12-2009

Comparison of Eye Movement Data to Direct Measures of Situation Awareness for Development of a Novel Measurement Technique in Dynamic, Uncontrolled Test Environments

Kristin Moore
Clemson University, ksmoore@clemson.edu

COMPARISON OF EYE MOVEMENT DATA TO DIRECT MEASURES OF SITUATION AWARENESS FOR DEVELOPMENT OF A NOVEL MEASUREMENT TECHNIQUE IN DYNAMIC, UNCONTROLLED TEST ENVIRONMENTS

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Human Factors Psychology

---

by
Kristin Suzanne Moore
December 2009

Accepted by:
Leo Gugerty, Committee Chair
Christopher Pagano
Richard Pak
Andrew Duchowski

ABSTRACT

Situation awareness (SA) is a measure of an individual's knowledge and understanding of the current and expected future states of a situation. While there are numerous options for SA measurement, none are currently suitable in dynamic, uncontrolled environments. Direct measures of SA are the most common, but require a large amount of researcher control as well as the ability to stop operators during a task in order to ask questions about their levels of SA. The current research explored the relationship between direct measures of SA and eye tracking measures as a first step in the development of an unobtrusive SA measure to be used in less controllable, dynamic environments. Two studies compared participant eye movements and SA in driving and air traffic control scenarios. Both studies showed that the more individuals fixated on an important, task-relevant event, the higher their SA for that event. The studies also provide evidence that the way operators allocate attention (i.e., distributed widely or narrowly) affects their SA as well as their task performance. In addition, study 2 results showed positive correlations between SA and task performance. The results indicate that eye tracking may be a viable option for measuring SA in environments not conducive to current direct SA measurement techniques. Future research should continue to explore which eye movement variables best predict participant SA, as well as to investigate the relationship between attention allocation and SA.

DEDICATION

This dissertation is dedicated to Dr. William Moroney, without whom I would have never discovered human factors psychology. Thank you for mentoring me throughout my undergraduate education and encouraging me to rethink my graduate field of study. Whenever I had questions or needed guidance, you were always there with sound advice. You have been and continue to be an inspiration for what I can achieve in both my career and life in general.

This dissertation is also dedicated to my dad, who always believed in me and supported me every step of the way.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the help of numerous individuals. I would like to thank Dr. Leo Gugerty for his advisement throughout the dissertation process, as well as his insight and ability to think through issues I would not have recognized. I would also like to thank Dr. Chris Pagano for his continued guidance and support throughout my graduate school career. In addition, I would like to thank my other committee members, Dr. Richard Pak and Dr. Andrew Duchowski. A special thank you also goes to Dr. Jay Steele, who programmed the simulator for the first study, and wrote several programs to aggregate the data for both studies.

I would like to thank all of the air traffic controllers who helped in the development of the scenarios used for my experiment, especially Edward Hawk who spent many hours explaining the ins and outs of air traffic control to me. In addition, thank you to the controllers who were willing to participate in the actual experiment. I would also like to convey thanks to my bosses and coworkers at SPAWAR Atlantic who gave me time and space to finish my graduate work before bombarding me with additional work.

Finally, special thanks go to my family and friends, without whom I would have gone crazy a long time ago. In particular, my mom and sisters and my graduate school friends.

TABLE OF CONTENTS

Table of Contents (Continued)

Page

LIST OF TABLES

LIST OF FIGURES

List of Figures (Continued)

CHAPTER ONE

INTRODUCTION

Whenever a task is performed, no matter how small, a person must coordinate a myriad of cognitive and physical processes. Consider, for example, a person simply cleaning his kitchen. He must know how to clean, what type of cleaning products to use, where he should clean, which areas have already been cleaned, and what is left to be cleaned. The cognitive processes involved in the act of cleaning alone involve long term memory for what types of products to use, short term memory for what surfaces have already been cleaned and are yet to be cleaned, attention to continue the cleaning process, and so on. Other, more multifaceted tasks require a more complex set of mental processes with higher consequences for errors. For example, a pilot of a commercial aircraft must use short and long term memory, attention and decision making to safely navigate the aircraft from take-off to landing. While all of these constructs are important, they are not the only processes involved when completing dynamic tasks. A mistake during the flight could cause injuries or deaths, therefore it is important for researchers to have an intricate understanding of what processes are involved and how errors occur. In addition to the other processes involved, situation awareness (SA) is one construct that has consistently been correlated with performance on a variety of tasks in various domains.

The current study explores the construct of SA and its measurement in two task domains. The introduction will survey the research on SA and its components, as well as current measurement methods. Physiological measures are rarely used to measure SA;

the current research examines the relationship between eye tracking and direct measures of SA to determine if eye tracking is a viable measurement option when other options are not. Two studies will compare different eye movement measures and direct SA measures in both driving and air traffic control scenarios. Research examining both the construct of SA and methods of measurement follows.

SA has been a topic of interest since World War I (Press, 1986; as cited by Endsley, 1995c), but only in the past three decades has it been extensively researched. Many researchers have operationally defined and measured SA, but further discussion is needed to better understand its meaning. From a global perspective, SA is attending to and understanding what is occurring in the environment immediately surrounding an individual during a dynamic (i.e., changing) situation. Clearly, this is ambiguous and in need of further clarification. Although the construct had been implied previously, a widely accepted, formal definition of SA was not introduced until 1987. Endsley defines SA as, "The perception of the 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, 1987, 1988, 1995c). While this is the most cited definition of SA, there are several other viable interpretations (e.g., Adams, Tenney, & Pew, 1995; Smith & Hancock, 1995). The definition is still not without disagreement, but it is important to first understand why the construct of SA is even a relevant component of performance.

