

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

ASSESSING SITUATION AWARENESS WITH
EYE MOVEMENT

NESİBE YAVUZ

A THESIS OF MASTER OF SCIENCE
DEPARTMENT OF COMPUTER ENGINEERING

ADVISOR: ASSOC. PROF. DR. MEHMET GÖKTÜRK

JUNE 2024

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

**ASSESSING SITUATION AWARENESS WITH EYE
MOVEMENT**

NESİBE YAVUZ

A THESIS OF MASTER OF SCIENCE
DEPARTMENT OF COMPUTER ENGINEERING

ADVISOR: ASSOC. PROF. DR. MEHMET GÖKTÜRK

JUNE 2024

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

GÖZ HAREKETİ İLE DURUM FARKINDALIĞININ
DEĞERLENDİRİLMESİ

NESİBE YAVUZ

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN: DOÇ. DR. MEHMET GÖKTÜRK

HAZİRAN 2024

MASTER of SCIENCE JURY APPROVAL FORM

A thesis submitted by Nesibe YAVUZ, defended on 11/06/2024 before the jury formed with the 12/02/2024 date and 2024/13 numbered decision of the GTU Graduate Administration Board, has been accepted as a MASTER of SCIENCE thesis in the Department of Computer Engineering.

JURY

MEMBER

(THESIS ADVISOR) : Assoc. Prof. Dr. Mehmet GÖKTÜRK

MEMBER : Prof. Dr. Kürşat ÇAĞILTAY

MEMBER : Assist. Prof. Dr. Alp Arslan BAYRAKÇI

APPROVAL

Gebze Technical University Graduate Administration Board

.../.../..... date and/.... numbered decision.

SIGNATURE/SEAL

To my Family

ABSTRACT

Over the past 36 years, significant attention has been given to situation awareness (SA) as a psychological concept critical to performance, particularly in aviation where it is vital for Air Traffic Control (ATC) safety. With the increasing density of air traffic, air traffic controllers must maintain focus while monitoring radar, weather, and electronic strip displays. This research aimed to evaluate SA using the Situation Awareness Global Technique (SAGAT), measuring individual performance, and eye movements. In this study, twenty-four subjects participated in an ATC simulation, were directed to observe aircraft and identify those near collision. The simulation paused at random intervals, prompting participants to answer an average of 11 SAGAT questions across three SA levels. Responses were recorded to compute SAGAT Q/A scores for assessing participants' SA levels. Results showed no significant change in missed clicks, performance scores, SAGAT scores, correct/incorrect questions, target missed fixation count, total fixation count, and eye tracker scores ($p > 0.05$) between intervals. A significant difference was found only in the number of resolved aircraft between interval 1 and interval 3 ($p < 0.05$). The target hit fixation count differed significantly between intervals 1 and 3, and intervals 1 and 4. There was a weak positive correlation ($r = 0.121$) between SAGAT and performance scores, and a weak negative correlation ($r = -0.207$) between performance and eye tracker scores. The study compared eye movement tracking methods with SAGAT for assessing ATC situational awareness. Statistical analysis revealed no significant difference between the two approaches ($p > 0.001$), despite initial expectations favoring eye movement tracking. The higher correlation magnitude of eye movement tracking over SAGAT does not provide sufficient evidence to consider it a more accurate SA measure. SAGAT's limitations, such as task interruption, suggest it may not be suitable for real-time scenarios. Exploring alternative approaches less dependent on memory could validate SAGAT's effectiveness. Further research is needed to explore the potential benefits of eye tracking, as it might offer unique insights that could enhance its utility as a superior SA indicator compared to SAGAT.

Keywords: Situational Awareness, Air Traffic Control, SAGAT, Eye Tracker

ÖZET

Geçtiğimiz 36 yıl boyunca, özellikle Hava Trafik Kontrol (ATC) güvenliği için hayati önem taşıyan havacılıkta, performans için kritik bir psikolojik kavram olarak durum farkındalığına (DF) büyük önem verilmiştir. Hava trafiğinin artan yoğunluğuyla birlikte, hava trafik kontrolörleri radar, hava durumu ve elektronik strip ekranlarını izlerken odaklanmayı sürdürmelidir. Bu araştırma, Durum Farkındalığı Küresel Tekniğini (SAGAT) kullanarak, bireysel performansı ve göz hareketlerini ölçerek SA'yı değerlendirmeyi amaçlamıştır. Bu çalışmada, yirmi dört katılımcı bir ATC simülasyonuna katılmış, uçakları gözlemlemeye ve çarpışmaya yakın olanları belirlemeye yönlendirilmiştir. Simülasyon rastgele aralıklarla duraklatılarak katılımcılardan üç DF seviyesinde ortalama 11 SAGAT sorusunu yanıtlamaları istenmiştir. Yanıtlar, katılımcıların DF seviyelerini değerlendirmek üzere SAGAT Q/A puanlarını hesaplamak için kaydedilmiştir. Sonuçlar, aralıklar arasında kaçırılan tıklamalarda, performans puanlarında, SAGAT puanlarında, doğru/yanlış sorularda, hedef kaçırılan fiksasyon sayısında, toplam fiksasyon sayısında ve göz izleyici puanlarında ($p > 0,05$) önemli bir değişiklik olmadığını göstermiştir. Aralık 1 ve aralık 3 arasında sadece çözülen uçak sayısında anlamlı bir fark bulunmuştur ($p < 0.05$). Hedefe isabet fiksasyon sayısı 1. ve 3. aralıklar ile 1. ve 4. aralıklar arasında önemli ölçüde farklılık göstermiştir. SAGAT ve performans puanları arasında zayıf bir pozitif korelasyon ($r = 0,121$) ve performans ile göz hareketleri izleme puanları arasında zayıf bir negatif korelasyon ($r = -0,207$) vardı. Çalışma, ATC durumsal farkındalığını değerlendirmek için göz hareketi izleme yöntemlerini SAGAT ile karşılaştırmıştır. İstatistiksel analiz, göz hareketi takibi lehine olan ilk beklentilere rağmen, iki yaklaşım arasında anlamlı bir fark olmadığını ortaya koymuştur ($p > 0.001$). Göz hareketi takibinin SAGAT'a göre daha yüksek korelasyon büyüklüğü, bunun daha doğru bir DF ölçümü olarak değerlendirilmesi için yeterli kanıt sağlamamaktadır. SAGAT'ın görev kesintisi gibi kısıtlamaları, gerçek zamanlı senaryolar için uygun olmayabileceğini düşündürmektedir. Hazıfayı daha az bağımlı alternatif yaklaşımların araştırılması SAGAT'ın etkinliğini doğrulayabilir. Göz takibinin potansiyel faydalarını keşfetmek için daha fazla araştırmaya ihtiyaç vardır, çünkü bu, SAGAT'tan daha üstün bir durumsal farkındalık göstergesi olarak kullanımını geliştirebilecek özel bilgiler sağlayabilir.

Anahtar Kelimeler: Durumsal Farkındalık, Hava Trafik Kontrol, SAGAT, Göz Takibi

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Assoc. Prof. Dr. Mehmet GÖKTÜRK from Gebze Technical University for his guidance and support. His knowledgeable advice, persistent patience, and perceptive criticism have all been invaluable in helping to shape this thesis. I am motivated to push the limits of my ability and pursue excellence by his commitment to my academic and personal development. It is also important that acknowledgment is given to every author contributing to the related studies and work mentioned in this thesis.

My sincere gratitude goes out to my family for their constant encouragement and support during this journey. My parents have always been a source of strength and inspiration because of their unending love, tolerance, and faith in my skills. The person I am today has been nurtured by their advice and sacrifices. I'm grateful to my family for being my pillar of support and for being there for me no matter what.

I am profoundly grateful to all the subjects who participated in my experiments. Their willingness to dedicate time and energy to this research has been crucial to its success. Their collaboration and excitement were invaluable in obtaining the required information and understanding, and I sincerely value their dedication to furthering this field's understanding. We could not have conducted this study without their involvement.

I want to sincerely thank TÜBİTAK BİLGEM, my place of employment, for giving me the tools and resources I needed to complete my experiments, as well as the chance to do so. The assistance has been of great value in making my research possible, and I am very grateful for the facilities, equipment, and cooperative atmosphere that allowed me to carry out my job efficiently. We appreciate their kindness and the environment they have created that promotes creativity and scientific research.

TABLE OF CONTENTS

	<u>Page</u>
ABSTRACT	vi
ÖZET	vii
ACKNOWLEDGEMENTS	viii
TABLE OF CONTENTS	ix
LIST OF SYMBOLS AND ABBREVIATIONS	x
LIST OF FIGURES	xi
LIST OF TABLES	xii
1. INTRODUCTION	1
1.1. General Knowledge of Air Traffic Control Management Systems	3
1.2. Situation Awareness	6
1.2.1. Definition of Situation Awareness	6
1.2.2. Measurement of Situation Awareness	8
1.2.2.1. SART	9
1.2.2.2. CARS	11
1.2.2.3. SAGAT	11
1.2.3. Eye Tracking	12
1.2.3.1. Eye Tracking Metrics	14
1.2.3.2. Positive Correlations with Eye Tracking Metrics and SA	17
2. PURPOSE OF THIS STUDY	19
2.1. Contribution of Thesis	19
2.2. Overview	20
3. EXPERIMENT	21
3.1. Hypothesis	21
3.2. System Design	21
3.3. Simulation	22
3.4. Eye Tracker	23
3.5. Methodology	26
3.5.1. Participants	28
3.5.2. Scenario	29
3.5.3. Collision Algorithm	30
3.5.4. SAGAT Queries	33
3.5.5. Performance Parameters	38
3.5.6. Eye Tracking Metrics Used	39
4. RESULTS	45
4.1. Performance Measurements, SAGAT and Eye Tracker Correlation Evaluation Using Paired Sample t-Test	55
4.2. Group-Based Analysis of Intervals Using One-Way ANOVA Analysis	62
5. CONCLUSIONS	67
REFERENCES	70
BIOGRAPHY	75
PUBLICATIONS AND PRESENTATIONS FROM THE THESIS	76
APPENDICES	77

LIST OF SYMBOLS AND ABBREVIATIONS

APW	: Air Proximity Warning
ATC	: Air Traffic Control
atcTRsim	: Air Traffic Control Simulator
BİLGEM	: Bilişim ve Bilgi Güvenliği İleri Teknolojiler Araştırma Merkezi
CARS	: Crew Awareness Rating Scale
DHMI	: Devlet Hava Meydanları İşletmesi
JSON	: JavaScript Object Notation
MSAW	: Minimum Safe Altitude Warning
SA	: Situational Awareness
SAGAT	: Situation Awareness Global Technique
SART	: Situation Awareness Rating Technique
SPAM	: Situation Present Assessment Method
STCA	: Short Term Conflict Alert
TÜBİTAK	: Türkiye Bilimsel ve Teknolojik Araştırma Kurumu

LIST OF FIGURES

	<u>Page</u>
Figure 1.1: Air traffic controllers working in a control tower [Labor et al., 2000].	4
Figure 1.2: Endsley's situational awareness model (Waidyanatha, 2020).	7
Figure 1.3: Different types of SA measurement techniques.	8
Figure 1.4: Tobii TX300 eye tracker device with removable 23" monitor.	14
Figure 1.5: One student's eye movements while examining a chest x-ray. Fixations are represented as circles and the duration of the fixations are represented by size. Saccades are represented as the lines between the circles [Kok and Jarodzka, 2017].	15
Figure 1.6: Example of eye movements on aircrafts. The circles are fixations; the size of the circle represents its duration. The lines between the circles are saccades.	16
Figure 3.1: Graphical representation of experiment environment.	22
Figure 3.2: atcTRsim project environment in Ankara (Web 5 2023).	23
Figure 3.3: Tobii X2-30 eye tracker device.	25
Figure 3.4: Thesis simulation where the participant is instructed to monitor colliding aircrafts.	26
Figure 3.5: System block diagram of the current thesis study.	27
Figure 3.6: Experiment where the participant is instructed to monitor colliding aircrafts.	28
Figure 3.7: a) Marked user interface of system design. b) Simulation scenario flowchart of user interface.	29
Figure 3.8: a) Marked aircraft collision detection of system design. b) Aircraft collision detection algorithm flowchart.	31
Figure 3.9: Representation of close proximity alarm colored in a red foreground.	32
Figure 3.10: Representation of close proximity warning with aircrafts colored in a yellow foreground.	32
Figure 3.11: Situation awareness level 1 question: region.	35
Figure 3.12: Situation awareness level 2 question: last resolved aircraft.	35
Figure 3.13: Situation awareness level 2 question: last conflicted aircraft.	36
Figure 3.14: Situation awareness level 3 question: future aircraft collision.	37
Figure 3.15: Correct and missed click algorithm flowchart.	39
Figure 3.16: Fixation geospatial coordinate conversion JSON file.	41
Figure 3.17: Aircrafts updated position coordinates JSON file.	42
Figure 3.18: Aircraft and fixation correspondence algorithm flowchart.	43
Figure 4.1: Cumulative performance measurements and SAGAT score result.	47
Figure 4.2: Cumulative performance score, SAGAT score and eye tracker score results.	48
Figure 4.3: Cumulative clicks and fixation count result.	50
Figure 4.4: Cumulative total fixation count result.	51
Figure 4.5: Results of cumulative SAGAT questions' accuracy based on the situational awareness levels.	52
Figure 4.6: Results of individual performance, SAGAT and eye tracker score.	54

LIST OF TABLES

	<u>Page</u>
Table 3.1: Participant information.	28
Table 3.2: SAGAT questionnaire inspired by Endsley.	34
Table 3.3: Example of exported excel file by Tobii Studio.	40
Table 4.1: Descriptive statistics of the experiment.	46
Table 4.2: Cumulative performance score and SAGAT score result.	48
Table 4.3: Cumulative performance score, SAGAT score and eye tracker score result.	49
Table 4.4: Cumulative clicks and fixation count result.	50
Table 4.5: Descriptions of the SAGAT questions and SA levels.	53
Table 4.6: Cumulative SAGAT question accuracy results.	54
Table 4.7: Paired sample correlations.	58
Table 4.8: Paired sample hypothesis test result.	59
Table 4.9: Test of homogeneity of variances for intervals.	62
Table 4.10: One-Way ANOVA test descriptive statistics.	63
Table 4.11: Results of one-way ANOVA test for the 8-minute intervals.	64
Table 4.12: Tukey post-hoc test for correct clicks.	65
Table 4.13: Tukey post-hoc test for target hit fixation count.	66

1. INTRODUCTION

In human-machine contexts, Situation Awareness (SA) may have a direct impact on human performance and decision-making [Zhang et al., 2023]. Majority of accidents are caused by human mistakes. According to the National Transportation Safety Board survey, it was discovered that 88% of accidents in the aviation field were due to errors made by humans [Jones and Endsley, 1996]. The International Maritime Organization (IMO) found that 85% of the accidents in the maritime industry were also caused by the loss of situation awareness. Additionally, the construction sector was responsible for nearly 20% of worker fatalities [Martinez-Marquez et al., 2021]. Situation Awareness is essential for all air traffic controllers who do not possess the luxury of making errors [Moore and Gugerty, 2010; Martinez-Marquez et al., 2021] but it is not limited to only life critical jobs such as ATC. The safety of ships in the maritime domain, safety of the workers in construction sites, and maintaining the health of patients while they undergo extensive surgery can all be taken into account of the importance of SA.

Pilots rely on air traffic controllers who are highly qualified professionals, to assist them to land safely and keep the flow of air traffic safe. Workplaces for air traffic controllers include route centers, approach control facilities, and control towers [Labor et al., 2000]. A controller's duty requires regular observation of radar systems, runways, and weather screens. If the controllers are stationed in a tower, they must also keep an eye out for departing and arriving aircraft from the tower windows. Because of the extreme concentration required, this constant monitoring can be a heavy workload and stressful task [Labor et al., 2000].

In Zhang et al. (2023) literature review where the authors investigated the relationship between direct and indirect SA methods, it was discovered that the aviation industry was the basis for more than half of the analyzed studies. Even though the area of investigation is aviation, this does not limit other fields such as maritime, education, driving and medicine.

Air traffic control management systems produce different alarms such as STCA, MSAW and APW alarms during dangerous situations. The presentation of these alarms is created using the EUROCONTROL guidelines [Web 1, 2017; Web 2, 2017;

Web 3, 2017]. Even though controllers are made aware that a dangerous situation is taking place by these visual and audio alarms, there is always a risk of the controller not being conscious of or aware of his/her surroundings. By measuring situation awareness of controllers while on the task or during training, appropriate safeguards may be implemented to raise awareness and reduce the possibility of a catastrophic disaster.

Situation Awareness measurement techniques can be divided into three categories: subjective, indirect, and direct techniques [Xinyao et al., 2020]. Measuring SA can be done using different methods such as Situation Awareness Global Technique (SAGAT) [Endsley et al., 1988], Situation Awareness Rating Technique (SART) [Taylor, 2017], Situation Present Assessment Method (SPAM) [Durso et al. 1998], Crew Awareness Rating Scale (CARS), and eye tracking [McGuinness and Foy, 2000; Moore and Gugerty, 2010; de Winter et al., 2019]. Through the use of questions during simulation freezes, SAGAT enables researchers to measure SA. SART and CARS are post-trial subjective rating techniques that are applied after the participant has executed the task. The premise behind SPAM is that sometimes, SA is more about knowing where to seek for information in the surroundings than it is about remembering what that information is [Durso et al., 1998]. Eye tracking, on the other hand, can also be used to assess situation awareness by measuring the movement of the eyes and then statistically analyzing the resulting data [Kok and Jarodzka, 2017]. Eye tracking has been shown in several research to be a valid approach of detecting SA without being obtrusive like SAGAT [Moore and Gugerty, 2010; Van De Merwe, Van Dijk and Zon, 2012; Hasanzadeh, Esmaeili and Dodd, 2016; de Winter et al., 2019].

In this study, the atcTRsim project that was developed by DHMİ and TÜBİTAK BİLGEM in 2010 was used to carry out the simulation. In previous studies, taking into account the aviation domain has also used different simulations such as the “TRACON II Air Traffic Control Simulator” in Moore and Gugerty’s study (2010). The atcTRsim project’s Airspace Editor application was used to create an air traffic simulation. This application normally is used to create airspaces that are needed for air traffic controllers to reference sectors, special use airspaces, fixes, routes, procedures and other important airspace components. By automatically generating random aircrafts and updating their positions every 1 second, the replication of flying aircrafts on a radar screen was made.

In the experiment the participants are responsible for detecting aircraft that are close to colliding. This detection can be made by observing the altitude and horizontal distance of two aircraft. If the altitude difference is below 10 Flight Level and the horizontal distance is less than 40 nautical miles, then this indicates that the aircrafts are at risk for colliding. At the time of this risk, the aircraft are presented with a “Close Proximity Warning” alarm with the aircrafts foreground being yellow. If the altitude difference is below 10 Flight Level and the horizontal distance is less than 20 NM then this indicates that the aircrafts are in serious danger and that immediate action must be taken place. When this situation takes place, aircraft are presented with a “Close Proximity Alarm” alarm with the aircrafts foreground being Red. The user must click on the aircraft that seem to be at risk and this needs to be done before any warning alarms are presented to the user. For every successful save of a potential colliding aircraft pair, the user’s correct click count is incremented. For every missed collision, then the user’s missed click count is incremented. During the experiment the user is asked a series of questions based on the SAGAT technique regarding the simulation. At random intervals the user is shown a full screen question forum which hides the simulation in the background. The user is given 1-2 minutes to answer the questions. Simultaneously the user’s eye movements are tracked by the Tobii X2-30 eye tracker. The duration of the experiment is 32 minutes with 8 minute intervals.

1.1. General Knowledge of Air Traffic Control Management Systems

Those with the training to maintain a safe flow of air traffic are known as air traffic controllers in the global air traffic control system [Web 4, 2018]. In their area of responsibility, air traffic controllers utilize a range of regulations to prevent airplanes from colliding with one another and to safely and efficiently navigate all aircraft through their assigned sector of airspace as well as on the ground. They do this by communicating instructions, advice, and information to pilots via radio frequency, working with various assisting divisions, and utilizing technological advancements (Figure 1.1).



Figure 1.1: Air traffic controllers working in a control tower [Labor et al., 2000].

Aircraft are not allowed to take off and fly at any given moment without complying to regulations and order given by the air traffic controller. The aircraft must follow the orders given by the air traffic controllers and fly along the pre-planned air routes. As a result, the pilots must request permission and instructions from the air traffic controllers before beginning an engine, taking off, flying across airspace, climbing, descending, adjusting speed, landing, and operating an aircraft on the ground. Many airplanes may be under the direction of one air traffic controller at once.

