Oberhauser, M., Dreyer, D., Braunstingl, R., & Koglbauer, I. (2018). What's Real About Virtual Reality Flight Simulation?. *Aviation Psychology And Applied Human Factors*, *8*(1), 22-34. doi: 10.1027/2192-0923/a000134

Oh, C. (2020). Pros and Cons of A VR-based Flight Training Simulator; Empirical Evaluations by Student and Instructor Pilots. *Proceedings Of The Human Factors And Ergonomics Society Annual Meeting*, *64*(1), 193-197. doi: 10.1177/1071181320641047

Preece, J., Rogers, Y., Sharp, H., & Rizzo, F. (2016). *Interaction design*. Lund: Studentlitteratur.

Salmon, P., Stanton, N., Walker, G., Jenkins, D., Ladva, D., Rafferty, L., & Young, M. (2009). Measuring Situation Awareness in complex systems: Comparison of measures study. *International Journal Of Industrial Ergonomics*, *39*(3), 490-500. doi: 10.1016/j.ergon.2008.10.010

Sasmito, G., Zulfiqar, L. and Nishom, M., 2019. Usability Testing based on System Usability Scale and Net Promoter Score. *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI),*

Hertzum, M. (2020). *Usability testing : a practitioner's guide to evaluating the user experience*. Morgan & Claypool.

Stanton, N. A., Salmon, P. M., & Walker, G. H. (2005). *Human Factors Methods : A Practical Guide for Engineering and Design*. Taylor & Francis Group. https://doi.org/10.1201/9781315587394

Stein, M., & Robinski, M. (2012). Simulator Sickness in Flight Simulators of the German Armed Forces. *Aviation Psychology and Applied Human Factors*, *2*(1), 11-19. https://doi.org/10.1027/2192-0923/a000022

Taylor, R.M., 1990. Situational Awareness Rating Technique (SART): the development of a tool for aircrew systems design (AGARD-CP-478) pp3/1 –3/17. In: Situational Awareness in Aerospace Operations. NATO-AGARD, Neuilly Sur Seine, France.