One way to illustrate the importance of SA is to describe situations where a loss of SA had negative consequences. A simple example of loss of SA is when an outfielder catches a fly ball but fails to throw a runner out because he does not realize it is not the last out of the inning (Tenney & Pew, 2006). Another, more severe example is the death of almost 5000 people between 1978 and 1992 from airplane accidents due to controlled flight into terrain. A lack of SA was determined to be the cause of 74% of those accidents (Woodhouse & Woodhouse, 1995; from Durso & Gronlund, 1999). In general, having an understanding of the past, present, and future components of a situation should lead to better performance, with a loss of this understanding potentially resulting in devastating consequences in high risk tasks. Even if the outcome is not catastrophic, costly errors may result from a loss of SA. Researchers continue to operationally define and measure the SA construct with the ultimate goal of designing interfaces and implementing training procedures that will increase operator SA and reduce human error.

Situation awareness research has been conducted in a variety of real-time, dynamic domains including air traffic control (ATC) (e.g., Endsley & Smolensky, 1998; Durso, Truitt, Hackworth, Crutchfield, & Manning, 1998b), aviation (e.g., Kaber, Endsley, Wright, & Warren, 2002), anesthesiology (e.g., Gaba, Howard, & Small, 1995), nuclear power plants (e.g., Hogg, Folleso, Strand-Volden, & Torralba, 1995), driving (e.g., Gugerty, 1997), military command and control (e.g., Gorman, Cooke, & Winner, 2006; Salmon et al., 2007; Stanton et al., 2006), and even football (e.g., Walker & Fisk, 1995). Endsley (1987, 1988, 1995c) distinguishes between three levels of SA: Level 1 – perception, Level 2 – comprehension, and Level 3 – projection. An example of the three levels of SA from ATC would be a controller perceiving the number of aircraft in a particular airspace on the radar screen (Level 1), integrating information about an aircraft's heading, altitude and airspeed in order to comprehend that it is beginning its arrival approach (Level 2), and projecting how long it will take the aircraft to reach its destination (Level 3). Though Endsley's definition may be the most widely accepted, it is by no means the only definition of SA. Additionally, her definition is not complete, as it is difficult to define in detail a complex construct. In that sense, SA is similar to mental constructs such as attention, memory, and consciousness. All are complex, difficult to define wholly, not directly observable and not without disagreement among experts.

Some researchers question whether SA should even be considered a psychological construct, separate from other clearly defined constructs (Crane, 1992; Dekker & Hollnagel, 2004). To those that criticize, the continued use and application of SA is a testament to its importance beyond already existing constructs (Wickens, 2008). Though most agree SA is a construct separate from others, researchers continue to debate the definition, and in turn, the processes which affect development and maintenance of SA.

Two frameworks, the information processing approach and the ecological view, are typically the basis of theories of SA. Endsley's definition of SA is based on the information processing approach, where SA is viewed as a product of a number of cognitive processes (Durso & Gronlund, 1999). Flach (1995) and others (Smith & Hancock, 1995; Adams, Tenney & Pew, 1995) advocate a more holistic, ecological approach to situation awareness, one that is based upon the perception-action cycle (Neisser, 1976). The ecological view defines SA as both a product and a process of the perception-action cycle (Durso & Gronlund, 1999). In Smith and Hancock's (1995) ecological view, SA is defined as "adaptive, externally directed consciousness" and is "directly related to stress, mental workload, and other energetic constructs that are facets of consciousness" (1995, pg. 138). Even though the theoretical framework of SA continues to be debated, specifically what components and processes should or should not be included in the definition, the processes which make up the information processing approach have been studied in a variety of task domains and add to the understanding of SA.

Perceptual and cognitive processes and structures in SA

The information-processing approach describes behavior and cognition underlying behavior in terms of processes (such as attention, comprehension, or memory retrieval) and the states of knowledge produced by these processes (such as a consciously recognized object or a retrieved memory). Applying this general approach to the dynamic situations addressed by SA, SA is viewed as knowledge of the current and expected future states of a situation (SA as knowledge or product) and is comprised of set of attentional and comprehension processes that gather, interpret and update this knowledge. Previous experiences and training, among other things, will affect knowledge of the current situation as well as what is expected to occur in the future. This view of SA as both processes and knowledge produced by these processes is exemplified in the following description: "By defining SA as a generative process of knowledge creation and informed action taking, we expressly deny that SA is merely a snapshot of the agent's current mental model. Rather, SA guides the process of modifying knowledge – that is, of constructing a representation of current and likely events" (Smith & Hancock, 1995, pg. 142).

However, Endsley (1995c) and others are careful to point out that SA does not involve processes underlying decision making and response execution; and although it may be influenced by constructs such as workload, working memory and attention, it is independent from them. Endsley (1995c) explains that if these constructs become a part of the definition of SA, its independence will be lost. One study found no relationship between mental workload and SA in a review of 23 experiments (Vidulich, 2000). After dividing the studies by interface manipulation type, varied results were found. When researchers added information to an interface to improve SA, the resulting mental workload scores were mixed. When researchers simply rearranged the available information, a majority of the studies found an increase in SA and a decrease in mental workload. Thus, in certain circumstances mental workload and SA may co-vary, but little consistency between the two constructs has been found.

Endsley's (1987, 1988, 1995c) high-level model of how SA fits into the stages of information processing is illustrated in Figure 1.1.