Although the responsibilities of air traffic controllers may vary depending on different areas such as tower, approach and ground, the essential roles can be listed as follows:

- Above everything else, ensure the safety of the flight.
- To provide descend and climb instructions in an orderly fashion and ensure safe flow of air traffic.
- To make sure aircraft maintain a safe distance from each other, ground vehicles and obstructions.
- To inform aircraft of all relevant information, such as weather conditions and anticipated approach procedures.
- To manage unforeseen circumstances, irregular traffic, and emergencies, as well as to coordinate with search and rescue teams.

According to the Directorate General of Civil Aviation's annual report, in Turkey as of the end of 2022, the total number of passengers exceeded 182.3 million. The number of domestic passengers increased by 14.9% compared to the previous year; the number of international passengers increased by 73%. As of the end of 2022, the total aircraft traffic was 1,883,186. Domestic aircraft traffic's percentage increased by 6.9% compared to the previous year; international air traffic by 49.9%; and transit aircraft traffic increased by 50.6% [Sivil Havacılık Genel Müdürlüğü, 2022].

With the increase of air traffic in both domestic and international flights, the mental workload for air traffic controllers inevitably also increases. A survey conducted by the National Transportation Safety Board revealed that human error was the cause of 88% of aviation-related incidents [Jones and Endsley, 1996]. Overall, humans make up the largest percentage of the causes of airplane accidents.

An example of losing situation awareness can be given regarding the plane crash of TransAsia Airways Flight 222 in 2014 [Salotti and Suhir, 2019]. Rainy weather and low visibility were present during the descent and landing. The crew finally turned off the autopilot after deciding to fly lower than the minimum permitted altitude. The crew concentrated on looking for the runway visually as the aircraft approached its destination, but due to a slight change in the trajectory, they missed it. The plane eventually crashed into a hill behind the airport because they did not pay enough attention to the altitude. "Controlled flight into terrain" was the accident's classification. When the corporation was questioned, it came to light that many teams were utilized to descend below the permissible limit when visibility was poor. It was also discovered that the crew had too many flights to handle in a short amount of time. Therefore, exhaustion played a role. This incident falls under the SA concerns of attentional tunneling and tiredness.

Another aviation accident that can be analyzed is the Turkish Airlines Flight 6491 [Salotti and Suhir, 2019]. The trip took off on January 16th, 2017. The destination airport's meteorology report indicated thick fog. When the "decision height" was very low, the crew sought to land using instruments. In air traffic control, if the pilot cannot land using visual rules, then he/she can land using instrumental devices that acts as the pilot's eyes. They decided to perform a go-around as soon as they reached the minimal height without visibility, but the altitude was too low, and the jet crashed just behind the runway. The crew failed to decide on and adhere to the proper landing technique

for that particular plane on that particular runway in those particular meteorological conditions. Therefore, the human error happened early in the planning stage. When there are numerous diverse scenarios, it is challenging to determine the proper decision height from charts and landing procedures.

There are many more accidents that have occurred because of human factors and the lack of situation awareness. Evaluating this type of awareness in human beings that work in safety critical jobs such as aviation, maritime, medicine or construction is crucial. If the person is not fully mentally present while managing their task, then this can be detected, and proper actions can then be taken to improve situation awareness whether it be reconstructing the software applications the occupants use or execute other methods.

1.2. Situation Awareness

1.2.1. Definition of Situation Awareness

The definition that is frequently used for SA is “the perception of elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future” [Endsley, 1988]. The ability of an evaluator to accurately assess and comprehend the context and surrounding situation is known as situational awareness [Garcia, 2016]. Only through focusing on, perceiving, and comprehending one's environment can one acquire SA [Martinez-Marquez et al., 2021]. In the fields of aviation, maritime, and construction, SA is the fourth most researched topic. A person's capacity to recognize errors depends on their capacity to notice changes, which is directly tied to situation awareness [Hasanzadeh, Esmaili and Dodd, 2016; Behrend and Dehais, 2020; Martinez-Marquez et al., 2021].

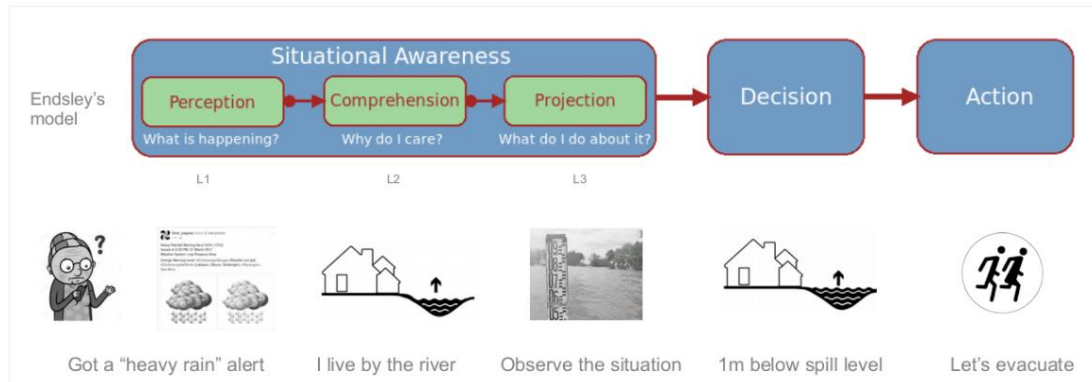


Figure 1.2: Endsley's situational awareness model (Waidyanatha, 2020).

According to Endsley [Endsley et al., 1988], perception (Level 1), comprehension (Level 2) and projection (Level 3) are key values for determining SA (**Figure 1.2**). The distinctions between these three levels are significant because they reveal variations in perceptual and cognitive processes ranging from basic perceptual (Level 1) to more complex cognitive (Level 2) processes. Higher SA (mental processing) levels are dependent on lower levels (perception of information). The three-level hierarchy structure can also be defined as so [Zhang et al., 2023];

- Level 1: Perception of the surroundings by an individual over a volume of time and place.
- Level 2: The individual's comprehension of the significance of the surrounding elements.
- Level 3: The projection of the environmental element's status in the near future.

In the aviation domain, for example, Indicated Air Speed (IAS) differences between the right and left main flight displays (PFDs) may be seen by a pilot, who may then recognize that this could be a sign of faulty equipment that could have an impact on flight performance [van Dijk, van de Merwe and Zon, 2011]. Another example of this three-level hierarchy can be given in the context of driving. Such a scenario can be used to analyze the levels of SA of an approaching car at a four-way stop intersection. The Level-1 SA relates to recognizing any cars on the other three streets, the Level-2 SA refers to the driver's comprehension of the order in which each vehicle would arrive at the crossing, and the driver's choice of when to cross the intersection is covered by Level 3 SA.

1.2.2. Measurement of Situation Awareness

Various methods have been developed to assess an individual's level of situation awareness. In some literature reviews, these measurements are grouped into two categories such as direct and indirect [Zhang et al., 2023] or are grouped into three categories (**Figure 1.3**) such as subjective techniques (e.g. observer-rating and self-rating techniques), indirect physiological measurements techniques (e.g., performance measures, physiological measures such as eye tracking) and direct techniques (e.g., freeze-probe techniques). There is also literature that divides these methods into four categories such as subjective techniques, performance measures, memory probe measures, and physiological techniques [Xinyao et al., 2020]. Specifically, direct measurement techniques gather an individual's perceptions. One way to get knowledge about a situation is by asking questions or making observations, which provides a direct understanding. Alternatively, indirect approaches may be used to infer situational awareness by analyzing the user's physiological reaction, behavior, and task performance.

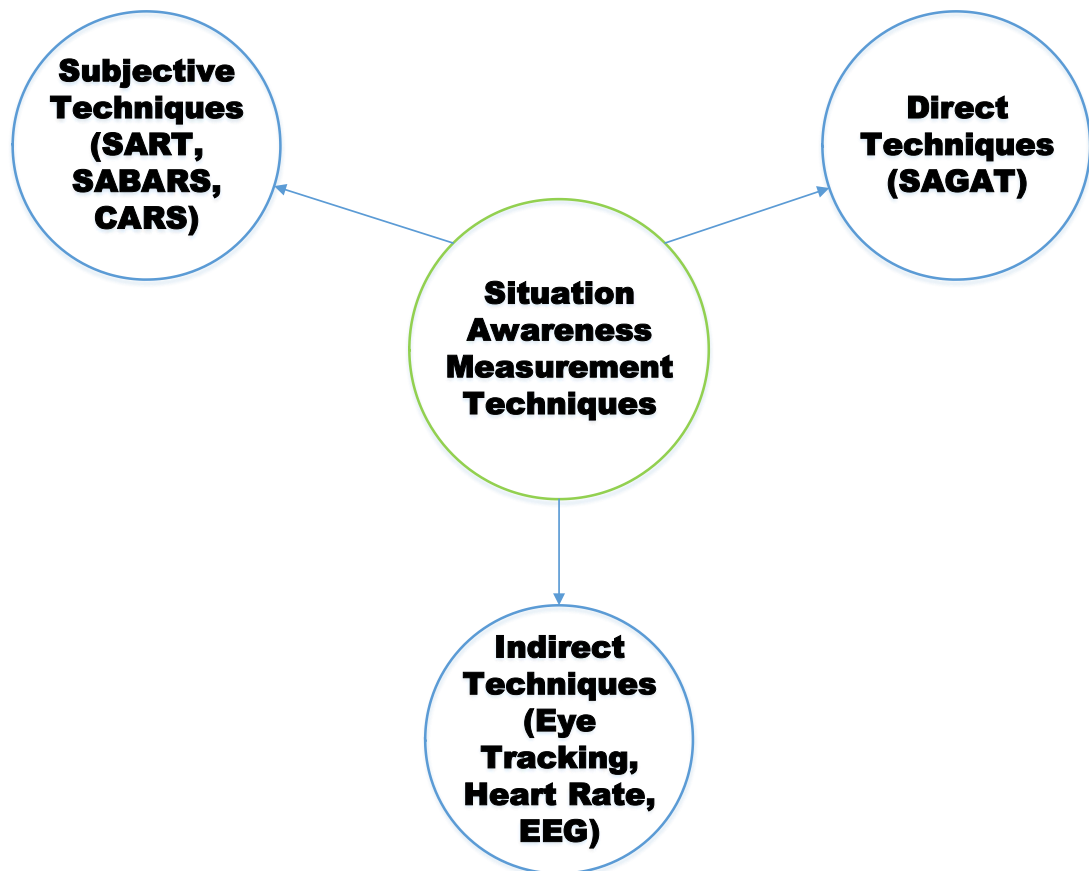


Figure 1.3: Different types of SA measurement techniques.

Subjective techniques are considered one of the most straightforward SA measurements. The SA measurement is produced by asking a domain expert to rate their own level of SA [Durso et al., 1998]. The commonly used self-ranking methods are the “Situational Awareness Rating Technique” (SART) and the SA behavioral rating scale (SABARS) [Matthews and Beal, 2002].

Direct measurement techniques can be labeled as query methods such as SAGAT which is an objective SA measurement. This scenario-based technique assesses SA by inquiring questions as the task at hand is being completed. Indirect measurements entail incorporating all relevant data into a realistic situation and evaluating the participant's SA based on how their behavior changes in that scenario [Hasanzadeh, Esmaeili and Dodd, 2016]. However, some researchers claim that analyzing performance is not an effective technique to learn about SA [Wickens, 1996]. Regarding physiological measuring methods, eye tracking was most frequently employed to evaluate SA. Eye tracking methods can be used to measure SA indirectly by using the collected metrics such as by examining eye metrics such as fixation count, fixation rate, and dwell time.

Each of these categories has its own drawbacks. For subjective techniques, because no two people have the same knowledge of a topic, it is impossible to ensure that the results of subjective procedures will be consistent among subjects. For direct measurement techniques, these methods can be intrusive and highly reliant on memory [Durso et al., 1998]. Indirect measurement techniques such as eye tracking also need more investigation of the reliability of the technique.

1.2.2.1. SART

“The situation awareness rating technique” (SART) [Taylor, 2017] was initially created for the evaluation of pilot SA. It is a straightforward post-trial subjective rating technique. Operator SA is determined by SART using the ten dimensions listed below: attentional focus, situational familiarity, information quality, information quantity, concentration of attention, situational instability, variability of the situation, arousal,

complexity of the situation, and spare mental capacity. There is also a quicker variation of the SART method called the 3D SART.

Following a trial or experiment, participants are asked to subjectively rate each dimension on a scale measuring seven points (1 being the lowest, and 7 being the highest) depending on how well they performed the task being studied. The ratings are then added together to derive a participant SA measure. The SART score can be calculated using the following formula: $SA = \text{Understanding} - (\text{Demand} - \text{Supply})$. The advantages of SART provide a generic model that can be used in other domains. It doesn't interfere with task performance and is quick and simple to administer. The drawback of SART is that it can cause memory degradation and poor recall because the questionnaire is presented to the participant after the trial or experiment. SABARS, similar to SART being a rating technique differs in the way that the participant's behavior is observed by a domain expert instead of the participants rating themselves.

1.2.2.2. CARS

The CARS is a post-trial questionnaire like SART that assesses four cognitive processes associated with SA: perception of pertinent information from outside sources, understanding of the significance on this information, prediction of potential future changes to the situation, and decision-making regarding the best course of action to take in order to attain a desired result. Eight questions make up the survey, two for each SA dimension. Each inquiry evaluates SA on a scale of 1 (very well) to 4 (very poor) [Van De Merwe, Van Dijk and Zon, 2012].

1.2.2.3. SAGAT

Endsley et al. (1988) created the Situation Awareness Global Assessment Technique, a query technique method. The foundation of SAGAT is information processing theory. Prior to making decisions and performing, Endsley et al. (1988) regards situation awareness as a model that is internalized and derived from the environment. One of the most well-publicized and well-known SA metrics is SAGAT [Center, E. E, 2000]. Using a freeze-probe technique called SAGAT, a simulation is paused at random intervals, and participants are then asked questions regarding the ongoing task. When there is a freeze period the screen is cleared (e.g., a blank screen). The SA is divided into three levels: perception, comprehension, and projection correspond with the questions. These questions can be asked and answered via pen and paper [de Winter et al., 2019] or on screen [Argyle et al., 2020]. Following task completion, the responses from the participants are compared to the simulation's real events. The total accurate response rate derived from questions is called a SAGAT score. Higher accuracy rates, or SAGAT scores, signify a higher degree of match between the circumstance and awareness of it [Zhang et al., 2023].

According to Endsley (1988), it's advised to freeze the scenario at random times so that the users cannot predict and prepare for the questions that are to come ahead. At each stop, a random pick from a fixed set of queries is advised. There is an informal rule that the freezes should not occur earlier than three to five minutes during the trial for the participants need time to adjust to the occurring experiment. Within each trial, multiple freezes can occur. During a 15-minute study, freezes that happened three times had no negative consequences. Freezes of up to two minutes in length have been employed without significant difficulty or impact on performance. Freezes can also

last up to five or six minutes without memory deterioration [Endsley and Garland, 2000]. Answers to each question should be rated as correct or incorrect depending on whether they fit within a reasonable tolerance range around the actual value. It might be appropriate, for instance, for a participant to be 10 MPH off than actual ground speed.

SAGAT is considered to be one of the most extensively validated SA techniques due to its proven reliability [de Winter et al., 2019]. It is typically necessary to have computers that can produce and assess SAGAT inquiries as well as a simulation that highly models the task being studied. The simulation and computer should be capable of randomly blanking all "window" displays, distributing pertinent SA inquiries at random, and computing subject SA scores. The simulation can be a scenario created for air traffic controllers, drivers, medical students etc. In the study done by van den Beukel and van der Voort (2017), driving systems that are partially automated were examined which altered the role of the driver. With this automated driving system, the driver's role was to oversee the automation while still having the ability to intervene on occasion. In the research stated, an evaluation methodology is proposed and analyzed to allow designers to assess driver-support in relevant real-world situations and provide interface solutions that help drivers in this new role. In the driving simulator experiment, SAGAT is used as a measurement method.

SAGAT is a very helpful technique in design review because this technique gives diagnostic data on how effectively the system in question fulfills the individual's varied SA criteria. While it has benefits like assessing the user's SA, it also has drawbacks including task disruptions and interruptions, deviating from the current task, and requiring the user to bring the current circumstance to conscious recollection, which can cause memory deterioration [de Winter et al., 2019]. In the real world there are tasks that cannot be halted so this method might not be feasible for all situations.

1.2.3. Eye Tracking

Direct SA tools have their limitations and because of this SA prediction can be improved by utilizing sensors that obtain behavioral and physiological trends. These measurements may be utilized to infer numerous SA-relevant constructs indirectly and in real time [Zhang et al., 2023]. The non-intrusive continuous nature of eye

movements is a significant advantage of studying them. Eye activity is fascinating to investigate for three reasons. The first reason is being that the eye activity can be measured without interfering with the current task at hand. Secondly, the action is continually measured and thirdly, it is objectively measured [Van De Merwe, Van Dijk and Zon, 2012].

With the use of an infrared light source and an eye video camera, most eye tracking systems in use today are video-based. Different video-based eye tracking systems are needed, depending on the activity being examined, the experiment, and the surroundings. Based on how they interact with the user and surroundings, eye tracking systems can be classified as mobile, tower-mounted, or remote [Martinez-Marquez et al., 2021]. Mobile eye tracking can also be eluded as wearable devices (eye tracking glasses) or head-mounted. Remote eye trackers measures eye movements from a distance and do not have to be touched by the person. An example of a remote or table eye tracking device can be seen in **Figure 1.4**.

Previous studies have demonstrated that eye movements can serve as indications of mental workload and SA [Glaholt, 2014]. Limitations that SAGAT has can be prevented by measuring eye movements. With eye movement tracking for continuous SA measurement, the user can work uninterrupted on the current task [Xinyao et al., 2020]. An additional benefit of employing eye movement as a measurement for SA is that it eliminates the need for the user to recall the task's aspects from conscious memory, hence reducing the risk of memory decay [Kok and Jarodzka, 2017; Wee et al., 2017].

There are also limitations of eye tracking methodology including tracking loss and sensitivity to a participant's movement. If participants wear narrow glasses, hard contact lenses, or have droopy eyelids, calibration may be challenging [Wee et al., 2017; Xinyao et al., 2020]. With faulty calibration, the data can be of poor quality and must be discarded. The ability to directly deduce cognitive processes from eye movements are also still poorly understood [Glaholt, 2014]. For example, a student is listening to a lecture and is fixated on the slides. The point where the student is fixated on may be due to the colorful illustrations of the slides, the eagerness to learn from the course or the tendency for people to center their attention on the center of a screen [Kok and Jarodzka, 2017].



Figure 1.4: Tobii TX300 eye tracker device with removable 23” monitor.

1.2.3.1. Eye Tracking Metrics

Eye tracking research have identified and related different eye metrics to different emotional, physiological, and cognitive states that can be utilized to gain a deeper knowledge of the human psyche. Fixation, saccades, pupillary response, and blink rate make up these eye metrics.

An eye metric known as "fixation" occurs when the eye stays still and fixed for between 180 and 300 milliseconds [Martinez-Marquez et al., 2021]. A dwell is a collection of fixations on one specific piece of information within a display [Glaholt, 2014]. People learn new information while fixating on an object, stimulus, or location. Different fixation metrics include fixation rate, fixation count, fixation duration, and fixation length. A study in the field of aviation research found that experienced pilots focus more of their attention on a range of flight instruments than they do on any one of them for a prolonged period of time [Ziv, 2016]. In the study conducted by Hasse, Grasshoff and Bruder [Hasse, Grasshoff and Bruder, 2012], it was discovered that

failure detection frequency is higher for air traffic controllers who have focused more intently on pertinent area of interest.

Saccades are quick eye movements that occur when someone alternates between fixations. Since visual information transfer is blocked during saccades, which last between 10 and 100 milliseconds. Saccades and cognitive functions are not directly related [Häggström, Englund, Lindroos, 2015]. However, apathy, tension, and fatigue can all be connected to saccade velocity. For instance, the rate of saccades decreased with fatigue and challenging activities [Yan and Tran, 2019]. The visual representation of saccades and fixation is presented in **Figure 1.5** and **Figure 1.6**.