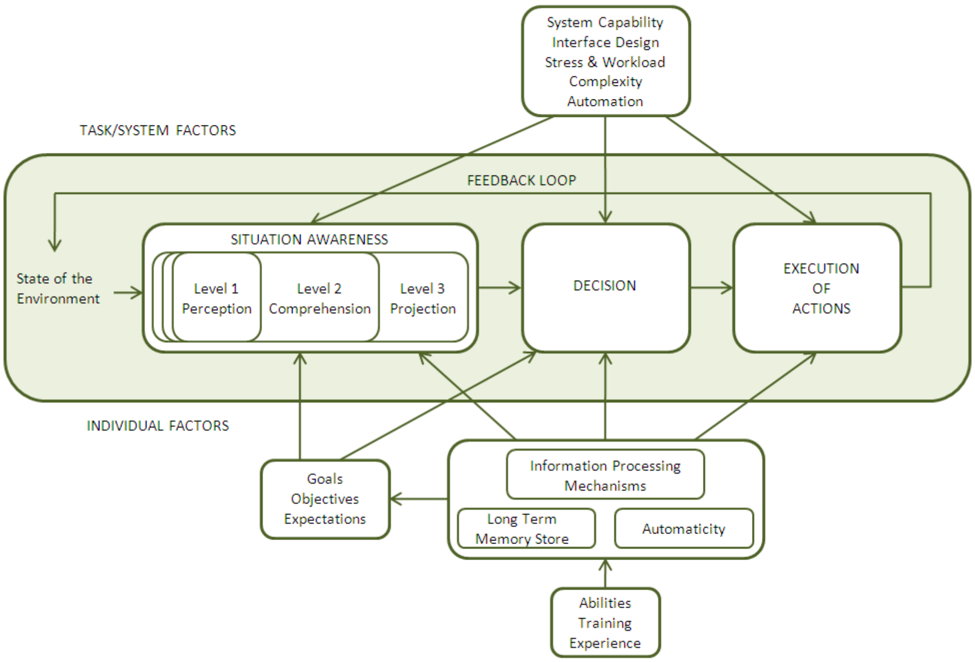


Figure 1.1. Endsley's model of situation awareness (adapted from Endsley, 1995c).

The model illustrates that multiple components are involved in the development and maintenance of SA. Even when people experience the same situation in the same environmental conditions, individual differences will likely lead to varying levels of SA due to variations in ability, experience and training. In addition, each of their specific goals and expectations will affect their perceptions. System factors also affect SA; if the system does not provide all of the necessary information for complete understanding of the environment, an individual is not going to be able to achieve higher levels of SA, regardless of other factors. Finally, environmental factors, such as varying levels of stress, will affect SA in different ways (Endsley, 1995c).

Recall that there are three levels of SA as described by Endsley. Level 1 SA is defined as the perception of elements in the environment and can be thought of as analogous to “word-level information prior to combining the words into phrases” (Durso & Gronlund, 1999, pg. 291). Level 2 SA is defined as the comprehension of the current situation. Comprehension occurs through the synthesis of the elements perceived in Level 1. Level 2 reflects the idea that the outcome of many perceptual processes is the recognition or comprehension of a meaningful object or event. Level 3 SA is defined as the projection of future status, and reflects the fact that the meaning of many dynamic events cannot be comprehended without anticipating how these events will play out in the near future.

The three levels of SA in Endsley’s definition are very broad in describing the high-level processes of perception, comprehension, and projection underlying SA. Several researchers have studied these processes in more detail in a variety of research domains. Perception and comprehension can be examined by considering what leads one to perceive and comprehend. Both the SEEV model of attention allocation and the Construction Integration model explore the components of perception and comprehension in more detail.

The SEEV model of how focal attention is allocated in real-time tasks is made up of four elements comprising the acronym SEEV – Salient events, Effort, Expectancy, and the Value of events (Wickens et al., 2005). The SEEV model includes both bottom-up and top-down processes. The salience of events in the environment is determined by their ability to capture attention in a bottom-up fashion. Effort (E), Expectancy (E) and Value (V) are top-down processes determined by operator understanding of the situation and previous experience among other factors. Effort refers to the physical difficulty of shifting attention to an object, e.g., the length of a saccade or head movement. Expectancy is proportional to how frequently information about an object is changing. As the frequency of information change increases, an operator will sample the object more often to attempt to avoid missing relevant information. Value refers to the priority or importance of an object. As the value increases, again the sampling should increase due to the higher importance level. The SEEV model predicts that people will allocate more attention to salient, high-value objects that are changing rapidly and that are easy to attend to. The model has been partially validated by empirical studies of driving whose results show that as the value and the rate of information change of objects increases, people allocate more attention to those objects (Horrey, Wickens & Consalus, 2006).

The construction-integration model was developed to better understand discourse comprehension (Kintsch, 1988, 2005). While the model has been primarily applied to how discourse is comprehended, it is also applicable to how information in dynamic environments is comprehended (Durso, Rawson & Girotto, 2007). In Kintsch's (1988) view, comprehension of words and sentences begins as a bottom-up process; the context is not considered until later stages. In the first stage, the sense-selection stage, when a word is read a network of propositions and connections are formed without consideration of the context. In this stage, understanding begins by rapidly reducing the number of potential word meanings to a manageable number; the potential meanings are initially selected based on a context-free approach to the meaning of the particular sentence component. In the second stage, which involves top-down processing, associations with the context (e.g., nearby words) helps reduce the number of potential meanings further.

This is the sense-elaboration stage. In the final stage, further understanding occurs based upon the long-term memory knowledge-base of the operator (further top-down processing). Someone with robust knowledge will likely obtain a quicker and more sound understanding as a situation progresses. The construction-integration model is aptly named because comprehension is made up of the integration of an understanding of word meaning constructed from what is in the environment (bottom-up) as well as what the operator already knows (top-down) (Kintsch, 1988). It is important to understand that this process is cyclical due to the limited cognitive capacity of humans. In terms of text comprehension, cycles are typically at the sentence level; integration occurs when nodes from one cycle are carried over and integrated into the next (Durso et al., 2007; Kintsch, 1988)

Durso et al. (2007) point out that the construction-integration model is analogous to the way operators develops SA over time through the bottom-up process of perception of information in the environment as well as the top-down processes of developing a situational model using environmental context and their own knowledge base. Operators must develop an *eventbase* in order to construct a representation of the environment around them (Durso et al., 2007). If SA develops in the same way as discourse comprehension, eventbase development begins through a strictly bottom-up process, similar to the salience component of the SEEV model. The integration of elements obtained from Level 1 perception would be initially context-free, with operator knowledge aiding in the winnowing out and eventual selection of an event meaning (i.e., comprehension). The need for context to guide comprehension may partially explain why SA must be built up over time and is not instantly obtained. Top-down processing is needed to suppress irrelevant information and only allow the appropriate meaning in a specific context to appear.