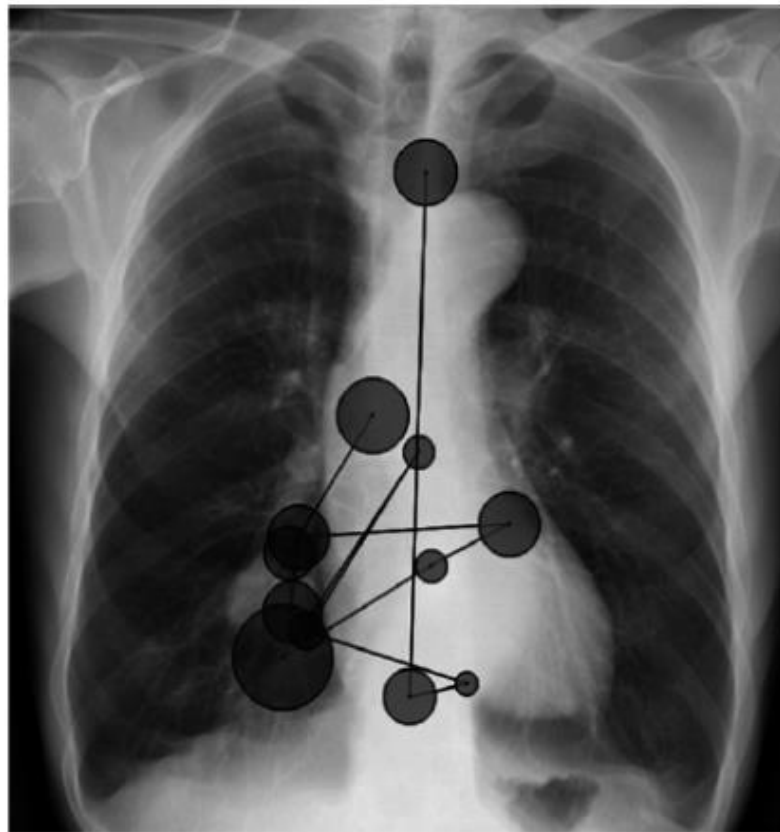


Figure 1.5: One student's eye movements while examining a chest x-ray. Fixations are represented as circles and the duration of the fixations are represented by size. Saccades are represented as the lines between the circles [Kok and Jarodzka, 2017].

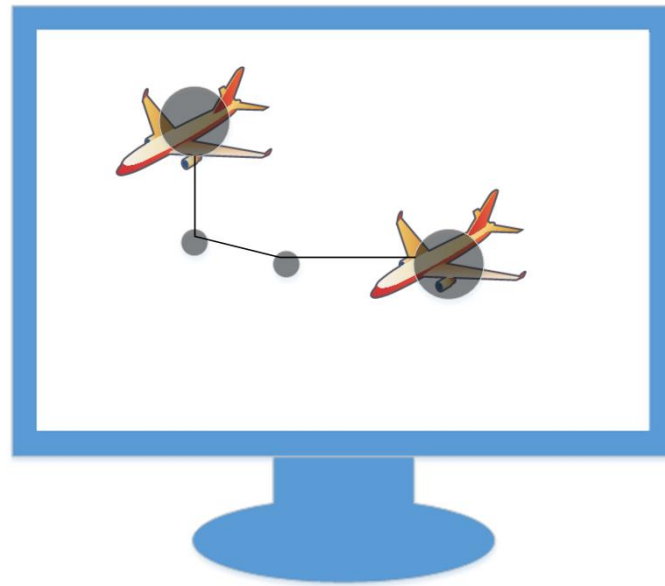


Figure 1.6: Example of eye movements on aircrafts. The circles are fixations; the size of the circle represents its duration. The lines between the circles are saccades.

Pupil size and the location of its center can be measured using video recordings. Lighting has an impact on pupil size, which controls how much light reaches the retina. Emotions, muscle exhaustion, cognitive processes, and mental workload all have an impact on pupil size [Li et al., 2019]. For instance, in the maritime domain, a greater pupillary dilation response has been linked to a greater mental processing burden [Bjørneseth et al., 2014; Orlandi and Brooks, 2018]. Additionally, the cognitive impacts of stress in aircraft incidents and combat have been studied using pupil size [Martin, Cegarra and Averty, 2011; Stankovic, Aitken, Clark, 2014].

Blink rate is defined as the number of blinks per second or minute, and it is influenced by physiological factors like mood, task demands, attentiveness, and stress. Mental workload and fatigue can be measured using blink rates. For instance, low mental workload and fatigue are both associated with high blink rates [Muczyński and Gucma, 2013]. As seen in various behavioral researches, eye movement metrics can offer a more accurate indicator of workload. Eye movements may be used to collect extra data and even more precise measurements of construction, maritime, and aviation-related topics. Eye movements may provide a more comprehensive view of the operator's workload, knowledge, and situational awareness when combined with performance metrics [Glaholt, 2014].

1.2.3.2. Positive Correlations with Eye Tracking Metrics and SA

Fixation count, dwell count, fixation rate, and dwell time eye metrics were all found to be positively correlated with situation awareness. In the study conducted by van de Merwe, Van Dijk and Zon (2012), before and after a system failure that was induced, the CARS scores and eye tracking data were compared. The CARS score showed that all SA levels had higher SA in the post-malfunction period compared to the pre-malfunction era. Eye tracking demonstrated that when the problem was displayed, operators' focus migrated from other displays to the “electronic centralized aircraft monitoring display” (ECAM), as shown by a higher fixation rate and longer dwell time on the ECAM. As a result, it can be inferred that SA, dwell length, and fixation rate as evaluated by CARS may be positively correlated.

Regarding fixation counts and dwell times, in one of four experimental settings, Hasanzadeh et al. Hasanzadeh, Esmaeili and Dodd's [Muczyński and Gucma, 2013; Hasanzadeh, Esmaeili and Dodd, 2018] found that feedforward-information AOIs had a greater fixation count and a longer dwell period. In the second scene, participants with higher SART scores had longer dwell durations in a different kind of environment related AOIs. In Hasanzadeh, Esmaeili and Dodd's (2018) earlier work, the findings in the construction domain demonstrated that it is more probable for workers to miss crucial details that could aid in situation perception if they spend too much time on one task or are preoccupied with something unimportant. The findings supported the hypothesis that employees with low levels of SA were too fixated on the AOI and hence missed additional threats present nearby. However, those with high SA levels focused less on a single issue and spent more time observing their surroundings to fully grasp the circumstances.

In the study conducted by Moore and Gugerty (2010), by using a direct measure of SA that was similar to the SAGAT, it was found that there was a positive link between the percentage (%) of time spent fixating on an AOI and both future and current SA. Even though no link between direct SA measurements and eye tracking metrics was identified in other experiments, researchers assert that rather than using direct methods like the SAGAT, eye tracking measurements may be more helpful in predicting performance in SA tasks.

In contrast to the SAGAT score, which had a significantly lower correlation with performance evaluations, the performance score was found to be positively associated to visual sampling score, according to the study made by de Winter et al. (2019).

Although the subject of measuring situation awareness (SA) with eye tracking technology is expanding, there are still a number of gaps in the literature. Even though eye tracking offers a wealth of objective, real-time data on visual attention and cognitive processes, it is still difficult to interpret eye movements in connection to SA. The absence of established measurements for directly connecting levels of SA to patterns of eye movement is one notable gap. Furthermore, little is known about how varied tasks and circumstances affect the eye movement indicators of SA, which makes it difficult to generalize results across multiple domains. On the contrary, the widely recognized SAGAT asks participants questions on how they perceive, understand, and project situational variables at random intervals, thereby offering a direct measure of SA. However, SAGAT is intrusive and it may introduce biases and interfere with the task's natural flow, changing the very SA that it is intended to measure.

2. PURPOSE OF THIS STUDY

In complex human computer interaction systems, it's essential for users to be aware of their present situation. Although this study focuses on aviation, it can also be implemented by various domains that demand attention and situation awareness. To evaluate a person's situation awareness, it needs to be measured and there are a few techniques such as SAGAT, SART, CARS, etc.... that serve this purpose.

The aim of this thesis is to assess situation awareness of air traffic controllers in the aviation domain using SAGAT the well-known SAGAT technique and eye tracking, a more unconventional technique. To elaborate more on the goal, we would like to statistically prove that evaluating a person's situation awareness by using eye tracking is a more effective technique than the more intrusive SAGAT technique, contribute to existing literature that has found positive correlations of eye tracking methodologies and situation awareness, and strengthen the need of measuring situation awareness for occupations that demand attention during heavy workload tasks such as air traffic controllers.

2.1. Contribution of Thesis

This study compares the eye tracking and SAGAT approaches for assessing SA in an effort to fill in the gaps in the literature. Through an organized investigation of the advantages and disadvantages of both methods, the research will advance a more sophisticated comprehension of the ways in which eye movements and SA are related in diverse settings. Standardizing measures for eye tracking in SA research and figuring out the accuracy that eye movement data may provide as a reliable alternative for SA without the invasive consequences of SAGAT are two examples of what this entails.

The contributions of this study provide resources that evaluate situation awareness techniques in the aviation field as well as decrease the possibility of catastrophic accidents because of low situation awareness resulting in human error. This study will also highlight information on the contextual and situational elements that affect the validity of eye tracking as a SA measure, which will be crucial knowledge for future applications in a variety of industries, including healthcare, aviation, and military

operations. In the end, the research will advance the creation of less invasive and more efficient SA measurement tools, improving the precision and usefulness of SA evaluations in practical contexts.

2.2. Overview

This study gives insight into different situation awareness measuring techniques and provides the result of how well eye tracking measures situation awareness compared to SAGAT. This study also puts emphasis on how important situation awareness is in critical situations tasks.

The first part of this study provides a background of aviation and the significance of situation awareness. Various situation awareness measuring techniques are also explained as well as providing information on the positive correlations found between situation awareness and eye tracking methodologies. The second part states the contribution of this thesis to literature related to situation awareness. The third part discusses the hypothesis and the methodology of the conducted experiments. The fourth part provides statistical descriptions of the results and the correlation between situation awareness and performance. The last section of this thesis discusses the findings of the experiment and concludes the observations made.

3. EXPERIMENT

3.1. Hypothesis

It has been seen that there are positive correlations between SA and eye tracking metrics but there needs to be more studies conducted to prove the validity of using eye tracking methods to accurately measure a person's situation awareness. The thesis will hopefully contribute to the works related to situation awareness being measured by eye movements.

- The hypothesis was “Eye movement (tracking) techniques provide a better measurement of situational awareness in Air Controllers than probe related techniques such as SAGAT”.
- The null hypothesis of the study is that there will be little to no correlation between eye movements and situation awareness and that SAGAT will prove to measure situation awareness in a better manner.

3.2. System Design

The simulation took place on a 25” laptop with a 1920 x 1080 resolution. The simulation was created to resemble live air traffic that is usually displayed on controller's radar monitors. The monitor that is usually used by controllers is 28” with a 2K x 2K resolution. The Tobii X2 – 30 eye tracker is placed just below the screen and above the keyboard. The layout of the experiment setup can be seen in **Figure 3.1**.



Figure 3.1: Graphical representation of experiment environment.

3.3. Simulation

The simulation that was used is an application that is part of the project atcTRsim (Air Traffic Controller Simulator) developed by TÜBİTAK BİLGEM and DHMİ (Devlet Hava Meydanları İşletmeciliği). atcTRsim is a comprehensive simulation system that satisfies both basic and advanced ATC training requirements. Support is provided for Approach, En-route, and Tower ATC trainings at all levels (**Figure 3.2**). atcTRsim is a great platform for research activities as well as training activities because it is easily adaptable to local conditions. In a real air traffic control system, there are systems that manage flight, radar, and trajectory data such as Radar Data Processing System (RDPS), synthetic traffic generator (TGEN), and Flight Data Processing System (FDPS). These modules have also been developed to be a part of atcTRsim.



Figure 3.2: atcTRsim project environment in Ankara (Web 5 2023).

3.4. Eye Tracker

The eye tracker that was used in the experiment is the Tobii X2 – 30 (**Figure 3.3**). Tobii is one of the global leaders in the eye tracking industry and is trusted by many world leading institutes and companies. Many studies [Moore and Gugerty 2010; Ratwani, McCurry and Trafton 2010; Aydin and Göktürk 2018; Hasanzadeh, Esmaeili and Dodd 2018; Shi and Rothrock 2023] have used Tobii’s different eye tracking devices as well. The sampling size is 30 Hz with a standard deviation of approximately 2 Hz. The recommended screen size to use is 25” which satisfies the current environment. The distance of operating (subject to eye tracker) is between 40 and 90 cm.

There are many data outputs that can be retrieved including the time stamp of the fixation, eye position, gaze point and pupil diameter. The available 83 data parameters can be listed below. The explanation of all the 83 data and the units used can be viewed in the Tobii Studio Application.

- General Data - Includes information on participant characteristics, recording details, studio version, project and test names, export data, and the current fixation filter.
- Media Data - Includes information about media position, size and name.

- Segment and Scene Data - Includes information about scenes and segments.
- Time Stamp Data - Includes information about the time stamps available for export.
- Recording Event Data - Includes information about recording events such as key presses, studio events, external events, and mouse clicks.
- Gaze Event Data and AOI (Area of Interest) Activity Information - Includes information about AOI activity and gaze event data such as saccades and fixations.
- Gaze Tracking Data - Includes information regarding the raw gaze point of each data sample, either as an average or for each eye separately, represented in millimeters or pixels.
- Eye Tracking Data and Validity Codes - Includes information such as validity codes, pupil size and eye position.

The outputs that were used are time stamp, fixation position point and pupil diameter. These outputs can be exported by using the Tobii Studio application to an excel or tab-separated values file format. The issues encountered and their solutions for the eye tracker device can be found in Appendix-A.

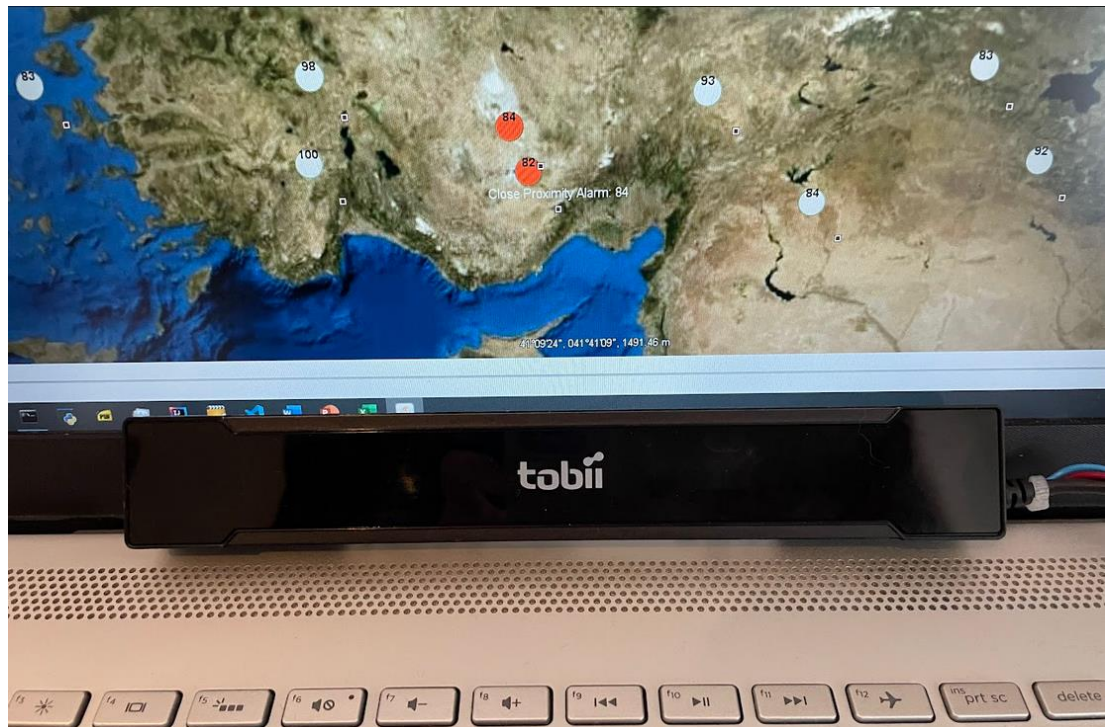


Figure 3.3: Tobii X2-30 eye tracker device.

This eye tracker is categorized as a remote eye tracker because the device stays in place while it is used. There are mobile eye trackers (wearable eyeglasses) which give the user greater freedom of movement as opposed to remote trackers. For simulations remote eye trackers are ideal but in the real world, mobile eye trackers can prove to be more of an advantage [Martinez-Marquez et al. 2021].

3.5. Methodology

The study includes a simulation related to tasks that are executed by air traffic controllers. The simulation has scenarios that test the participant's situation awareness. The scenario has flying aircrafts represented as circles and the participant must select a pair of aircrafts that are in conflict with each other or are at risk in being in a conflict (**Figure 3.4**). When the aircraft are close to colliding a warning or alarm appeared. The user's task was to monitor and resolve the aircraft that are losing separation. During the simulation, at random times the SAGAT technique was applied meaning that the scenario stopped, and the users were expected to answer questions related to the experiment. Simultaneously the Tobii X2-30 eye tracking device captured the participants' eye movements. After the simulation, the correlation between the measured eye metrics, SAGAT score, and participants' performance score were evaluated. If there are no significant correlations found between eye metrics and the performance measurements, then the initial hypothesis will be rejected.

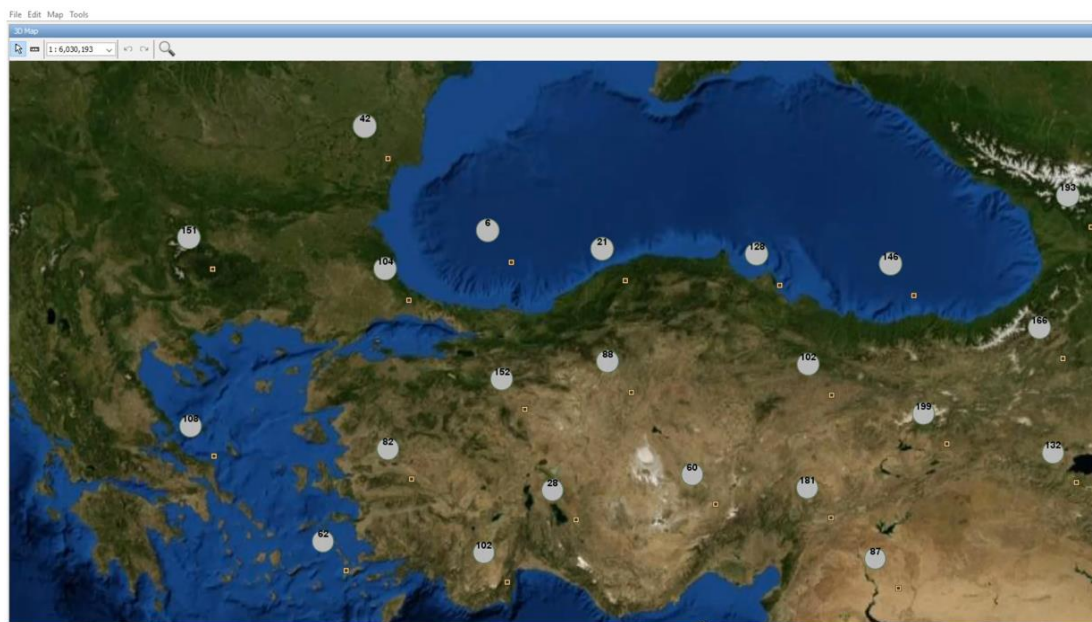


Figure 3.4: Thesis simulation where the participant is instructed to monitor colliding aircrafts.

The performance measurements were the number of missed clicks (the aircraft that have collided but were not detected by the user) and correct clicks (the aircraft that have collided and were detected by the user).

The overview of the experiment can be shown in **Figure 3.5** as a system block diagram.

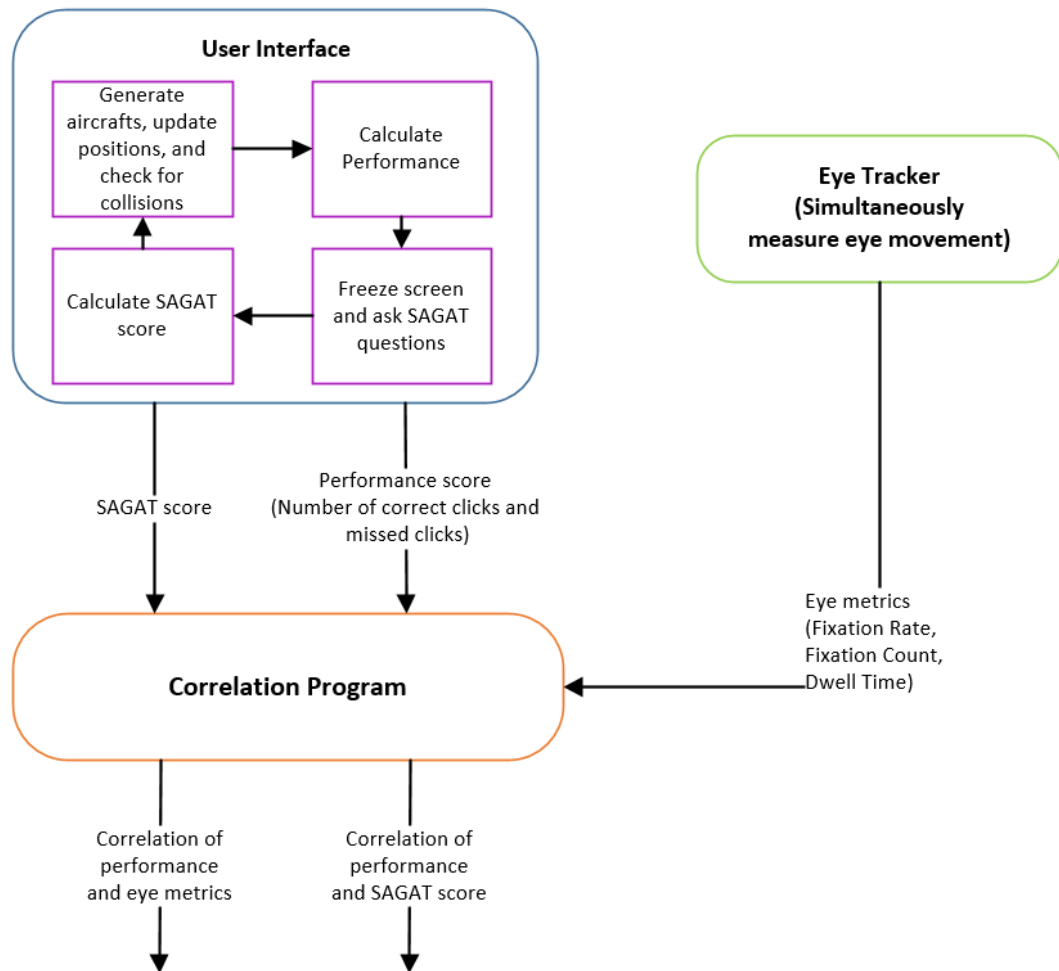


Figure 3.5: System block diagram of the current thesis study.

Since both eye trackers (Tobii X2 – 30 and TX300) were not able to be used for the experiments, the hypothesis of this study was modified. The new hypothesis can be seen below;

- The hypothesis will be “Performance measurements are closely correlated with probe related techniques such as SAGAT”. This hypothesis states that performance can also be a measurement of situation awareness.
- The null hypothesis of the study is that there will be no correlation between performance and situation awareness.

The independent variables are the SAGAT question types and the interval number. The dependent variables are correct clicks, missed clicks and the SAGAT score.

3.5.1. Participants

A total of 24 subjects participated in the experiment. The participants that the experiment was executed on consisted of 12 females and 12 males ranging from ages 22 – 42 (**Table 3.1**). Most of these participants have a decent knowledge and background of the aviation domain. Most participants majored in computer or electrical engineering.

Table 3.1: Participant information.

Number of Participants	Gender	Mean Age	Standard Deviation
12	Female	28.88	2.66
12	Male	28.71	5.25
24	Female+Male	28.88	6.83

Figure 3.6 shows one of the participants participating in the experiment.



Figure 3.6: Experiment where the participant is instructed to monitor colliding aircrafts.

3.5.2. Scenario

The scenario of the simulation consists of the generation of random aircrafts as aircrafts leave the screen and the generation of the freeze time of the SAGAT queries. **Figure 3.7** shows the flow of the scenario's algorithm. The simulation starts by loading 20 random aircraft. The aircraft altitude, location and heading information is randomly generated. The reason behind the locations and heading being randomly generated is to simulate real life air traffic. During poor weather conditions, air traffic controllers sometimes won't have immediate knowledge of where different aircraft are located at.

As the scenario progressed, the scenario was stopped with the screens blanking out at random times and the user had to answer a SAGAT questionnaire based on the 3 levels of situation awareness; perception, comprehension and projection. According to Endsley and Garland (2000), it's advised to freeze the scenario at random times so that the users cannot predict and prepare for the questions that are to come ahead. After the user has submitted his/her answer, the scenario resumed until the advisor terminated the scenario. The duration of the experiment was 32 minutes.

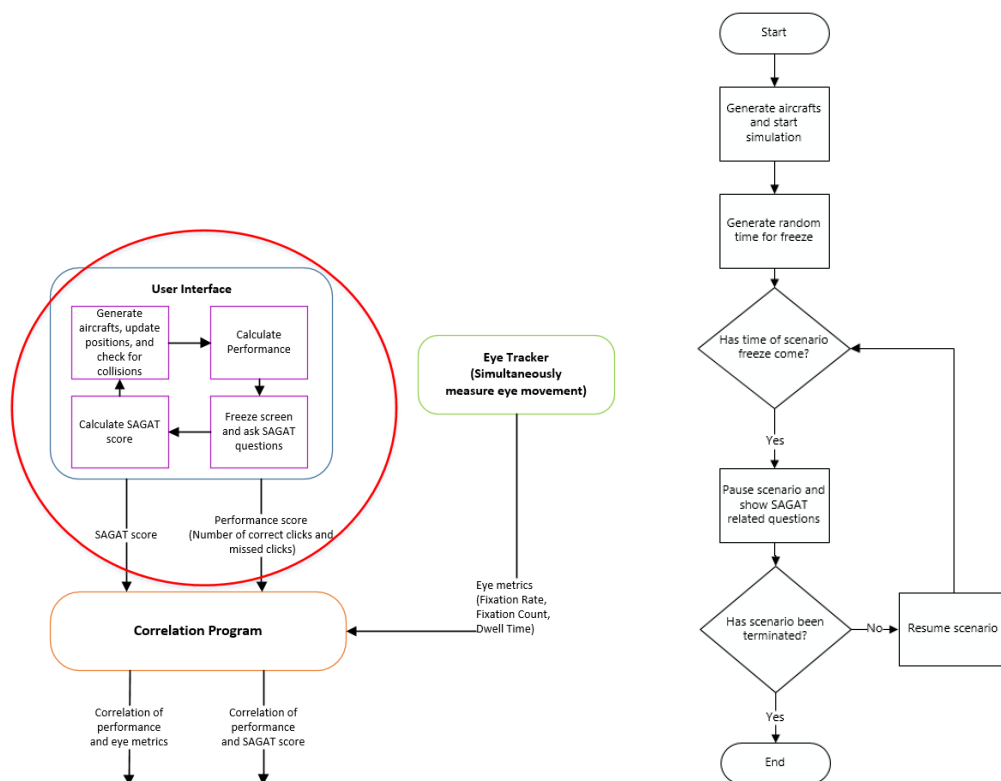


Figure 3.7: a) Marked user interface of system design. b) Simulation scenario flowchart of user interface.

The summarization of the scenario of the experiment can be seen below.

1. The subjects were given instructions on the experiment before the actual trials started. Before starting the real trial, the subjects took part in a 10-minute test run. This was done to help the person become accustomed to the task.
2. By clicking on either pair of the concerning aircraft, the subjects began to resolve the aircrafts that were about to collide during the experiment. The subjects would know that the conflict had been solved if the aircraft did not turn yellow and if they went in opposite directions.
3. The application at random times would pause the simulation and then the pop-up window containing SAGAT questions would appear. The user would try their best to give accurate answers by clicking on the given options and then submitting their response.
4. The simulation would resume, and the subject would continue on resolving aircrafts that were close to colliding until the next freeze would occur.

3.5.3. Collision Algorithm

The algorithm of the determination of the aircraft collision can be shown in **Figure 3.8**. The aircraft's altitude is shown as the number written in the circle resembling the aircraft. The aircraft approaching each other is deemed in conflict if the difference of the altitude is less than 10 Nautical Miles (NM) and if their horizontal distance difference is less than 20 NM. The conflict is presented to the user as the aircrafts foreground color being red and the label being set to "Close Proximity Alarm" (**Figure 3.9**). The aircraft's losing separation is considered to be at a risk for conflict if the difference of the altitude is less than 10 Nautical Miles (NM) and if their horizontal distance difference is less than 40 NM but above 20 NM. The risk of conflict is presented to the user as the aircrafts foreground color being yellow and the label being set to "Close Proximity Warning" (**Figure 3.10**). The aircraft's positions are updated every one second and the distance traveled for each iteration is 2 NM.

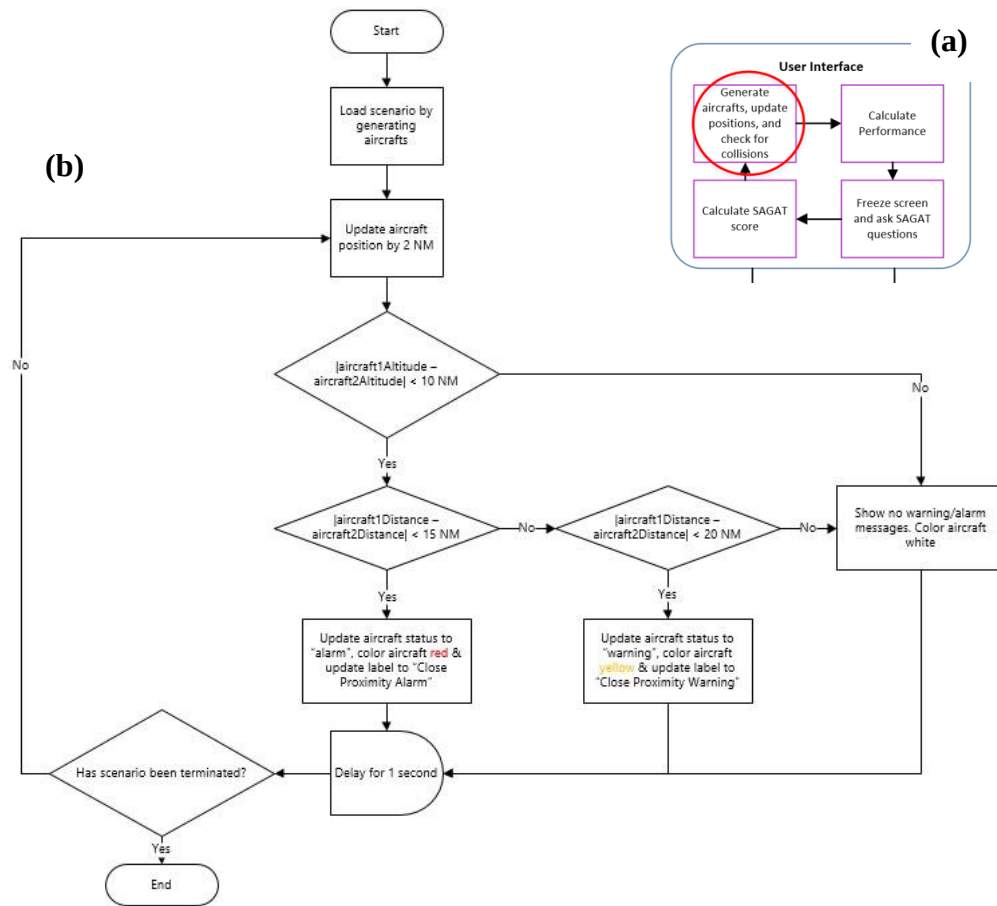


Figure 3.8: a) Marked aircraft collision detection of system design. b) Aircraft collision detection algorithm flowchart.



Figure 3.9: Representation of close proximity alarm colored in a red foreground.

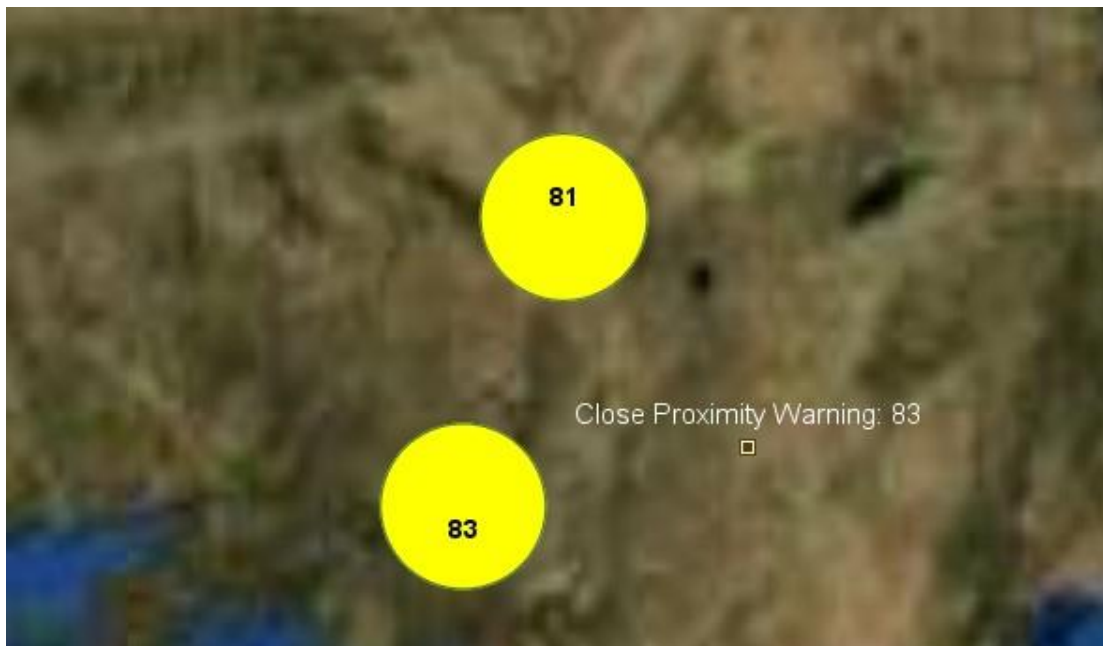


Figure 3.10: Representation of close proximity warning with aircrafts colored in a yellow foreground.

3.5.4. SAGAT Queries

According to these recommendations given by Endsley (1988) and Endsley and Garland (2000), in the thesis scenario, the freeze duration was about one to two minutes long meaning that the user would have one to two minutes to answer the questions. There are at least 4 freezes in the experiment. The queries that appeared to the user can be seen in **Table 3.2** where each question corresponds to one of the three levels of situation awareness. The questions that were asked consisted of 5 different question types.

The random time generation range of the freeze probe is between 3 – 8 minutes [Endsley 1988; Endsley and Garland 2000] for the SAGAT queries to appear to the user on the screen. The reason that the random generation of the freeze probe query time is between 3 and 8 minutes is that according to Endsley and Garland (2000), there is an informal rule says that states that freezes should not happen earlier than three to five minutes into an experiment. The reason for this is that the participant needs to start creating a mental picture of the ongoing task and comprehend it before answering the given queries. In conjunction to the informal rule, no two freezes should happen in a minute or less. With the range given in this study, even if the random generator in the algorithm generated a 3-minute time frame consecutively, this suggestion would still be valid.

The key difference between the interval time and the random freeze time of the SAGAT questions lies in their roles and variability. The interval time is a fixed duration of 8 minutes, ensuring that the entire experiment is divided into four equal parts. Systematic analysis and comparison over several intervals are made easier by this consistent categorization. The random freeze time introduces variability and unpredictability into the experiment. These freeze moments, which are generated, may occur only once or more than once in an 8-minute period. This randomness ensures that participants are queried at different times, preventing predictability for the participant. In summary, the random freeze time adds an element of unpredictability that is essential for appropriately assessing participants' situational awareness in a variety of settings, while the interval time gives the experiment a fixed and consistent structure.

Table 3.2: SAGAT questionnaire inspired by Endsley.

Level of SA	Questions
1) Perception	1.1. Which region has the most aircrafts?
2) Comprehension	2.1. Did you recently resolve an aircraft?
	2.2. Which was the last aircraft you resolved?
	2.3. Which aircraft has lost separation?
3) Projection	3.1. Which aircraft pair is close to colliding if they continue their current paths

The different levels of SA questions asked during the experiment can be seen in **Figure 3.11**, **Figure 3.12**, **Figure 3.13** and **Figure 3.14**. In previous studies [de Winter et al., 2019; Argyle et al. 2020], questionnaires were answered with pen and paper, or on a computer. In this study, users were presented with questions on screen and were expected to answer multiple choice questions according to their recollection of the executed tasks.

The formation of the questions is inspired by the advice given by Endsley where an x-y coordinate map is given that includes the aircraft location. Each option that includes an aircraft is mapped to the given spatial map. This way the user can recognize the aircraft instead of recalling from memory.

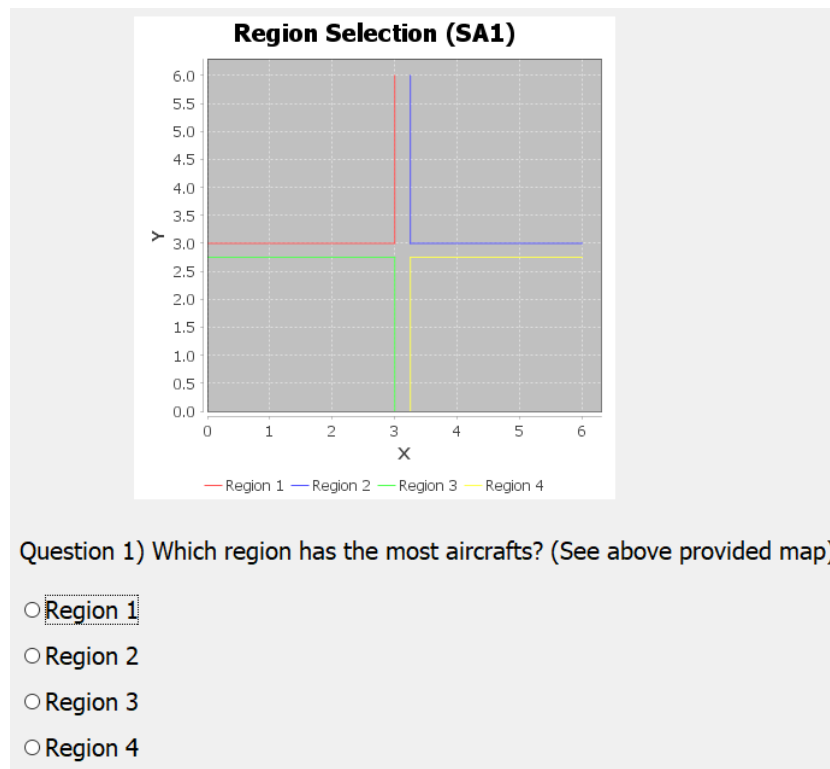


Figure 3.11: Situation awareness level 1 question: region.

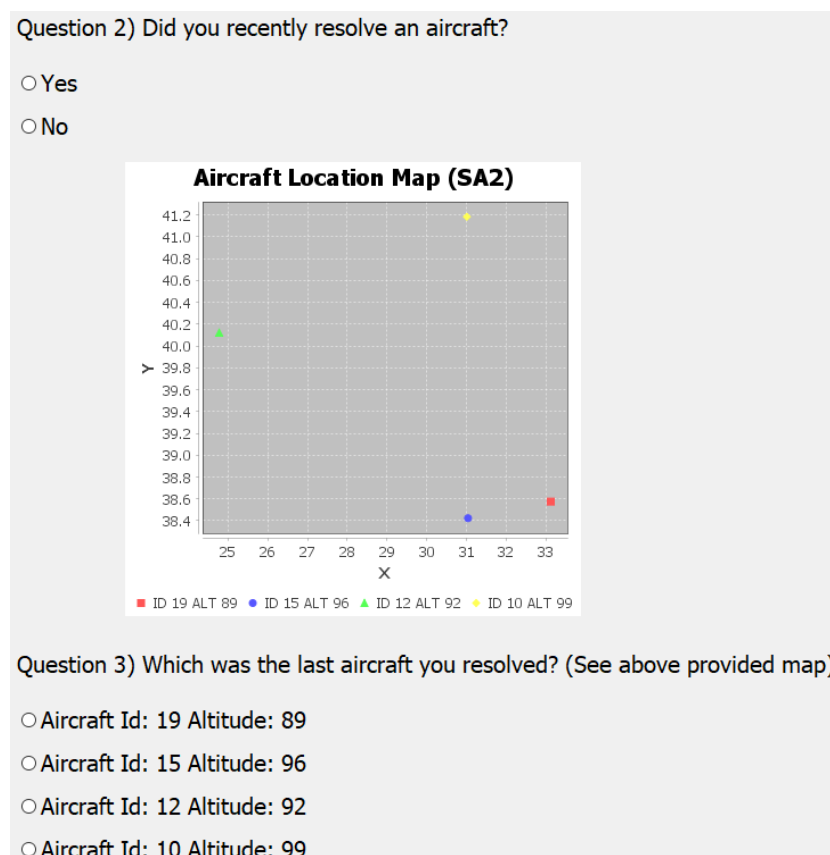
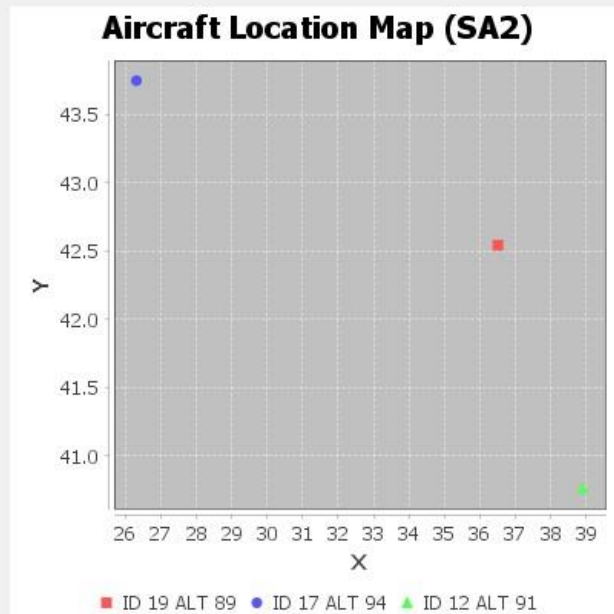


Figure 3.12: Situation awareness level 2 question: last resolved aircraft.