By looking at the process of developing SA from the perspective of the SEEV model and the Construction-Integration model, the following view of SA emerges. SA will be improved to the extent that operators use cues like task-priority and rate of information change to guide their attention allocation to dynamic events as these events change over space and time. Then once a high-priority event is focused on, SA will be improved to the extent that operators' comprehension process allows quick and accurate comprehension of this event.

Turning from the processes used to maintain SA to the cognitive structures underlying SA, the product of comprehension is commonly thought to be stored and updated in a situation model residing in working memory. It is easiest to understand situation models in the context of text comprehension, which is made up of both situation models and textbase. The textbase consists of the elements that allow an individual to have a word-level understanding of the text, or understanding simply the words without any additional inputs. The situation model of an individual is necessary to interpret and have a higher understanding of the words and their relationship to one another to form meaning and make the text coherent. The components involved in a situation model include an understanding of the language, knowledge of the world, and past experiences of the individual (Kintsch, 1998). As pointed out by Durso et al. (2007), the construct of a situation model should translate well from comprehending text to comprehending real-time situations.

Effects of SA on Performance

While it is important to understand the theoretical underpinnings of SA, it is equally important to understand how not obtaining higher SA or losing it once obtained can affect performance. Endsley (1995a) reviewed 24 accident reports from the National Transportation Safety Board (NTSB) from 1989 – 1992. Of the 24 accidents, it was determined that 17 were the result of human error with 15 of those related to SA. A further analysis of the accidents involving SA revealed that there were 32 SA errors (several accidents involved more than one error). From these reports, a taxonomy of errors was developed, with the number of recorded errors for each failure listed in Table 1.1 below (From Endsley, 1995a).

Table 1.1
Frequency and percentage of errors leading to accidents between 1989 and 1992

	Frequency	*Percentage*
Level 1: Failure to correctly perceive information	**23**	**71.9**
• Data not available	3	9.4
• Data difficult to detect or perceive	5	15.6
• Failure to monitor or observe data	10	31.3
• Misperception of data	4	12.5
• Memory failure	1	3.1
Level 2: Failure to comprehend situation	**7**	**21.9**
• Lack of or poor mental model	1	3.1
• Use of incorrect mental model	2	6.3
• Over-reliance on default values in mental model	1	3.1
• Other	3	9.4
Level 3: Failure to project situation into the future	**2**	**6.3**
• Lack of or poor mental model	1	3.1
• Overprojection of current trends	1	3.1

One of the key indicators of SA while driving involves hazard perception (Horswill & McKenna, 2004). In driving tasks, hazard perception is the only skill that has correlated with performance across numerous studies. A review of the literature on hazard perception and performance revealed that hazard perception ability is a good predictor of on-road crashes. In a large scale study of 100,000 drivers that measured the predictability of a hazard perception test, Hull and Christie (1992) found that drivers who scored low on the test were twice as likely as those who scored high to be involved in a fatal accident within one year (Horswill & McKenna, 2004). In driving research, SA (measured by hazard perception) has continually been positively correlated with good driving performance.

Even though one might assume that high levels of SA would be equated with higher performance levels, this is not always the case. Instead, SA should be viewed as a factor that affects performance, with high SA typically, though not always, leading to high levels of performance (Endsley, 1995c). A high level of SA can occur during low levels of performance and vice versa. For example, a novice system operator may be aware of a problem but may not have the expertise to solve it before an error occurs. Also, with high levels of automation, system performance may be high even if an operator experiences a loss of SA.

Situation awareness measurement techniques

There are a large variety of measurement techniques that have been employed to determine an individual's level of SA. Three types of methods are typically discussed: subjective measures, implicit measures, and explicit (direct) measures (Sarter & Woods, 1995). A fourth method of measurement that has received considerably less attention, but warrants further investigation, is the use of physiological measures. Each method has several advantages and disadvantages; currently no method is clearly superior to the others.

Subjective measures (including self- and observer-rating techniques) simply determine an operator's SA by asking the operator after the task is completed or by having a subject matter expert (SME) observe the operator and rate his SA. The most common subjective SA measure is the Situation Awareness Rating Technique (SART) (Taylor, 1990). Subjective measures, such as the SART, are favorable because they are relatively easy to implement and do not require a large amount of preparation beforehand. In addition, they can be used in dynamic, field-based research. There are several drawbacks to subjective measures; the main one being that studies have shown that SART neither correlates with performance or other measures of SA (Endsley, 1995b; Salmon et al. 2008a). Other issues with subjective ratings include the possibility that participants' task performance may affect SA ratings afterward. Participants may, for example, take the result of the task (i.e., pass or fail) and rate their SA based on their performance. Additionally, participants may not have an understanding of what their true SA is, believing that they were very aware when in fact they missed pertinent information in the environment. Observer ratings (typically from SMEs) are also not ideal because SA is an internal construct, making it inherently difficult to observe in others (Endsley, 1995b; Salmon, Stanton, Walker, & Green, 2006; Sarter & Woods, 1995).

Implicit performance measurements are imbedded within the primary task: high levels of SA are assumed when operators' actions indicate that they found the imbedded information. They are based on the assumption that performance is directly related to SA (Sarter & Woods, 1995). Researchers choose imbedded tasks that should lead operators with low SA to perform poorly and operators with high SA to perform well (Durso & Gronlund, 1999; Sarter & Woods, 1991; Wickens, 1996). For example, Gugerty (1997) imbedded SA measures in driving simulation tasks by measuring participant ability to detect hazards and to detect cars in lanes to their right or left. Detection of hazards and cars indicated that participants had adequate SA of their environment. The main benefit of implicit performance measures is that they are not intrusive because they are imbedded in the primary task. The main drawback is that it is difficult to parse out SA from the additional factors affecting performance on the imbedded tasks. Given that SA is not directly related to performance, it is possible to have good SA with poor performance and vice versa.

Explicit performance measures were developed to specifically measure SA during a task. The most common explicit measures of SA are query methods, which are developed based on the task or scenario and measure the SA of an operator by asking them situation specific questions throughout the task (Durso, Bleckley & Dattel, 2006). A frequently used query method is the Situation Awareness Global Assessment Technique (SAGAT) (Endsley, 1988, 1990). Researchers use the SAGAT to measure SA by blanking the screen at unpredictable times throughout the scenario and asking questions related to the three levels of SA. Another query measure, the Situation Present

Assessment Method (SPAM), is similar to SAGAT except the screen is not blanked and the scenario is still visible (Durso et al., 1995). In SPAM, the scenario may be stopped (frozen) or it may continue. As with all SA measurement methods, there are advantages and disadvantages when using explicit performance measures. The main advantage is that they are direct measures; they do not rely on inference or opinion to determine an operator's SA. They have also been used in a variety of domains and task scenarios and have shown a high degree of reliability and validity. The main disadvantages of techniques like SAGAT include the intrusion of stopping the task and also the necessary control over the task environment in order to design a task specific assessment.

Physiological measures of SA include measures of heart rate, brain activity, and eye movements. Physiological measurements are similar to implicit performance measures in that it is difficult to parse out SA from the other constructs that are likely affecting performance. Electroencephalographic (EEG) measurements of brain activity may be able to show within a very precise time window if information is being attended to (e.g., Mecklinger, Kramer, & Strayer, 1992); but these measures cannot identify the location or identity of the attended information. Eye-tracking devices can determine where a participant is looking, which is often where the participant is attending and perceiving. The use of eye tracking to measure SA has not been extensively studied. Several studies have shown support for a look-but-not-see phenomenon of attention; where someone may fixate on an object but cannot recall information about it (Salmon et al., 2006; Strayer, Cooper, & Drews, 2004). More research is needed using physiological measures of SA in order to determine their potential benefits.

One problem with current SA measurement techniques is the inability to measure SA in an uncontrolled environment (e.g., during training or on-the-job operations). Although subjective measures suffer from the previously mentioned disadvantages, they are currently the best option when field testing with no pre-determined scenarios. The SAGAT technique is the most common direct measurement method, but it requires a priori knowledge and the capability to freeze the test situation in order to measure SA. Thus, it cannot be used in an uncontrolled environment.

Goals of the current study

One previously unexplored option, which will be examined in the current study, is to use physiological measures such as eye tracking in combination with a direct measurement method in a controlled test situation to determine how eye movements and performance on SA measures correlate. If studies like this reveal patterns of eye movements that predict operator SA, then it may be possible to use eye tracking in uncontrolled test environments to measure operator SA. Even if the results do not support the measurement of SA using only eye tracking data in uncontrolled environments, the current research will help to further develop measurement methods of the processes of SA. The current studies will assess operators' SA in dynamic scenarios, including driving and ATC, using both direct query measures of SA and eye tracking measures. The eye tracking data and the SA query data will then be compared to see what patterns of eye movements predict whether an operator maintains accurate or inaccurate SA. Given this focus, the next sections will review how query measures of SA and eye tracking methods have been used for ATC and other tasks.

Use of direct query measures for SA measurement

SAGAT is the most commonly used and validated measure of SA available. It is domain specific and requires extensive preparation for each domain to develop detailed queries (Endsley, 2000). Task analyses are typically used to develop queries for the SAGAT. Task analyses are developed by determining the major goals of the user. Once those are established, the major subgoals required to meet those goals are identified (Endsley, 2000). Researchers agree that the focus of the queries should not be on static information in the environment, but rather on the dynamic situations (Endsley, 2000, Endsley & Rodgers, 1994; Wickens, 2008). Due to the domain-specific, complex nature of goal-directed task analyses (GDTAs), they may take up to a person-year to complete (Endsley, 2000).

SAGAT queries are typically administered several times throughout a task. For each administration, the scenario is frozen and the screen is blanked. In some applications, scoring for each query is binary (correct or incorrect) and based on what was happening when the scenario was blanked. Pre-specified tolerances levels are included for queries when necessary (e.g., altitude within 1000 ft.). Although the development of queries is time intensive, SAGAT remains the most commonly used direct measure of SA. In an experiment assessing SA in fighter pilots, SAGAT and SART were administered to measure SA and the NASA TLX was administered to measure workload. SAGAT was the only measure that correlated with performance (Endsley, Selcon, Hardiman, & Croft, 1998).

The SPAM method is a direct measurement method that is similar to SAGAT except the screen is not blanked during the scenario freezes. Durso, Bleckley, and Dattel (2006) tested participants on ATC performance to determine if SA measurements add to the predictive ability of a battery of tests of personality and cognitive ability tests (e.g., working memory span; Big Five). SA was measured using SAGAT and SPAM. Handoff delay times were predicted only by general fluid ability in the base model of cognitive tests. After fitting general cognitive abilities in a base model, SPAM increased the ability to predict variance in handoff delay times by 9% and ATC errors by 15%, whereas SAGAT did not account for any additional variance in these variables. However, SAGAT predicted en route times better than SPAM. SPAM and SAGAT had high convergent validity (Cronbach's alpha = .623), which is interesting considering they differed in the kinds of ATC performance variance that each accounted for. Durso et al. conclude that SPAM does a better job than SAGAT in explaining variance in ATC performance.