Question 1) Which aircraft has lost separation?



- ☐ Aircraft Id: 19 Altitude: 89
- ☐ Aircraft Id: 17 Altitude: 94
- ☐ Aircraft Id: 12 Altitude: 91
- ☐ None of the above

Figure 3.13: Situation awareness level 2 question: last conflicted aircraft.

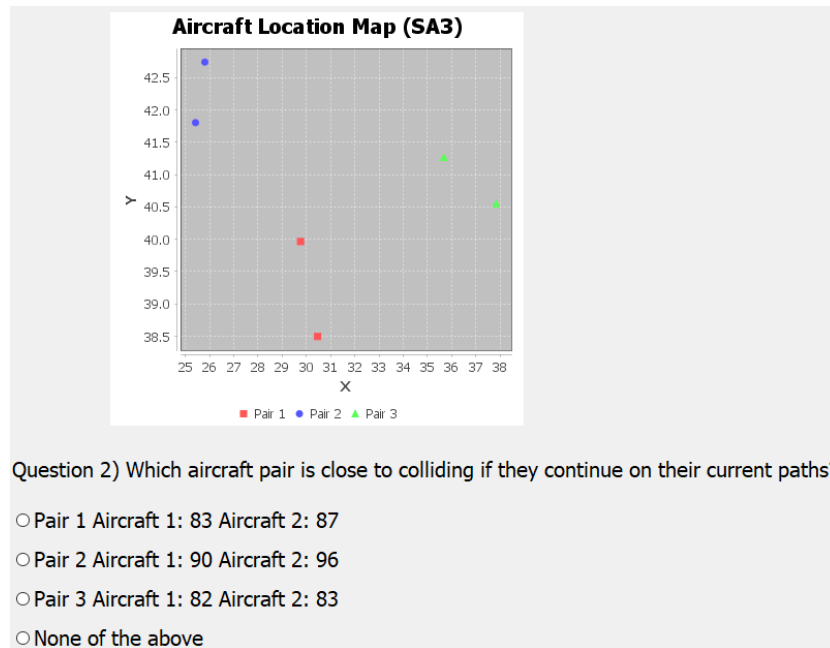


Figure 3.14: Situation awareness level 3 question: future aircraft collision.

The answers to the questions were evaluated by the advisor. According to Endsley (2021), SAGAT questions are recommended to be scored as incorrect or correct. This correctness of the answer however can be based upon whether the answer falls into an acceptable range around the actual correct answer. Studies such as the work of O'Brien & O'Hare (2007) and Fujino et al. (2020) used points to score correct answers.

If the participants are not sure about an answer to a given question, they are advised to answer as best as they can. Questions that are not answered were labeled as incorrect. The SAGAT score were evaluated as a percentage based on the correctness of each query. The reason for this is that Endsley (2021) also advises to score queries with percent correct and that in this study, all participants freeze time varies yielding in a different number of questions asked. The acceptable range of a correct answer is if the user is off by only 10 flight level when answering the altitude questions. For example, if the user has answered the question “**Which aircraft has lost separation?**” with an aircraft’s altitude being 93 but the correct answer is 98, the answer would still be marked as correct.

The questions asked in the study were evaluated as follows.

Level 1: Question **Which aircraft has the most aircrafts**. The answer was sought out by looking at the screen recordings.

Level 2: Questions *Did you recently resolve an aircraft?, Which was the last aircraft you resolved?, Which aircraft has lost separation?* correct answers were found in the application logs. The application continuously keeps track of aircrafts that are about to collide (resolvable, these aircrafts are deemed correct clicks if the user has clicked on it) and aircrafts that have already collided (missed clicks). It also keeps track of which aircraft collision that the user resolved. When this question is asked, the application already knows if the user has resolved an aircraft, so the resolved aircraft is picked and placed as one of the options of the question. The other options are randomly selected.

Level 3: Question *Which aircraft pair is close to colliding if they continue their current paths?* correct answers were found in the application logs. The application picks which aircraft pair is resolvable as well as three other random aircraft pairs as options. Before the question is asked, the application logs which aircraft pair is close to colliding, proving the correct answer in the application log file.

3.5.5. Performance Parameters

In addition to scoring the SAGAT queries, the user's actual performance also is scored. The performance parameters that were evaluated are the number of correct clicks and missed clicks relating to detecting an aircraft collision. A participant has successfully resolved an aircraft if the aircraft has not yet displayed the warning message shown in **Figure 3.10**. If the distance between the two colliding aircrafts is less than the default threshold (this is the distance where the aircrafts no longer are considered in conflict), larger than the warning threshold and the altitude difference is less than 10 NM, then the aircraft pair is considered resolvable. If the user clicked this resolvable aircraft, then it was considered a correct click. If this condition was not satisfied and the warning message appeared, then it was decided that the user had missed this collision. The algorithm can be seen in **Figure 3.15**.

The performance score was determined by the percentage of number of correct clicks and missed clicks. If the user resolved 5 aircrafts but missed 10 in the process, the score would be 33.33%. If the user resolved 10 aircrafts but only missed 5 then the score would be 66.66%. For each quarter (A total of four quarters each being 8

minutes) of the scenario, the user's performance were evaluated by this scoring technique.

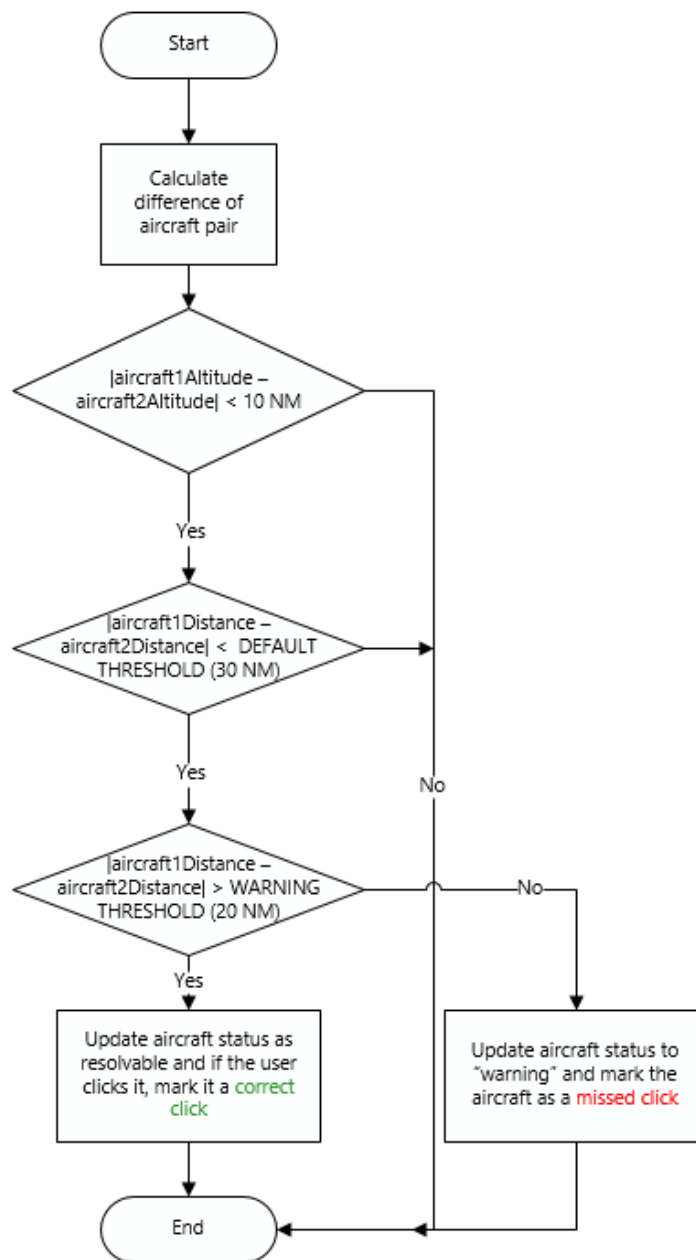


Figure 3.15: Correct and missed click algorithm flowchart.

3.5.6. Eye Tracking Metrics Used

The eye tracking metrics that were used in this study were the participant's fixation point. This metric is collected by the Tobii eye tracker and the data can be exported in

either an excel format or tab-separated values file format using the Tobii Studio application (**Table 3.3**). The exact parameters extracted by the Tobii Studio application were Local Time Stamp, Gaze Event Type, Gaze Event Duration, Fixation Index, Fixation Point X, Fixation Point Y, Pupil Right and Pupil Left.

The fixation points are given in x – y pixel format. The aircraft location are given in geospatial coordinates. To match these coordinates, the eye fixation pixel coordinates were converted to geospatial coordinates. The library that was used in the simulation also had a pixel to geospatial coordinates converter which considered the projection and the view of the given airspace map. It's important for the current view of the map to not change during the experiment or else the transformation of the fixation coordinates would not match with the aircraft's coordinates. To make sure the fixation pixel coordinates are transformed correctly to geospatial coordinates with the current view, the eye metric data is quickly exported into an excel format without closing the simulation application. Then the advisor imports this file to the simulation application and then it starts to convert the fixation points to geospatial coordinates. After this process is completed, the converted coordinates are saved into a JSON format file (**Figure 3.16**).

Table 3.3: Example of exported excel file by Tobii Studio.

LocalTimeStamp	FixationIndex	GazeEventType	GazeEventDuration	FixationPointX	FixationPointY	PupilLeft	PupilRight
21:11:30.798	3	Fixation	452	486	759	3.98	4.57
21:11:30.832		Saccade	71			4.01	4.59
21:11:30.866		Unclassified	228			4.08	4.68

```
[
  {
    "gazeEventType": "Fixation",
    "gazeEventDuration": "932",
    "pupilLeft": "4.27",
    "localTimeStamp": "20:49:44.328",
    "pupilRight": "4.74",
    "fixationLongitude": 30.80719625536115,
    "fixationIndex": 2,
    "fixationLatitude": 40.430952834806064,
    "fixationPointY": "409",
    "fixationPointX": "762"
  },
  {
    "gazeEventType": "Fixation",
    "gazeEventDuration": "217",
    "pupilLeft": "4.26",
    "localTimeStamp": "20:49:45.230",
    "pupilRight": "4.66",
    "fixationLongitude": 33.931685099143955,
    "fixationIndex": 3,
    "fixationLatitude": 39.62664897947922,
    "fixationPointY": "482",
    "fixationPointX": "980"
  },
  {
    "gazeEventType": "Fixation",
    "gazeEventDuration": "468",
    "pupilLeft": "4.32",
    "localTimeStamp": "20:49:45.482",
    "pupilRight": "4.74",
    "fixationLongitude": 31.52382213696271,
    "fixationIndex": 4,
    "fixationLatitude": 40.704199394825054,
    "fixationPointY": "384",
    "fixationPointX": "812"
  }
]
```

Figure 3.16: Fixation geospatial coordinate conversion JSON file.

During the experiment, the aircrafts updated positions are also recorded to a JSON file (**Figure 3.17**). After the experiment, the exported fixations in geospatial coordinates file and aircraft updated positions file are compared to determine which resolvable aircrafts have been looked at by the user. The number of resolvable aircrafts that have been seen by the user is defined as the target hit fixation count. The number of aircrafts that were resolvable but have not been looked at by the user is the target miss fixation count. The eye tracker score was determined by the percentage of number of target hit fixation count and target miss fixation count. If the user resolved 5 aircrafts and was fixating on those aircrafts but missed 10 in the process, the eye tracker score would be 33.33%. The total number of fixation count was also calculated. This parameter includes all fixations made by the user on the screen. For each quarter of the scenario, the user's eye tracker score was evaluated.

The comparison algorithm can be seen in **Figure 3.18**. The corresponding fixation and aircraft can be seen when evaluating the simulation log files.

```
{
  "aircraft": {
    "altitude": 95,
    "details": [
      {
        "latitude": 40.32342525532502,
        "time": "16:37:02.074",
        "longitude": 27.12229283848863
      },
      {
        "latitude": 40.30024917968805,
        "time": "16:37:03.080",
        "longitude": 27.09095343240245
      },
      {
        "latitude": 40.27707301439082,
        "time": "16:37:04.088",
        "longitude": 27.05962472891065
      },
      {
        "latitude": 40.25389675944302,
        "time": "16:37:05.097",
        "longitude": 27.028306715601587
      }
    ],
    "id": "11"
  }
},
```

Figure 3.17: Aircrafts updated position coordinates JSON file.

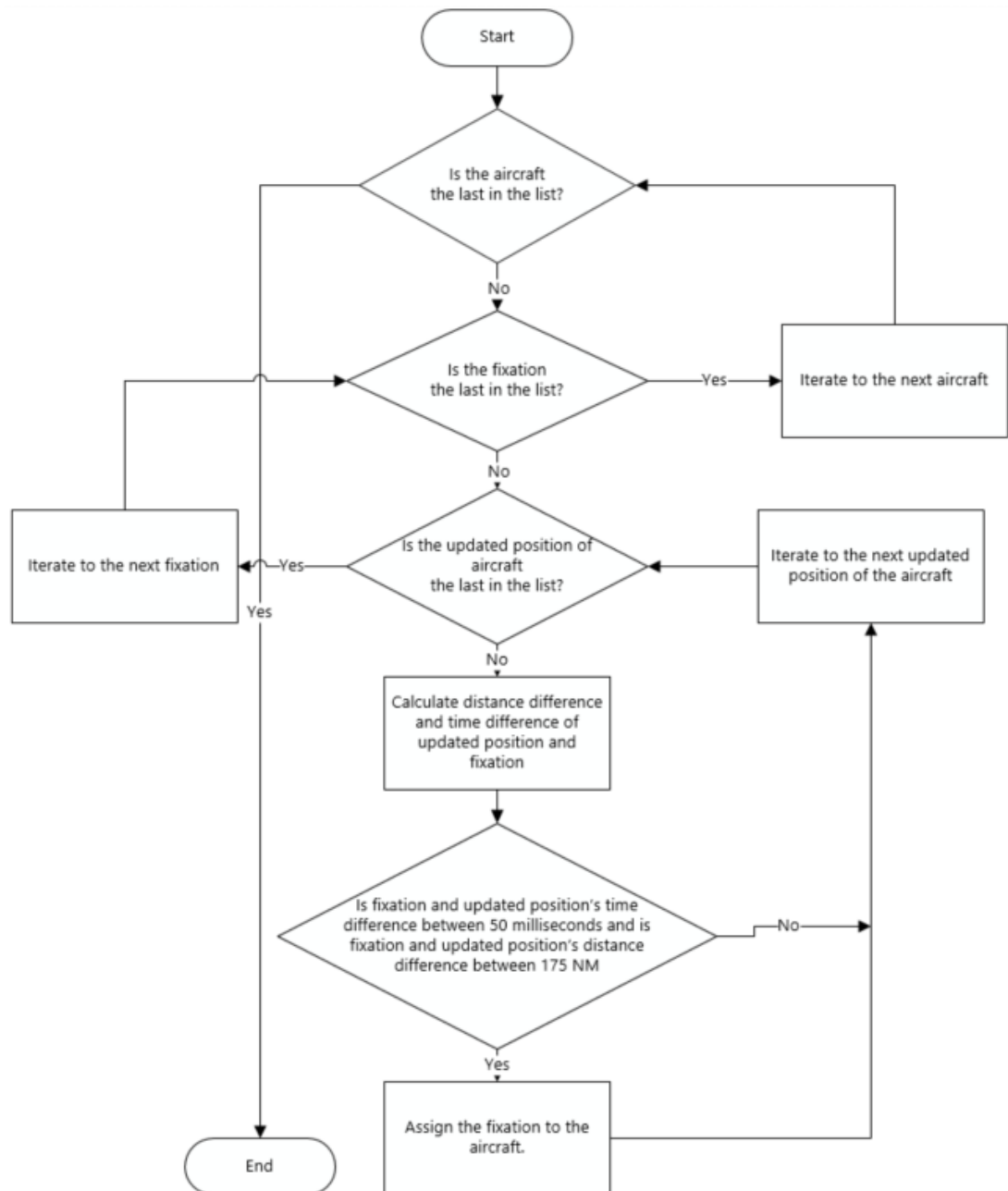


Figure 3.18: Aircraft and fixation correspondence algorithm flowchart.

The complete steps of the aircraft and fixation correspondence are summarized below.

1. Start the scenario and Tobii eye tracking recording.
2. After finishing the experiment, stop the scenario. This allowed all the aircraft's updated positions to be saved as a JSON file. Each aircraft's position that is updated every second is logged with its then current position and time of the updated position.
3. Export the participants eye metrics from Tobii Studio as an excel file.

4. Convert the position of the fixations from pixel coordinates to geographical coordinates by clicking File -> Open Tobii File in the simulation program. This would read Tobii's exported excel file and write to a new Tobii eye metric results JSON file with the converted fixation coordinates.
5. To compare which fixations are fixated on which aircrafts, click the Compare Fixations & Aircrafts button. This button allows the evaluator to select the updated aircrafts positions JSON file and Tobii eye metric results JSON file. After selecting these two files, the application compares the distance between the fixation and aircraft positions as well as the time of which the fixations and aircraft positions took place.

4. RESULTS

A total of 24 participants were included in the experiment conducted in this research. Four of which were air traffic controllers.

The descriptive statistics table in **Table 4.1** represents the comprehensive statistical information for the experiments. On average, participants accurately resolved 38 aircraft. The minimum number of aircraft resolved was 25, while the maximum number of aircraft was 59. On average, participants were unable to resolve 15 aircraft. The minimum number of unresolved aircraft was 3, while the maximum number of aircraft was 30. The average performance score of the participants was 71.01. The minimum performance score recorded was 53.85, while the maximum score was 95.16. The average number of questions answered correctly by participants was 8. The minimum number of accurately answered questions was 6, whilst the maximum was 13. The average number of questions that participants answered wrong was 2. The minimum number of incorrectly answered questions was 1, while the maximum was 7. Throughout the course of the experiment, an average of 11 SAGAT questions were asked. The minimum number of questions asked was 7, and the maximum number of questions was 17. The average score obtained by participants on the SAGAT questions was 77.36. The minimum score was 58.33, while the maximum score was 91.67.

The descriptive statistics for the eye metrics were also calculated. On average, participants fixated on 42 aircrafts that were resolvable (target hit fixation count). The minimum number of targets hit fixation counts was 15, while the maximum number was 66. On average, participants failed to fixate on 13 aircrafts that were resolvable (target miss fixation count). The minimum number of targets miss fixation count was 0, while the maximum number was 42. The eye tracker score was calculated by taking the percentage of fixation counts on the target that were successfully hit, out of the total fixation counts (hits and misses) on the target, and then multiplying by 100. The overall eye tracker score average regarding the target hit and miss fixation count was 75.85. The minimum eye tracker score was 26.32, while the maximum score was 100. The average total fixation count of the participants during the experiment was 18,744. The minimum total fixation count was 2,375, while the maximum total fixation count was 49,134.

Each participant's situation was distinct, but the task and questions remained the same throughout all scenarios.

Table 4.1: Descriptive statistics of the experiment.

	N	Min.	Max.	Mean	Std. Dev.	Var.	Skewness	Kurtosis
Total_Correct_Clicks	24	25	59	38.46	9.2	84.69	0.5	-0.39
Total_Missed_Clicks	24	3	30	15.75	7.5	56.2	0.02	-0.71
Success_Rate	24	53.85	95.16	71.01	12.51	156.61	0.4	-0.91
Total_Correct_Answer	24	6	13	8.75	1.96	3.85	0.34	-0.95
Total_Incorrect_Answer	24	1	7	2.58	1.41	1.99	1.33	3.07
Total_Questions	24	7	17	11.33	2.18	4.75	0.63	0.88
SAGAT_QA_Score	24	58.33	91.67	77.36	10.22	104.53	-0.12	-0.97
Target_Missed_Fixation_Count	24	0	42	12.67	11.13	123.97	0.91	0.49
Target_Hit_Fixation_Count	24	15	66	41.58	14.88	221.47	0.19	-1.07
Eye_Tracker_Score	24	26.32	100	75.85	20.48	419.63	-0.56	-0.41
Total_Fixation_Count	24	2375	49134	18744.54	12631.83	160000000	0.74	-0.16
Over_Fixation	24	0	132	29.5	40.48	1638.61	1.66	1.88
Valid N (listwise)	25							
Missing N (listwise)	1							

A cumulative performance score was obtained for all individuals at each 8-minute interval, resulting in a total of four folds/quarters being examined. The average was computed for the SAGAT score due to the scores being expressed as percentages. **Figure 4.1** displays a clustered column chart representing the scores. The correct and missed clicks of each participant were aggregated for each period. The subject's SAGAT score was also calculated as an average. The data shown in **Table 4.2** serves as the basis for the chart seen in **Figure 4.1**.

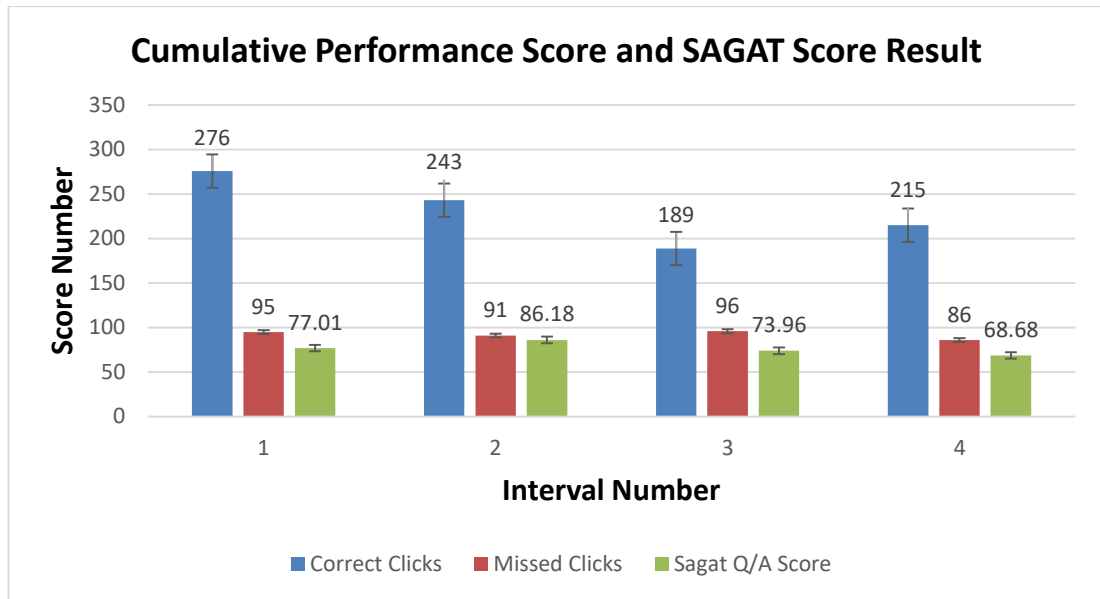


Figure 4.1: Cumulative performance measurements and SAGAT score result.

During interval 1, which lasted for the first 8 minutes of the experiment, the number of accurate clicks was the largest. In contrast, interval 3 had the lowest count of correct clicks. The frequency of missed clicks was greatest during interval 3, gradually decreasing as the trial progressed, with the lowest number of clicks missed occurring during interval 4. The decline of the number of correct clicks between interval 1 and interval 3 could be an indicator of an increase of mental workload resulting in poor performance. The mean SAGAT score exhibited a rise from intervals 1 to 2, followed by a decline from intervals 3 to 4. The period with the highest average SAGAT score was interval 2, while the interval with the lowest points was interval 4. The rising SAGAT score suggests that participants have begun to prioritize attention to detail and prepare for the impending freezes (**Table 4.2**). The decline of the SAGAT scores from interval 3 to interval 4 may suggest that the participant was starting to dismiss details about the ongoing task.

Table 4.2: Cumulative performance score and SAGAT score result.

Interval	Number of Correct Clicks	Number of Missed Clicks	SAGAT Q/A Score Average (%)
1	276	95	77.01
2	243	91	86.18
3	189	96	73.96
4	215	86	68.68

The cumulative measurement for the performance score, SAGAT score and eye tracker score for each interval was also calculated and can be seen in **Figure 4.2**. The subject's performance score, SAGAT score and eye tracker score were calculated as an average. The lowest eye tracker score took place in interval 2 and the highest eye tracker score occurred in interval 1. The highest performance scores were calculated during interval 1 whereas the lowest score was calculated during interval 3. The SAGAT score was highest during interval 2 and the lowest during interval 4.

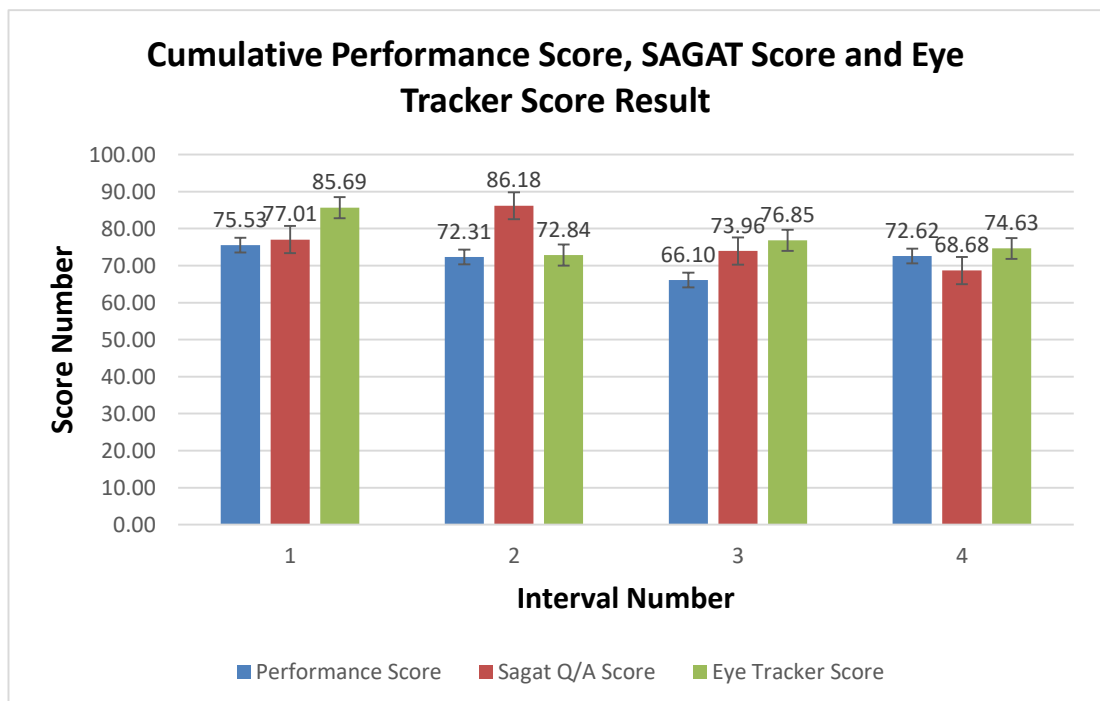


Figure 4.2: Cumulative performance score, SAGAT score and eye tracker score results.

The cumulative score of performance, SAGAT and eye tracking were all depicted in **Table 4.3**. The overall eye tracker scores range from 72.84% to 85.69%.

Table 4.3: Cumulative performance score, SAGAT score and eye tracker score result.

Interval	Performance Score Average (%)	SAGAT Q/A Score Average (%)	Eye Tracker Score Average (%)
1	75.53	77.01	85.69
2	72.31	86.18	72.84
3	66.10	73.96	76.85
4	72.62	68.68	74.63

The cumulative number of clicks and fixation count for each interval were shown in **Figure 4.3**. The highest number of fixation count that corresponded with the resolvable aircrafts position that were looked at by the participants occurred in interval 1 and this interval also holds the highest correct clicks (number of resolved aircrafts). The lowest number of fixation count that corresponded with resolvable aircrafts that were looked at took place in interval 4. Participants' target hit fixation count decreased between intervals 1 and 4. The highest number of fixation count that corresponded with missed resolvable aircraft took place in interval 2 and the lowest count number took place in interval 1. Intervals 1 through 3 showed a decrease of correct clicks and target hit fixation count, which could indicate mental fatigue and attention decline.

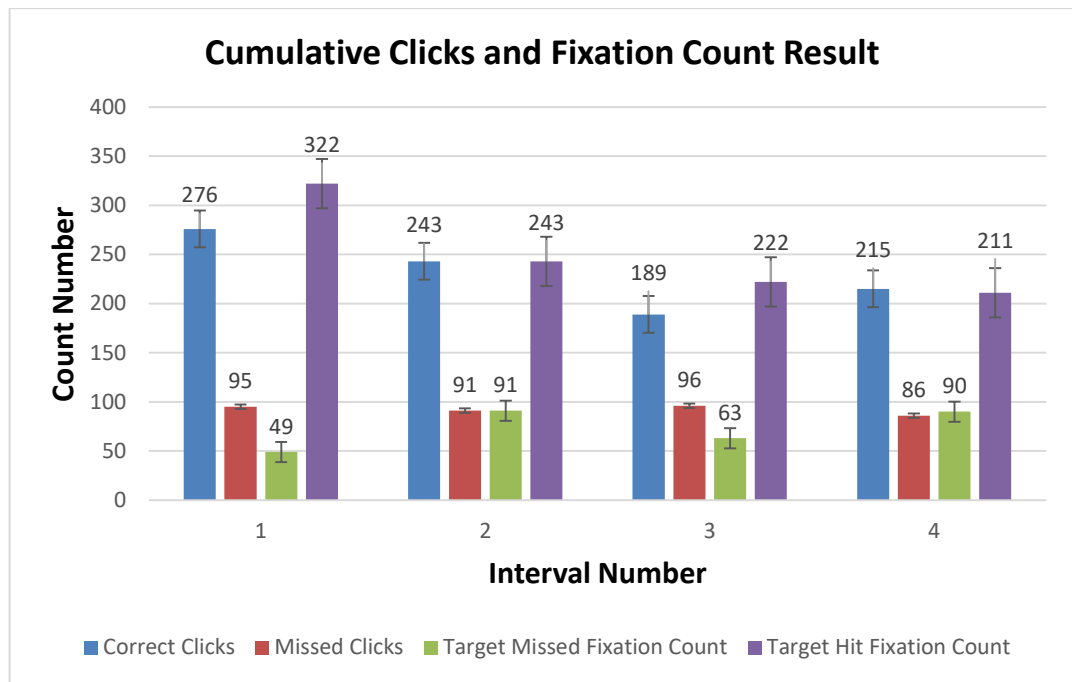


Figure 4.3: Cumulative clicks and fixation count result.

The cumulative correct/missed clicks and target hit/missed fixation count were shown in **Table 4.4**.

Table 4.4: Cumulative clicks and fixation count result.

Interval	Correct Clicks	Missed Clicks	Target Missed Fixation Count	Target Hit Fixation Count
1	276	95	49	322
2	243	91	91	243
3	189	96	63	222
4	215	86	90	211

Figure 4.4 represents the cumulative number of total fixations count for each interval. This includes all eye activity that is considered a fixation. It can be seen that the most fixation count was calculated for interval 2 while interval 1 had the lowest fixation

count. The total fixation count increased between interval 1 and interval 2 following a decrease of fixation count between interval 2 and interval 4.

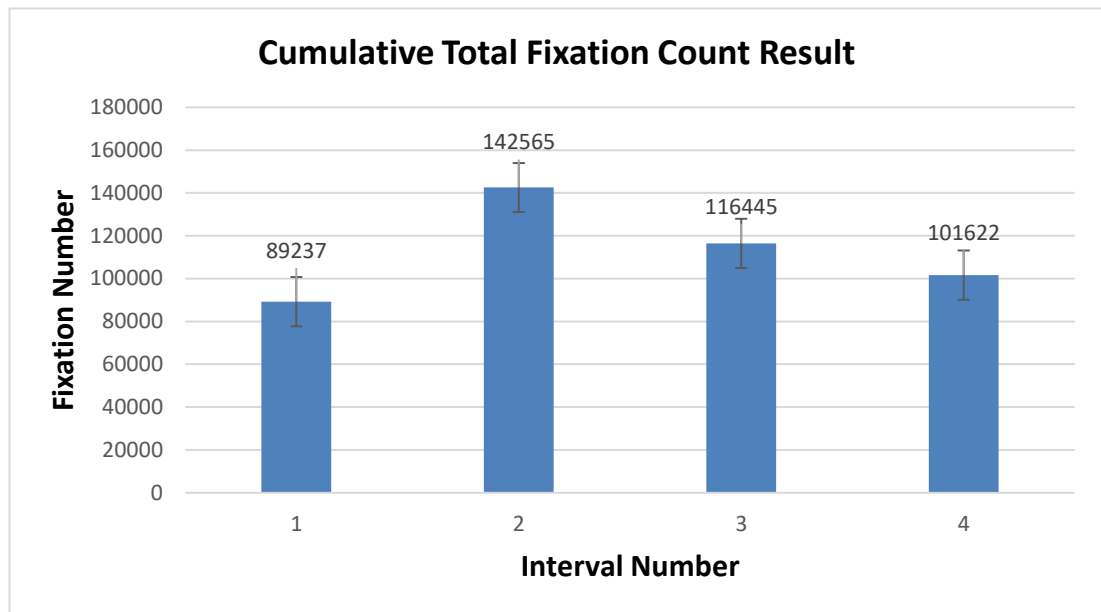


Figure 4.4: Cumulative total fixation count result.

The SAGAT questions asked were based on different levels of SA (Perception, Comprehension and Projection). While evaluating the correctness of each answer, a chart was created to show the accuracy of each question's answer made by the subject shown in **Figure 4.5**. The questions matching the question numbers can be seen in **Table 4.5**. The chart was based on **Table 4.6** where this table shows cumulative number of incorrect and correct answers made by the subjects.

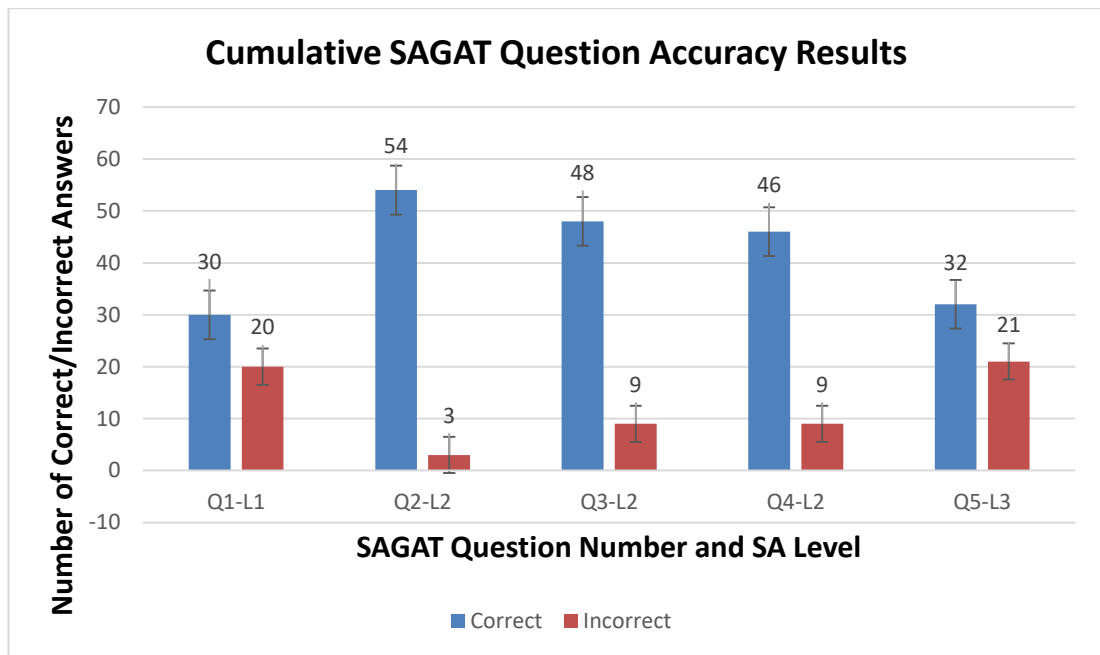


Figure 4.5: Results of cumulative SAGAT questions' accuracy based on the situational awareness levels.

A total of 5 types of SAGAT questions were asked, the reason for this small collection of questions is that the task given to the participants is a basic task. If a complex job was given for example giving actual commands (Changing a climb/descend flight level, adjusting the trajectory of the flights, or monitoring different kind of alerts such as STCA, MSAW or MONA) then a wider variety of questions could be asked. Due to technical issues of the simulation, sometimes the user could not click on a resolvable aircraft which would lead to a false collision. If the time frame of an aircraft being resolvable and being at risk of a conflict was extended longer than maybe the user would have a higher chance of saving the aircraft.

Table 4.5: Descriptions of the SAGAT questions and SA levels.

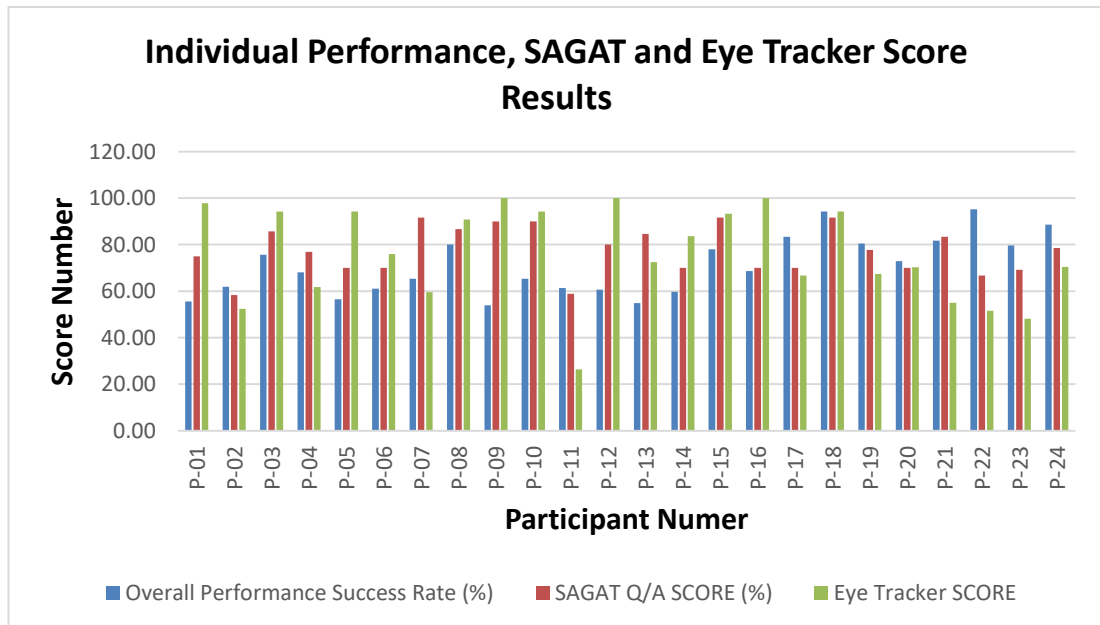
Question No	SA Level	Question Description
Question 1	Level 1	Which region has the most aircrafts?
Question 2	Level 2	Did you recently resolve an aircraft?
Question 3	Level 2	Which was the last aircraft you resolved?
Question 4	Level 2	Which aircraft has lost separation?
Question 5	Level 3	Which aircraft pair is close to colliding if they continue on their current paths?