Query techniques (e.g., SAGAT and SPAM) are ideal when the task is understood a priori and the experimenter is able to interrupt or completely stop the scenario. In situations where the experimenter does not have experimental control, e.g., on-the-job operations, a different technique would be better suited (Salmon et al., 2008a). Physiological measurements have been discussed for measurement in less structured environments, but few empirical studies have been done. As the affordability and portability of devices required to gather physiological measures has improved, more and more researchers have begun to incorporate these measures into their data collection.

One promising physiological measure that has been used in several SA experiments is measuring eye movements.

Eye tracking research

The use of eye tracking as a measure in psychological research has become more common in recent decades. Some researchers argue against using physiological measures due to their intrusive nature, but with new technological advancements eye trackers are now able to capture data passively, without any disturbance to the participant. A majority of eye trackers gather similar raw data, but differences arise in how researchers operationally define eye movement variables. Eye trackers collect thousands of data points, researchers must determine which data are applicable to their research interests and decide how to separate relevant from irrelevant points. Jacob and Karn (2003) report four typical eye tracking metrics, along with their definitions:

1. *Fixation*: A relatively stable eye-in-head position within some threshold of dispersion (typically ~2°) over some minimum duration (typically 100–200 ms), and with a velocity below some threshold (typically 15–100 degrees per second).

2. *Gaze Duration*: cumulative duration and average spatial location of a series of consecutive fixations within an area of interest. Gaze duration typically includes several fixations and may include the relatively small amount of time for the short saccades between these fixations. A fixation occurring outside the area of interest marks the end of the gaze.

3. *Area of interest (AOI)*: Area of a display or visual environment that is of interest to the research or design team and thus defined by them (not by the participant).

4. *Scan path*: Spatial arrangement of a sequence of fixations (pg. 583-584).

While AOI is not by itself an eye tracking metric, AOIs are used to calculate the other metrics.

Jacob and Karn (2003) reviewed 24 usability studies using eye trackers and recorded the eye-tracking metrics used in each study. The most commonly used metrics are discussed in Table 1.2, along with their relationship to performance in the studies.

Table 1.2
Common eye tracking metrics and their relationship to performance (adapted from Jacob & Karn, 2003, pg. 584-584)

Metric	# times cited	Relationship to performance
Number of fixations, overall	11	The number of fixations overall is thought to be negatively correlated with search efficiency (Goldberg & Kotval, 1998; Kotval & Goldberg, 1998). A larger number of fixations indicates less efficient search possibly resulting from a poor arrangement of display elements. The experimenter should consider the relationship of the number of fixations to task time (i.e., longer tasks will usually require more fixations).
Gaze % (proportion of time) on each area of interest	7	The proportion of time looking at a particular display element (of interest to the design team) could reflect the importance of that element. Researchers using this metric should be careful to note that it confounds frequency of gazing on a display element with the duration of those gazes. According to Fitts *et al*. (1950) these should be treated as separate metrics, with duration reflecting difficulty of information extraction and frequency reflecting the importance of that area of the display.
Fixation duration mean, overall	6	Longer fixations (and perhaps even more so, longer gazes) are generally believed to be an indication of a participant's difficulty extracting information from a display (Fitts *et al*, 1950; Goldberg & Kotval, 1998).
Number of fixations on each area of interest	6	This metric is closely related to gaze rate, which is used to study the number of fixations across tasks of differing overall duration. The number of fixations on a particular display element (of interest to the design team) should reflect the importance of that element. More important display elements will be fixated more frequently (Fitts *et al*, 1950).
Gaze duration mean, on each area of interest	5	This is one of the original metrics in Fitts *et al.* (1950). They predicted that gazes on a specific display element would be longer if the participant experiences difficulty extracting or interpreting information from that display element.
Fixation rate overall (fixations/s)	5	This metric is closely related to fixation duration. Since the time between fixations (typically short duration saccadic eye movements) is relatively small compared with the time spent fixating, fixation rate should be approximately the inverse of the mean fixation duration.

While usability studies have different goals compared with experimental studies, the eye movement metrics are similar across types of studies. While all of the above metrics are applicable to the current study, three seem especially relevant: the percentage of time gazing at (or fixating on) an AOI, the number of fixations on an extended event or a whole scenario, and the mean fixation duration during an extended event or a whole scenario. Reviews of studies using these eye movement variables are presented in the next section.

Eye tracking studies

Hauland (2002, 2008) examined process-oriented measures of SA using ATC students' eye movement data collected during simulator training. Measurement probes were used to measure SA; the probes were essentially implicit performance measures, imbedded into the experimental (termed 'abnormal') scenarios. Similar to the current experiment, Hauland was interested in developing novel SA measurement methods using eye tracking. The measurement probes were considered to be process-oriented measures, which differ from direct performance measures (e.g., SAGAT) which more likely measure the product of SA at various stopping points throughout a scenario. In the experiment, Hauland (2008) did not explicitly compare eye movements and established measures of SA; instead he hypothesized that the visual attention strategies of controllers would capture aspects of SA; and, in turn, the SA measures would predict performance. He hypothesized that SA measures would be validated if they predicted ATC performance and if they varied based on differences in the traffic situations (manipulated using implicit performance measures).

Teams of two controllers, one radar and one planner, participated in the experiment. Radar controllers handle the current traffic in their own sector, whereas planner controllers handle the monitoring of adjacent sectors and potential conflicts, as well as the flightstrips. Although specific tasks are allocated to each controller, both are involved in all monitoring tasks. The dependent measure was dwell time on AOIs; which was defined as fixations lasting longer than 250 ms within an AOI. AOIs were static areas of the radarscope, and several objects, such as aircraft, could be in an AOI at the same time.