According to **Table 4.6**, participants provided the highest number of correct answers (54) for Question 2 (Level 2), while Question 1 (Level 1) had the lowest number of correct responses (30). The answers that had the greatest number of incorrect responses were for Question 5 (Level 3), with a total of 21 incorrect answers. On the other hand, Question 2 (Level 2) had the fewest inaccurate responses, with just 3 incorrect answers. According to the collected data, it seemed that the comprehension (Level 2) questions were the ones that had the highest number of right answers, suggesting that the participants were fully aware about the state of the aircraft. The question with the lowest accuracy rate was Question 1, which pertains to perception (Level 1). It may be inferred that participants are unable to recognize which regions had the highest aircraft count. During the experiment, participants often attempted to focus their attention on a single pair of colliding airplanes, even though there may really be many collisions occurring simultaneously, ranging from 2 to 3. This may also explain why several participants were unable to accurately respond to the question at level 1.

Table 4.6: Cumulative SAGAT question accuracy results.

Number of Correct/Incorrect Answers	Question 1 (Level 1)	Question 2 (Level 2)	Question 3 (Level 2)	Question 4 (Level 2)	Question 5 (Level 3)
Correct	30	54	48	46	32
Incorrect	20	3	9	9	21

In **Figure 4.6**, the participants' individual performance score, SAGAT score, and eye tracker score can be seen. These results show variations among the participants of the overall performance rate. Some participants show higher overall performance score while others have relatively lower scores. The SAGAT scores based on these results typically follow the trend of the overall performance score. This may indicate a positive correlation between performance score and SAGAT score. There is also variation observed among participants of the eye tracker score. Higher eye tracker scores are frequently associated with higher overall performance success rates, indicating a possible positive correlation between eye tracking data and performance. However, there are a few participants that do not show this relationship.

**Figure 4.6:** Results of individual performance, SAGAT and eye tracker score.

The most common feedback given was that the resolvable range (the distance at which the user has to click in order for the aircraft to be saved) should be increased. The reason for this is that in the real world, air traffic controllers would not wait for an aircraft to be in a critical position. They would intervene immediately.

Other participants sometimes found it hard to focus on all the aircraft because there would sometimes be a few collisions happening at the same time. Every time an aircraft would go out of frame, another aircraft would be generated so there would be at least 20 aircraft in the view. There were few feedback given, one of them being that the participants would've liked the altitude label to be more centered in the circle (representation of the aircraft). Another feedback that participants had was that they would like an indicator of a solved aircraft. Normally if the aircraft was successfully resolved, the aircraft would go in opposite directions but a visual cue such as the aircraft changing to a different color was suggested.

In regard to measuring the users eye movement, it was seen that people who wore prescription eyeglasses had lower target hit fixation count. This could be the fact that the user's eye movements weren't being measured accurately by the eye tracker device. There were also calibration issues when first starting the experiments. Sometimes it would take a few calibrations measurement tries for the device to properly calibrate the eyes.

4.1. Performance Measurements, SAGAT and Eye Tracker Correlation Evaluation Using Paired Sample t-Test

The paired sample t-Test analysis [Kent State University Libraries, 2023] was used to determine the Pearson correlation coefficients between the means of the SAGAT score, correct and missed clicks, and the number of aircraft (**Table 4.7**). The available dependent variables are continuous; hence it was appropriate to use the Pearson Correlation [Kent State University Libraries, 2023]. The level of significance (α) was set to 0.001 in order to conduct a two-tailed test that would detect any significant changes, regardless of their direction. The primary aim was to identify any association, irrespective of whether it was positive or negative. The paired sample t-Test was used since all the samples belonged to the same group of individuals. The objective of this statistical technique is to assess the disparities between two variables within the same set of individuals. **Table 4.7** displays the test results for pairings.

The performance score (success rate) and SAGAT score had a weak and positive correlation ($r = 0.121, p > 0.001$). There was no significant average difference between the performance score and SAGAT score ($t = 2.05, p > 0.001$).

The performance score (success rate) and eye tracker score had a weak and negative correlation ($r = -0.207, p > 0.001$). There was no significant average difference between the performance score and SAGAT score ($t = 0.92, p > 0.001$).

The total fixation count and performance score had a weak and negative correlation ($r = -0.253, p > 0.001$). There was significant average difference between the total fixation count and performance score ($t = 7.24, p < 0.001$).

The total number of correct clicks and performance score had a strong and positive correlation ($r = 0.723, p < 0.001$). There was a significant average difference between total correct clicks and performance score ($t = -18.44, p \leq 0.001$). The total number of missed clicks and performance score had a strong and negative correlation ($r = -0.924, p < 0.001$). There was a significant average difference between total missed clicks and performance score ($t = -13.77, p < 0.001$).

The total number of correct clicks (resolved aircrafts) and the SAGAT score had a weak and positive correlation ($r = 0.207, p > 0.001$). There was a significant average difference between correct clicks and the SAGAT score ($t = -15.55, p < 0.001$). The total number of missed clicks and the SAGAT score had a very weak and negative correlation ($r = -0.064, p > 0.001$). There was a significant average difference between missed clicks and the SAGAT score ($t = -23.11, p < 0.001$).

The total number of correct clicks and total number of missed clicks had a moderate and negative correlation ($r = -0.438, p > 0.001$). There was a significant average difference between the performance score and SAGAT score ($t = 7.84, p < 0.001$).

The total number of correct clicks and target hit fixation count had a moderate and positive correlation ($r = 0.307, p > 0.001$). There was no significant average difference between the total number of correct clicks and target hit fixation count ($t = -1.03, p > 0.001$). The total number of missed clicks and target hit fixation count had a moderate and positive correlation ($r = 0.42, p > 0.001$). There was a significant average difference between the total number of missed clicks and target hit fixation count ($t = -9.31, p < 0.001$).

The total number of correct clicks and target miss fixation count had a weak and positive correlation ($r = 0.12, p > 0.001$). There was a significant average difference between the total number of correct clicks and target miss fixation count ($t = 9.31, p < 0.001$). The total number of missed clicks and target miss fixation count had a weak and negative correlation ($r = -0.251, p > 0.001$). There was no significant average difference between the total number of missed clicks and target miss fixation count ($t = 0.103, p > 0.001$).

The performance score and target hit fixation count had a weak and negative correlation ($r = -0.182, p > 0.001$). There was a significant average difference between the performance score and target hit fixation count ($t = -6.81, p < 0.001$). The performance score and target miss fixation count had a weak and positive correlation ($r = 0.22, p > 0.001$). There was a significant average difference between the performance score and target miss fixation count ($t = -19.28, p < 0.001$).

The target hit fixation count and target miss fixation count had a strong and negative correlation ($r = -0.8, p < 0.001$). There was a significant average difference between target hit fixation count and target miss fixation count ($t = -5.73, p \leq 0.001$).

In summary, there are strong, positive, and statistically significant correlations for total correct clicks and performance score. Strong, negative, and statistically significant correlations were found for total missed clicks and performance score, total correct clicks and total missed clicks, and target hit fixation count and target missed fixation count. Target hit fixation count and correct clicks showed a moderate and positive correlation but was not statistically significant. Target hit fixation count and missed clicks showed a moderate and positive correlation and was statistically significant. Most other pairs show weak or very weak correlations with some being statistically significant.

Table 4.7: Paired sample correlations.

		N	Correlation	Sig.
Pair 1	SAGAT_QA_Score & Success_rate	24	0.121	0.572
Pair 2	Success_Rate & Eye_Tracker_Score	24	-0.207	0.332
Pair 3	Total_Fixation_Count & Success_rate	24	-0.253	0.233
Pair 4	Total_Correct_Clicks & Success_rate	24	0.723	0
Pair 5	Total_Missed_Clicks & Success_rate	24	-0.924	0
Pair 6	Total_Correct_Clicks & SAGAT_QA_Score	24	0.207	0.332
Pair 7	Total_Missed_Clicks & SAGAT_QA_Score	24	-0.064	0.765
Pair 8	Total_Correct_Clicks & Total_Missed_Clicks	24	-0.438	0.033
Pair 9	Total_Correct_Clicks & Target_Hit_Fixation_Count	24	0.307	0.145
Pair 10	Total_Missed_Clicks & Target_Hit_Fixation_Count	24	0.42	0.041
Pair 11	Total_Correct_Clicks & Target_Missed_Fixation_Count	24	0.12	0.575
Pair 12	Total_Missed_Clicks & Target_Missed_Fixation_Count	24	-0.251	0.236
Pair 13	Success_Rate & Target_Hit_Fixation_Count	24	-0.182	0.394
Pair 14	Target_Missed_Fixation_Count & Success_Rate	24	0.22	0.301
Pair 15	Target_Missed_Fixation_Count & Target_Hit_Fixation_Count	24	-0.8	0

In the Paired Sample Test result (**Table 4.8**), the average difference, standard deviation, standard error mean, t-test statistic (t), degree of freedom (df) and the p-value ($Sig.$) can be seen. Pairs 3, 4, 5, 6, 7, 8, 10, 11, 13, and 15 show statistically significant differences in means ($p < 0.001$) whereas pairs 1, 2, 9, and 12 do not show statistically significant differences in means ($p > 0.001$).

Table 4.8: Paired sample hypothesis test result.

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Dev.	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	SAGAT_QA_Score - Success_rate	6.36	15.17	3.1	-0.05	12.76	2.05	23	0.052
Pair 2	Success_Rate - Eye_Tracker_Score	4.9	26.13	5.33	-6.13	15.93	0.92	23	0.368
Pair 3	Total_Fixation_Count - Success_rate	18673.54	12635.01	2579.11	13338.24	24008.83	7.24	23	0
Pair 4	Total_Correct_Clicks - Success_rate	-32.55	8.65	1.76	-36.2	-28.9	-18.44	23	0
Pair 5	Total_Missed_Clicks - Success_rate	-55.26	19.65	4.01	-63.55	-46.96	-13.77	23	0
Pair 6	Total_Correct_Clicks - SAGAT_QA_Score	-38.9	12.26	2.5	-44.08	-33.73	-15.55	23	0
Pair 7	Total_Missed_Clicks - SAGAT_QA_Score	-61.61	13.06	2.67	-67.13	-56.1	-23.11	23	0
Pair 8	Total_Correct_Clicks - Total_Missed_Clicks	22.71	14.19	2.9	16.72	28.7	7.84	23	0
Pair 9	Total_Correct_Clicks - Target_Hit_Fixation_C ount	-3.13	14.9	3.04	-9.42	3.17	-1.03	23	0.315
Pair 10	Total_Missed_Clicks - Target_Hit_Fixation_C ount	-25.79	13.56	2.77	-31.52	-20.06	-9.31	23	0
Pair 11	Total_Correct_Clicks - Target_Missed_Fixatio n_Count	25.79	13.56	2.77	20.06	31.52	9.31	23	0
Pair 12	Total_Missed_Clicks - Target_Missed_Fixatio n_Count	3.13	14.9	3.04	-3.17	9.42	1.03	23	0.315
Pair 13	Success_Rate - Target_Hit_Fixation_C ount	-29.36	21.12	4.31	-38.29	-20.44	-6.81	23	0
Pair 14	Target_Missed_Fixatio n_Count - Success_Rate	-58.28	14.81	3.02	-64.54	-52.03	-19.28	23	0
Pair 15	Target_Missed_Fixatio n_Count - Target_Hit_Fixation_C ount	-28.92	24.71	5.04	-39.35	-18.48	-5.73	23	0

The research conducted by de Winter et al. revealed a weak and positive correlation ($r = 0.200$) between the freeze-probe score (SAGAT score) and the performance score. The study found that there was a weak and positive correlation ($r = 0.207$) between the total number of correct clicks and SAGAT scores. Additionally, there was a very weak and negative correlation ($r = -0.064$) between the total number of missed clicks and SAGAT score. Furthermore, there was a weak and positive correlation ($r = 0.121$) between the SAGAT score and the overall performance score, which is calculated as the ratio of correct clicks to the sum of correct clicks and missed clicks. There was a significant moderate association ($r = -0.438$) between the total number of correct clicks and the total number of missed clicks. The link between correct clicks, wrong clicks, total performance score, and SAGAT score shown a greater magnitude than de Winter's discovery. The link between the SAGAT and performance score was weak and positive, the same as de Winter's findings which also showed a weak and positive correlation between the performance score and SAGAT score. Moderate positive correlation ($r = 0.31$) was found between visual sampling score and performance score by de Winter et al. while this study found a weak and negative correlation ($r = -0.207$) between the eye tracker score and performance score. Possible reason for the lower magnitude correlation result could be due to the fact that de Winter et al.'s experiment were conducted on more participants (86 university students), resulting in more reliable observations. With more participants, the study was less likely to be influenced by outliers or anomalies.

In the experiments conducted by Hasanzadeh, Esmaeili and Dodd (2016), they found a moderate and negative correlation ($r = -0.401$) between fixation count and performance but the result was not significantly different ($p > 0.05$). The results of this thesis also had found a negative correlation between performance and eye tracker score. The difference in the thresholds and algorithms used to interpret eye movements might have played a role of their study finding a higher correlation magnitude. Another factor that may be relevant is the particular domain or setting in which the studies were done (for example, construction work versus aviation). How eye movements relate to performance can vary depending on the tasks involved and the surrounding conditions.

In Desvergez et al. (2019) study, in order to evaluate the variations in SA and visual perception between resident and senior anesthetists, the study used eye tracking data. According to their eye tracking metrics, they discovered that anesthetists with more

experience had superior situation awareness and visual perception. This suggests a positive association between eye tracking measures and expertise, which might be thought of as an alternative for performance. Compared to this thesis study, the difference of performance score having a negative correlation between eye tracker score, might be due to the nature of the tasks or the metrics used for performance assessment. The study highlighted the relationship between improved SA as determined by eye tracking and improved PPH scenario management. This is consistent with the observation that there is a positive relationship between SAGAT and performance score, suggesting that performance is typically improved by higher SA. Desvergez et al. emphasizes the value of experience in analyzing eye tracking data, arguing that higher levels of experience correlate with improved performance and SA. This thesis study did not specifically take experience into account.

Argyle et al. (2020), higher fixation counts have been associated with better performance and higher SA. The study found that weather forecasters with higher SA performed better on tasks because they were more fixated, especially on key information. When contrasting this study's findings with those of Argyle et al., we found discrepancies in the preparation of the data. This study included all data points in the analysis without removing anomalies or outliers, in contrast to Argyle et al. who used a 70% gaze sample criterion to guarantee good data quality and produced a sample of 68 recordings. As the inclusion of all data points in this investigation may have added noise and variability, altering the observed relationship between fixation count, SA, and performance, this variation in data filtering may account for the disparities in the findings.

To determine whether "Eye movement (tracking) techniques provide a better measurement of situational awareness in Air Controllers than probe related techniques such as SAGAT" holds true, we conducted paired sample t-testing of the performance score, SAGAT score and eye tracker score. The results indicate that the performance score and eye tracker score had a higher correlation magnitude ($r = -0.207$) than the correlation found between performance score and SAGAT score ($r = 0.121$). There were no statistical differences found for both correlations ($p > 0.001$). This evidence suggests that the hypothesis is rejected.

4.2. Group-Based Analysis of Intervals Using One-Way ANOVA Analysis

A one-way ANOVA test was used to assess the statistics of the intervals in the experiment, which comprised of four intervals. The independent variable in this study is the interval count. Since there are several intervals, the one-way ANOVA test, which assesses significant differences between two or more groups, was used. The sample size is 96 due to the presence of 4 intervals for every 24 participants.

Prior to running the one-way ANOVA test, it was necessary to examine the homogeneity of variance in the dataset. The testing (Levene Statistic) indicated (**Table 4.9**) that there is homogeneity of variance ($p > 0.05$) for all variables. This demonstrates that the ANOVA test is applicable.

Table 4.9: Test of homogeneity of variances for intervals.

	Levene Statistic	df1	df2	Sig.
Total_Correct_Clicks	0.98	3	92	0.406
Total_Missed_Clicks	0.98	3	92	0.407
Performance_Score	0.20	3	92	0.895
Correctly_Answered_Questions	5.63	3	92	0.001
Incorrectly_Answered_Questions	0.18	3	92	0.906
SAGAT_QA_Score	2.70	3	92	0.050
Target_Missed_Fixation_Count	2.32	3	92	0.080
Target_Hit_Fixation_Count	0.60	3	92	0.616
Total_Fixation_Count	3.20	3	92	0.027
Eye_Tracker_Score	1.42	3	92	0.242

Table 4.10 depicts the ANOVA test's descriptive statistics. According to the descriptive data, individuals missed an average of four clicks while correctly clicking an average of ten aircraft in each period. The average performance score of the participants was 71.64% in each period, the average score on the SAGAT was 76.46% and the average eye tracker score was 77.5%. Between intervals, the highest number of aircraft resolved by a participant was 22 while the lowest was 2. The highest number of aircraft missed by a participant was 11 and the lowest was 0. These findings indicate that every subject resolved at least 2 aircrafts in a given period and resolved at most 22. Participants scored at most 100% for the performance score in a given interval while the least score obtained was 26.67%.

The highest SAGAT Q/A score acquired was 100% while the lowest was 0%. This can infer that at least one participant in a given interval was able to correctly answer all the questions whereas at least one participant failed to correctly answer all the questions. The average correctly answered questions were 2 while the incorrectly answered question average was 1. In a given interval, participants correctly answered at least 0 and at most 7 questions. For the incorrect questions, the minimum question incorrectly answered was 0 while the maximum was 3.

Participants on average fixated on 10 resolvable aircrafts (target hit fixation count) and missed 3. The minimum target hit fixation count and target miss fixation count was 0 for a given interval. The maximum target hit fixation count and target miss fixation count was 25 and 18. The total fixation count was an average of 4,686. The analysis of the eye tracker score showed that at least one participant could not fixate on any resolvable aircraft while at least one participant fixated on all resolvable aircrafts in a given interval.

Table 4.10: One-Way ANOVA test descriptive statistics.

	Minimum	Maximum	Mean	Std. Dev.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Interval	1	4	2.5	1.12	1.26	0	0.25	-1.37	0.49
Total_Correct_Clicks	2	22	9.61	3.88	15.04	0.58	0.25	0.35	0.49
Total_Missed_Clicks	0	11	3.83	2.62	6.88	0.7	0.25	-0.02	0.49
Performance_Score	26.67	100	71.64	17.11	292.84	-0.36	0.25	-0.42	0.49
Correctly_Answered_Questions	0	7	2.19	1.16	1.35	1.02	0.25	2.59	0.49
Incorrectly_Answered_Questions	0	3	0.65	0.74	0.55	1	0.25	0.67	0.49
SAGAT_QA_Score	0	100	76.46	27.51	756.53	-0.98	0.25	0.35	0.49
Target_Missed_Fixation_Count	0	18	3.05	4.32	18.7	1.44	0.25	1.25	0.49
Target_Hit_Fixation_Count	0	25	10.4	5.46	29.8	0.23	0.25	-0.15	0.49
Total_Fixation_Count	213	27808	4686.14	4943.84	24441543	2.47	0.25	7.84	0.49
Eye_Tracker_Score	0	100	77.5	30.48	928.87	-1.23	0.25	0.34	0.49

The F-statistic value of the total correct clicks between the intervals was 4.27 according to the One-Way ANOVA hypothesis test for the 8-minute intervals (**Table 4.11**). The total number of correct clicks did vary significantly across the groups ($p < 0.05$). For all missed clicks across intervals, the F-statistic value was 0.12, meaning there was no statistically significant difference ($p > 0.05$). The F-statistic for the

performance score was 1.3, meaning there was no statistically significant difference ($p > 0.05$).

Table 4.11: Results of one-way ANOVA test for the 8-minute intervals.

		Sum of Squares	df	Mean Square	F	Sig.
Total_Correct_Clicks	Between Groups	174.53	3	58.18	4.27	0.007
	Within Groups	1254.21	92	13.63		
	Total	1428.74	95			
Total_Missed_Clicks	Between Groups	2.58	3	0.86	0.12	0.947
	Within Groups	650.75	92	7.07		
	Total	653.33	95			
Performance_Score	Between Groups	1133.95	3	377.98	1.3	0.278
	Within Groups	26685.51	92	290.06		
	Total	27819.46	95			
Correctly_Answered_Questions	Between Groups	8.37	3	2.79	2.14	0.101
	Within Groups	120.25	92	1.31		
	Total	128.63	95			
Incorrectly_Answered_Questions	Between Groups	1.21	3	0.4	0.73	0.537
	Within Groups	50.75	92	0.55		
	Total	51.96	95			
SAGAT_QA_Score	Between Groups	3877.61	3	1292.54	1.75	0.163
	Within Groups	67992.91	92	739.05		
	Total	71870.51	95			
Target_Missed_Fixation_Count	Between Groups	53.7	3	17.9	0.96	0.417
	Within Groups	1723.04	92	18.73		
	Total	1776.74	95			
Target_Hit_Fixation_Count	Between Groups	314.04	3	104.68	3.83	0.012
	Within Groups	2516.92	92	27.36		
	Total	2830.96	95			
Total_Fixation_Count	Between Groups	65790042	3	21930014	0.89	0.447
	Within Groups	2.30E+09	92	24523441		
	Total	2.30E+09	95			
Eye_Tracker_Score	Between Groups	2337.01	3	779	0.83	0.478
	Within Groups	85905.34	92	933.75		
	Total	88242.35	95			

The SAGAT score had a F-statistic value of 1.75. As with the other results, there was no statistically significant difference ($p > 0.05$) in the SAGAT score between the groups. The F-statistic for correctly answered and incorrectly answered questions were 2.14 and 0.73. These two variables also showed no significant differences ($p > 0.05$).