Within the AOIs, focused and distributed attention were analyzed. When participants' fixations were longer and they fixated on only one or two objects within the AOI it was considered to be a focused attention strategy – defined as fixating for at least one second on one or two objects within an AOI. When participants' fixations were shorter and they moved around within an AOI it was considered to be a distributed attention strategy – defined as fixating for no more than 1 second on at least three objects within an AOI. Focused and distributed attention should not be confused with focal and ambient visual channels. Focused and distributed attention both involve fixating on objects and are both components of focal vision; whereas ambient vision involves the periphery of one's visual scene and is thought to be a separate visual system than focal vision.

Positive correlations were found between several ATC performance measures, including ratings of response time and sufficiency of radio transmissions, and a distributed attention strategy for the planner controller position. The radar controller used the distributed attention strategy longer in the control (termed 'normal') scenarios, compared with the abnormal scenarios which included the implicit performance measures. Conversely, the planner controller's use of the focused attention strategy was shorter in the normal scenarios compared with the abnormal ones. Positive correlations between only one ATC performance measure, the logged number of radio transmissions, and a focused attention strategy were found for the radar controller position. Thus, radar controllers had better ATC performance with the distributed attention strategy, which they used more often in normal scenarios (where SA was thought to be higher); whereas planner controllers had less success with the focused attention strategy, which they used more often in the abnormal scenarios (where SA was thought to be lower) (Hauland, 2008). Results validated the idea that process-oriented measures such as eye movements can be used to measure SA. Similar to Hauland (2008), amount of time spent fixating on areas of interest will be used as an independent measure in the current studies.

The Attention-Situation Awareness (A-SA) model discussed earlier used eye movement data to predict performance in several flight simulations (Wickens et al., 2005).

> The underlying theoretical structure of the A-SA model is contained in two modules, one governing the allocation of attention to events and channels in the environment, and the second drawing an inference or understanding of the current and future state of the aircraft within that environment. The first module corresponds roughly to Endsley's (1995c) Stage 1 situation awareness, the second corresponds to her Stages 2 and 3 (Wickens et al., 2005, pg. 2).

The A-SA model is based on the SEEV model. The researchers were interested in predicting dwell time percentages in AOIs using the model. Bandwidth, which involves the amount and frequency of new information provided by a channel, and relevance were as good or better predictors of scanning behavior and performance by themselves compared to when Effort was included in the model. They found that Effort, i.e., length of scans, was not a predictor of scanning behavior. In addition, pilots who closely matched the optimal expected value model of scanning predicted by the experimenters had better performance for both flight path tracking and detecting traffic (Wickens et al., 2005).

In a similar set of experiments validating the SEEV model, researchers found that percent dwell time for the environment outside of a car in a driving simulator decreased as bandwidth for an in-vehicle display increased (Horrey, Wickens & Consalas, 2006). Bandwidth of the in-vehicle display was increased by increasing the speed of a number task presentation. As bandwidth increases, the environment changes at a faster pace; when bandwidth becomes too high, the amount of processing required to keep up will be too much on the operator and SA will suffer (Durso et al., 2007). Horrey, Wickens, and Consalas (2006) used the following variation of the SEEV formula to predict the likelihood of scanning on particular AOIs:

$$P(AOI_j) = \sum_{t=1}^{n} [(B_t)(R_t)(P_t) - Ef_t]$$

"where t = task, B = information bandwidth, R = relevance, P = priority, and Ef = effort associated with accessing the AOI. In this formula, Expectancy is expressed as information bandwidth and Value is expressed as the product of Relevance and Priority" (pg. 75).

The model fit for the Expectancy (bandwidth) and Value (relevance of priority) parameters. Effort was null in the model because its influence could not be predicted based on the experimental design. The predicted values were determined a priori using well-specified rules and rank ordered values of the task conditions and AOIs assigned based on a lowest ordinal algorithm (Horrey, Wickens, & Consalas, 2006). When comparing their actual experimental results to the predicted values, they found a high correlation between the predicted and actual values of percent dwell time ($r = .98$). When only Expectancy was included in the model, 63% of the variance in percent dwell time was accounted for, compared to 74% of the variance when only Value was included in the model. These findings suggest that two of the key environmental parameters people use to allocate attention in dynamic scenarios are frequency of information change and task value. They also suggest that when a task has a high value, drivers will examine AOIs relevant to this task more frequently at the cost of other AOIs (Horrey, Wickens, & Consalas, 2006). In the A-SA and SEEV validation experiments, the percent dwell time in AOIs was again a relevant measure of SA.

The above studies are relevant to the eye movement variable percent of time fixating on an important event. Other studies have looked at mean fixation durations as a measure of performance. Chapman and Underwood (1998) recorded the eye movements of novice and experienced drivers while they watched films of realistic driving scenarios that sometimes contained hazardous events. They found that the duration of novices' fixations were longer than that of experts. Overall, participants had longer mean fixation durations when fixating on hazards compared with the rest of the scene. Recarte and Nunes (2000) measured several eye movement variables describing how drivers scanned the road as they drove on highways and concurrently performed verbal and spatial-imagery tasks that did not require visual perception. The spatial-imagery task produced longer mean fixation durations on road objects than both the verbal task and no task. In these two studies, as well as the studies reviewed earlier by Jacob and Karn (2003), longer fixation durations are associated with difficulty processing stimuli due to lack of task expertise or attentional overload.

Another eye movement variable examined in the current study was the number of fixations on an event or a whole scenario. The studies reviewed by Jacob and Karn (2003) suggest that an excessive number of fixations is associated with difficulty in gathering information about a scenario. Rahimi, Briggs, and Thom (1990) found that in real driving tasks participants had a greater number of fixations at busy intersections compared with quiet ones. A review of previous research concluded that typically as task demand and/or visual complexity increases, fixation rate increases (Crundall, Underwood, & Chapman, 1998). The studies just reviewed suggest that effective tracking of a dynamic scenario (as done by expert or non-overloaded operators) will be associated with more time fixating on the event as well as fewer and shorter fixations than in the case of ineffective tracking.