The F-statistic for target missed fixation count, target hit fixation count and total fixation count were 0.96, 3.83 and 0.89, respectively. The F-statistic for the eye tracker was 0.83, and was not statistically significant difference ($p > 0.05$). The only eye

metric parameter that showed statistically significant difference was the target hit fixation count ($p < 0.05$). For the total number of correct clicks and target hit fixation count that were found to be statistically different, a Tukey post-hoc test was performed.

The post hoc test findings (**Table 4.12**) indicated a significant difference in the number of aircraft resolved provided for interval 1 compared to interval 3 ($p < 0.05$). There was no statistically significant difference between intervals 1 and 2, interval 2 and interval 3, interval 3 and interval 4, interval 1 and interval 4, as well as between intervals 2 and interval 4, with a p-value greater than 0.05.

In summary, the statistical analysis demonstrates that there was only a significant difference in the resolving of aircraft in interval 1 compared to interval 3. No other statistical significance was exhibited between the other intervals.

Table 4.12: Tukey post-hoc test for correct clicks.

(I) Interval	(J) Interval	Mean Difference (I - J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	1.38	1.07	0.572	-1.41	4.16
	3	3.63	1.07	0.005	0.84	6.41
	4	2.54	1.07	0.087	-0.25	5.33
2	1	-1.38	1.07	0.572	-4.16	1.41
	3	2.25	1.07	0.157	-0.54	5.04
	4	1.17	1.07	0.694	-1.62	3.96
3	1	-3.63	1.07	0.005	-6.41	-0.84
	2	-2.25	1.07	0.157	-5.04	0.54
	4	-1.08	1.07	0.74	-3.87	1.71
4	1	-2.54	1.07	0.087	-5.33	0.25
	2	-1.17	1.07	0.694	-3.96	1.62
	3	1.08	1.07	0.74	-1.71	3.87

The post hoc test findings (**Table 4.13**) indicated a significant difference in the number of resolvable aircrafts that the user looked at (target hit fixation count) for interval 1 and interval 3 as well as interval 1 and interval 4 ($p < 0.05$). There was no statistically significant difference found between the other intervals because the p-value was greater than 0.05.

Table 4.13: Tukey post-hoc test for target hit fixation count.

(I) Family	(J) Family	Mean Difference (I - J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	3.29	1.51	0.136	-0.66	7.24
	3	4.17	1.51	0.035	0.22	8.12
	4	4.63	1.51	0.015	0.67	8.58
2	1	-3.29	1.51	0.136	-7.24	0.66
	3	0.88	1.51	0.938	-3.08	4.83
	4	1.33	1.51	0.814	-2.62	5.28
3	1	-4.17	1.51	0.035	-8.12	-0.22
	2	-0.88	1.51	0.938	-4.83	3.08
	4	0.46	1.51	0.99	-3.49	4.41
4	1	-4.63	1.51	0.015	-8.58	-0.67
	2	-1.33	1.51	0.814	-5.28	2.62
	3	-0.46	1.51	0.99	-4.41	3.49

5. CONCLUSIONS

This study investigated the potential of eye movements being a better situation awareness indicator than other methods such as SAGAT. The malfunction of the eye tracker led to a different aim in this study. The current research's goal was to investigate if performance measurements had a correlation with the measurements made by SAGAT. Aviation is one industry where awareness is essential, hence this domain was considered.

In the experiment, 24 subjects participated in a simulation where the goal to detect colliding aircraft. During the simulation, the participant was asked at random times to answer questions (SAGAT queries) regarding the task at hand. For each interval that lasted 8 minutes, the user's performance score and SAGAT score were evaluated. The SAGAT questions consisted of 5 different question types. The most questions that would be asked during the entire experiment was 17 where the least number of questions was 7. There showed a decrease of correct clicks and target hit fixation count between intervals 1 through 3 which could indicate mental fatigue and attention decline.

The results showed that the correlations between the total number of correct clicks and SAGAT score, had a weak and positive correlation ($r = 0.207$). The total number of missed clicks and SAGAT score had a very weak and negative correlation ($r = -0.064$). The SAGAT score and performance score had a weak and positive correlation ($r = 0.121$). The total number of correct clicks and total number of missed clicks had a moderate and negative correlation ($r = -0.438$). The results of weak correlations could be the fact that the user had a hard time recalling the situation of the aircrafts which the participants did mention themselves. Target hit fixation count and target missed fixation count showed a strong and negative correlation ($r = -0.8$). This can be interpreted as while the users fixated more on resolvable aircrafts, the less they missed crucial incidents. There was a moderate and positive correlation ($r = 0.307$) found between target hit fixation count and total correct clicks. This might suggest that the more participants fixated on resolvable aircrafts, the more resolvable aircrafts they clicked.

Compared to the correlation between the SAGAT score and performance score, the correlation between performance score and eye tracker score showed a higher correlation magnitude ($r = -0.207$). These results show that a weak and negative correlation was found between performance score and eye tracker score. A positive correlation found between the performance score and SAGAT score indicates that participants with higher performance scores also demonstrated higher SAGAT scores. This result implies that situation awareness (SA), as determined by the SAGAT approach, may have a moderate impact on task performance. Negative correlation between the performance score and eye tracker score can be due to eye tracking data containing errors or noise. This could happen if the calibration wasn't correct or if the eye tracker is sensitive to slight head movements. Data from eye trackers may be impacted by the testing environment. For example, inconsistent eye tracker results that don't match performance could be caused by inadequate lighting or distractions in the environment.

According to the results, the initial hypothesis, “Eye movement (tracking) techniques provide a better measurement of situational awareness in Air Controllers than probe related techniques such as SAGAT”, is rejected because no significant differences could be found between the correlations of performance score, SAGAT score and eye tracker score.

In the One-Way ANOVA test while evaluating the intervals of the experiment, there showed no significant difference in missed clicks, performance score, SAGAT score, correctly answered questions, incorrectly answered questions, target missed fixation count, total fixation count, and eye tracker score ($p > 0.05$) between intervals. Based on these ANOVA results, the variable correct clicks showed a significant difference among the interval groups, as indicated by a p-value less than 0.05. A Tukey post-hoc test was executed and it was found that only interval 1 and interval 3 showed statistically significant difference while the other interval groups showed no statistically significant difference. The total hit fixation count variable also showed a significant difference ($p < 0.05$) among the interval groups. The Tukey post-hoc test showed there was a statistically significant difference between interval 1 and interval 3, as well as interval 1 and interval 4.

In the experiment, only the aircraft's altitude and position were present in the user interface. More detailed information like speed or heading could have maybe increased

the chances of the user recalling the situation. It is recommended that a large collection of questions should be picked and asked in SAGAT because participants usually start to memorize and prepare for the upcoming questions. This can alter the true measurement of situation awareness. During the experiment, difficulties were encountered with calibration while using the eye tracker. The calibration issues might have affected the accuracy of the eye tracking data, making it harder to draw definitive conclusions about its effectiveness relative to situation awareness measurement.

This research has shown that there is some correlation between performance, SAGAT scores and eye tracker scores but showed no significant differences ($p > 0.001$). There is no variation in missed click, performance score, SAGAT score and eye tracker score in between the 4 intervals measured. Participant's correct click number showed variation between intervals 1 and interval 3. This research also raises the question of how SAGAT may not be feasible in real life scenarios. Methods such as eye tracking may be more feasible in situations where task interruption is not permitted because of life critical situations. Further research on the comparisons of direct and indirect situation awareness methods is needed to increase the precision of situation awareness measurement.

REFERENCES

- Argyle, E. M., Gourley, J. J., Kang, Z., & Shehab, R. L. (2020). Investigating the relationship between eye movements and situation awareness in weather forecasting. *Applied ergonomics*, 85, 103071.
- Aydin, T., & Gokturk, M. (2018). Measurement of situational awareness in the supervisory tasks. In 2018 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-6). IEEE.
- Behrend, J., & Dehais, F. (2020). How role assignment impacts decision-making in high-risk environments: Evidence from eye-tracking in aviation. *Safety science*, 127, 104738.
- Bjørneseth, F. B., Clarke, L., Dunlop, M., & Komandur, S. (2014). Towards an understanding of operator focus using eye-tracking in safety-critical maritime settings. In *International Conference on Human Factors in Ship Design & Operation*.
- Center, E. E. (2000). Situation Awareness-Synthesis of Literature Search. EEC Note, (16/00).
- de Winter, J. C., Eisma, Y. B., Cabrall, C. D. D., Hancock, P. A., & Stanton, N. A. (2019). Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology & Work*, 21(1), 99-111.
- Desvergez, A., Winer, A., Gouyon, J. B., & Descoins, M. (2019). An observational study using eye tracking to assess resident and senior anesthetists' situation awareness and visual perception in postpartum hemorrhage high fidelity simulation. *PloS one*, 14(8), e0221515.
- Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D., & Manning, C. A. (1998). Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly*, 6(1), 1-20.
- Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). In Proceedings of the IEEE 1988 national aerospace and electronics conference (pp. 789-795).
- Endsley, M. R. (2021). A systematic review and meta-analysis of direct objective measures of situation awareness: a comparison of SAGAT and SPAM. *Human factors*, 63(1), 124-150.
- Endsley, M. R., & Garland, D. J. (Eds.). (2000). Situation awareness analysis and measurement. CRC press.
- Fujino, M., Lee, J., Hirano, T., Saito, Y., & Itoh, M. (2020, December). Comparison of sagat and spam for seeking effective way to evaluate situation awareness and workload during air traffic control task. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 64, No. 1, pp. 1836-1840). Sage CA: Los Angeles, CA: SAGE Publications.

- Garcia, G. L. (2016). Understanding and defining situational awareness and interpersonal competence as essential evaluator competencies. PhD Dissertation, Graduate School of University of Illinois at Urbana-Champaign, Illinois, USA, pp. 138.
- Glaholt, M. G. (2014). Eye tracking in the cockpit: a review of the relationships between eye movements and the aviator's cognitive state. Defence Research and Development Toronto (Canada).
- Häggström, C., Englund, M., & Lindroos, O. (2015). Examining the gaze behaviors of harvester operators: an eye-tracking study. *International journal of forest engineering*, 26(2), 96-113.
- Hasanzadeh, S., Esmaeili, B., & Dodd, M. D. (2016). Measuring construction workers' real-time situation awareness using mobile eye-tracking. In Construction research congress (pp. 2894-2904).
- Hasanzadeh, S., Esmaeili, B., & Dodd, M. D. (2018). Examining the relationship between construction workers' visual attention and situation awareness under fall and tripping hazard conditions: Using mobile eye tracking. *Journal of construction engineering and management*, 144(7), 04018060.
- Hasse, C., Grasshoff, D., & Bruder, C. (2012). Eye-tracking parameters as a predictor of human performance in the detection of automation failures. In *Proceedings HFES Europe Chapter Conference Toulouse* (pp. 133-144).
- Jones, D. G., & Endsley, M. R. (1996). Sources of situation awareness errors in aviation. *Aviation, space, and environmental medicine*, 67(6), 507-512.
- Kent State University Libraries. (2023). SPSS tutorials: Paired Samples t Test. Retrieved January 28, 2024, from <https://libguides.library.kent.edu/spss/pairedsamplesttest>.
- Kent State University Libraries. (2023). SPSS tutorials: Pearson Correlation. Retrieved January 28, 2024, from <https://libguides.library.kent.edu/SPSS/PearsonCorr>.
- Kok, E. M., & Jarodzka, H. (2017). Before your very eyes: The value and limitations of eye tracking in medical education. *Medical education*, 51(1), 114-122.
- Li, W. C., Zhang, J., Le Minh, T., Cao, J., & Wang, L. (2019). Visual scan patterns reflect to human-computer interactions on processing different types of messages in the flight deck. *International Journal of Industrial Ergonomics*, 72, 54-60.
- Martin, C., Cegarra, J., & Averty, P. (2011). Analysis of mental workload during en-route air traffic control task execution based on eye-tracking technique. In *Engineering Psychology and Cognitive Ergonomics: 9th International Conference, EPCE 2011, Held as Part of HCI International 2011, Orlando, FL, USA, July 9-14, 2011. Proceedings 9* (pp. 592-597). Springer Berlin Heidelberg.

Martinez-Marquez, D., Pingali, S., Panuwatwanich, K., Stewart, R. A., & Mohamed, S. (2021). Application of eye tracking technology in aviation, maritime, and construction industries: a systematic review. *Sensors*, 21(13), 4289.

Matthews, M. D., & Beal, S. A. (2002). Assessing situation awareness in field training exercises. US Army Research Institute for the Behavioral and Social Sciences, 31.

McGuinness, B., & Foy, L. (2000). A Subjective Measure of SA The Crew Awareness Rating Scale-GetInfo. In Proceedings of the first human performance, situation awareness, and automation conference, Savannah, Georgia, USA.

Moore, K., & Gugerty, L. (2010). Development of a novel measure of situation awareness: The case for eye movement analysis. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 54, No. 19, pp. 1650-1654). Sage CA: Los Angeles, CA: SAGE Publications.

Muczyński, B., & Gucma, M. (2013). Application of eye-tracking techniques in human factor research in marine operations. Challenges and methodology. *Zeszyty Naukowe/Akademia Morska w Szczecinie*. pp.116-120.

O'Brien, K. S., & O'Hare, D. (2007). Situational awareness ability and cognitive skills training in a complex real-world task. *Ergonomics*, 50(7), 1064-1091.

Orlandi, L., & Brooks, B. (2018). Measuring mental workload and physiological reactions in marine pilots: Building bridges towards redlines of performance. *Applied Ergonomics*, 69, 74-92.

Ratwani, R. M., McCurry, J. M., & Trafton, J. G. (2010). Single operator, multiple robots: an eye movement based theoretic model of operator situation awareness. In 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 235-242).

Salotti, J. M., & Suhir, E. (2019). Degraded situation awareness risk assessment in the aerospace domain. In 2019 IEEE 5th International Workshop on Metrology for AeroSpace (MetroAeroSpace) (pp. 39-43).

SHGM, (2022), Directorate General of Civil Aviation Annual Report 2022. Ankara, Turkey.

Shi, C., & Rothrock, L. (2023). Using eye movements to evaluate the effectiveness of the situation awareness rating technique scale in measuring situation awareness for smart manufacturing. *Ergonomics*, 66(8), 1090-1098.

Stankovic, A., Aitken, M. R., & Clark, L. (2014). An eye-tracking study of information sampling and decision-making under stress: Implications for alarms in aviation emergencies. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 58, No. 1, pp. 125-129). Sage CA: Los Angeles, CA: SAGE Publications.

Taylor, R. M. (2017). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In *Situational awareness* (pp. 111-128). Routledge.

US Department of Labor, Bureau of Statistics Staff, & United States. Bureau of Labor Statistics. (2000). Occupational outlook handbook. Bernan Press (PA).

Van De Merwe, K., Van Dijk, H., & Zon, R. (2012). Eye movements as an indicator of situation awareness in a flight simulator experiment. *The International Journal of Aviation Psychology*, 22(1), 78-95.

van den Beukel, A. P., & van der Voort, M. C. (2017). How to assess driver's interaction with partially automated driving systems—A framework for early concept assessment. *Applied Ergonomics*, 59, 302-312.

van Dijk, H., van de Merwe, K., & Zon, R. (2011). A Coherent Impression of the Pilots' Situation Awareness: Studying relevant human factors tools. *The International Journal of Aviation Psychology*, 21(4), 343-356.

Web 1, (2017). EUROCONTROL Guidelines for short term conflict alert (STCA) EUROCONTROL-GUID-159. <https://www.eurocontrol.int/publication/eurocontrol-guidelines-short-term-conflict-alert-stca>, (Last Accessed Date: 8/10/2023)

Web 2, (2017). EUROCONTROL Guidelines for minimum safe altitude warning (MSAW)EUROCONTROL-GUID-160. <https://www.eurocontrol.int/publication/eurocontrol-guidelines-minimum-safe-altitude-warning-msaw>, (Last Accessed Date: 8/10/2023)

Web 3, (2017). EUROCONTROL Guidelines for area proximity warning (APW) EUROCONTROL-GUID-16. <https://www.eurocontrol.int/publication/eurocontrol-guidelines-area-proximity-warning>, (Last Accessed Date: 8/10/2023)

Web 4, (2018). Personal Licences and Training for Air Traffic Controllers. <https://web.shgm.gov.tr/en/personali-licences-and-training/2129-air-traffic-controller>, (Last Accessed Date: 17/10/2023)

Web 5, (2023). ATC Tower Control and Approach/Enroute Control Radar Simulator <https://bilgem.tubitak.gov.tr/atctrsim/>, (Last Accessed Date: 28/10/2023)

Web 6, Waidyanatha, N. (2020, August 17). Situational Awareness can Save People with Disabilities in Crises <https://lirneasia.net/2020/08/situational-awareness-PWD-crises/>, (Last Accessed Date: 22/2/2024)

Wee, H. J., Trapsilawati, F., Lye, S. W., Chen, C. H., & Pinheiro, J. P. (2017). Real time bio signal interface for visual monitoring of radar controllers. In ISPE International Conference on Transdisciplinary (pp. 394-401).

Wickens, C. D. (1996). Situation awareness: impact of automation and display technology. Situation awareness: limitations and enhancement in the aviation environment, k2-1. AGARD-CP-575, Canada.

Xinyao, G., Yawei, L., Qingmin, S., Linqing, N., & Huibin, J. (2020). Measuring the Situation Awareness of Tower Controllers by Using Eye Movement Analysis. *Journal of Engineering Science & Technology Review*, 13(1). 133 - 140

Yan, S., Wei, Y., & Tran, C. C. (2019). Evaluation and prediction mental workload in user interface of maritime operations using eye response. *International Journal of Industrial Ergonomics*, 71, 117-127.

Zhang, T., Yang, J., Liang, N., Pitts, B.J., Prakah-Asante, K., Curry, R., Duerstock, B., Wachs, J.P. and Yu, D., (2023). Physiological measurements of situation awareness: a systematic review. *Human factors*, 65(5), 737-758.

Ziv, G. (2016). Gaze behavior and visual attention: A review of eye tracking studies in aviation. *The International Journal of Aviation Psychology*, 26(3-4), 75-104.

BIOGRAPHY

Nesibe YAVUZ attended elementary and middle school in Syracuse. She along with her family moved back to Turkey in 2009. After graduating high school in 2012, she was accepted into the undergraduate program in the Computer Engineering Department at Karadeniz Technical University. After graduating in 2016, she was hired as a software engineer at TÜBİTAK BİLGEM in Gebze, Kocaeli, where she continues to work. Her focus is on aviation, specifically developing safety-critical software for the aviation domain. In 2020, she began her master's studies at Gebze Technical University.

PUBLICATIONS AND PRESENTATIONS FROM THE THESIS

Yavuz, N. and Göktürk, M. (2023). Assessing situation awareness with eye movement. In proceeding of GTU 7th Annual Symposium (p. 19). May 31 - June 1, 2023, Gebze, Kocaeli, Türkiye.

APPENDICES

APPENDIX-A: Problems and Solutions for Eye Tracker Device Usage

Table A.1: Problems and solutions for eye tracker device usage.

Device Name	Problem	Solution
Tobii X2-30	Disconnecting from computer because the capacitor had worn out (Figure A.1).	Replaced the capacitor (Figure A.2).
Tobii TX300	Device was not calibrating properly. Cameras had shifted causing inaccurate calibrations (Figure A.3).	Cameras have to be replaced (Figure A.4). No technical support is provided for this model, therefore the cameras could not be replaced.



Figure A.1: Casing opened for Tobii X2-30 device with original capacitor.



Figure A.2: Tobii X2-30 with replaced capacitor.



Figure A.3: Calibration issue for Tobii TX300 device.



Figure A.4: Casing opened for Tobii TX300 device