Several authors argue that eye tracking data should not be used to measure SA because it is not possible to determine what it is actually measuring (Salmon et al., 2008a). Cooke, Stout, and Salas (2001) note that eye movement data may not perfectly correlate with an individual's thoughts, but the information afforded by it can still be beneficial to researchers. Eye tracking data can be compared to verbal reports to determine which cues were attended to that operators stated were important. They explain how information not attended to by participants is stronger evidence of a lack of visual attention than cues that are attended to in the environment (Cooke, Stout, & Salas, 2001). Their reasoning for placing little emphasis on cues actually attended to is likely based on a look-but-don't-see phenomenon, where an individual may fixate on an object in his or her visual array but does not perceive or recognize it (Salmon et al., 2006; Strayer, Drews, & Johnston, 2003).

On the other hand, researchers have argued in favor of using eye tracking information to infer visual attention by claiming that where a person is looking is an indicator of what they are attending to, also known as the "eye-mind hypothesis" (Guan, Lee, Cuddihy, & Ramey, 2006; Williams et al., 2005). Even though people may not always perceive and recognize what they are looking at, it can be assumed that they are able to do so for the majority of the time. This is even more likely in situations where an operator is required to pay attention to information in order to adequately perform a task. Thus, while it cannot be concluded that all eye movements equate to operator perception, they must at least perceive a majority of the information they are looking at while performing a task in order to achieve even a minimal level of performance. In addition, many of the eye tracking studies reviewed above (Chapman & Underwood, 1998; Horrey, Wickens & Consalus, 2006; Recarte & Nunes, 2000; Hauland, 2008; Wickens et al., 2005) provide evidence that eye movements are responsive to changes in information in dynamic tasks, and show regular and plausible associations with changes in operator expertise, workload and task performance.

Rationale for the current research

To date, no experiment has compared eye movement data with validated, explicit measures of SA to explore the relationship between eye movements and SA and to determine if eye movements predict SA. SA and its measurement have been the focus of a number of human factors researchers and practitioners for over 20 years. Although there are established measures for various situations and scenarios, there are no direct measures of SA for realistic test environments where the researcher has little or no control over the information being presented to the participant. The inability to directly measure SA in uncontrolled environments has become clear to the author in the development of a test plan for the U.S. Navy for Maritime Domain Awareness (MDA) effectiveness. The MDA suite of tools was developed for the U.S. military and implemented with the goal of increasing operator SA and threat awareness of water crafts throughout the world. Measuring the effectiveness of MDA has proven to be difficult because tests must occur during a regular work shift, without interrupting operators from their work. Researchers are currently not able to develop a simulation of the MDA suite of tools and conduct testing outside of the actual work environment. Instead, they gather information about operator preference and performance by observation during operator shifts and questionnaires after the shifts are over. Even though SA is explicitly stated as one of the goals of MDA, it is currently not able to be effectively measured.

As discussed previously, the direct measures of SA (including the SAGAT and SPAM techniques) can only be used when the researcher has control over the testing environment and scenarios to be carried out by the operators. Subjective measures of SA, in the form of a questionnaire given to the participant or observations of SMEs during the testing, are not ideal because participants have difficulty assessing their own SA and because it is an internal construct. Consequently, outside observers cannot adequately determine the level of SA of another individual.

Eye tracking measures have been discussed as a viable SA measure, but have not been validated for this purpose. Hauland (2008) began to explore the relationship of eye tracking and SA, but did not compare eye movements to established measures of SA. It can be argued that eye tracking data is not a direct measure of SA and can only indicate where a person is looking, which has several drawbacks. One objection is that although a person may be looking at a particular portion of a monitor, it cannot be assumed that the person perceives and comprehends the information that they are seeing. This objection will be addressed in the current research by correlating patterns of eye movements with direct SA measures that depend on perception and comprehension.

A second objection is that eye movement data can only describe a portion of the processes occurring when measuring SA. Of the three levels of SA, some argue that it is likely only effective at describing Level 1 Perception, and that, because important aspects of comprehension and projection may take place cognitively, these processes may be less likely to be captured using overt eye movement measures. Study 1 and Study 2 of this research project address this objection by assessing correlations between eye movements and measures of SA that require comprehension and projection (planning).

Rationale for Study 1

The first part of the current project involved further analysis of data from a previous experiment conducted by the author and others (Balk, Moore, Steele, & Spearman, 2006). In this experiment, participants' eye movements were captured as they performed a simulated driving task that required SA. Participants watched 30 second scenarios containing a hazardous event in a low-fidelity driving simulator. At the end of each scenario, participants answered a question related to an event that occurred during the scenario (e.g., tailgating) using an SA measure similar to SAGAT. Originally, the focus of the experiment was on driver distraction while talking on a mobile phone. Participants were placed in one of two groups, distraction present (simulated mobile phone condition) or distraction absent (control group).

Since the experimental design included collection of both direct SA measures and eye movement data, I re-analyzed the data in light of the goals of the current project. The distraction and control groups were compared in terms of participant response to the SA queries and eye movement data. No prior studies have compared eye movement data to direct query measures of SA, so the analysis was exploratory in nature. Three eye movement variables were analyzed: percent time fixating on an event, number of fixations during an event, and mean fixation duration during an event. Analysis of the first study helped to guide the development of the second study's design and analysis.

Previous research led to several hypotheses. First, it was expected that SA accuracy for an event would improve as percent time fixating on events increased. In other words, participants who fixated on an event more often would have higher SA compared to those who did not fixate on the event as often (see Horrey, Wickens, & Consalas, 2006). Second, it was expected that increased SA accuracy for an event would be associated with fewer fixations and shorter fixation durations during the event. These hypotheses are supported by the studies of Jacob and Karn (2003), Chapman and Underwood (1998), and Recarte and Nunes (2000) discussed earlier, which suggested that more effective event processing is associated with fewer and shorter fixations.

CHAPTER TWO

STUDY 1

The first study's data set was from an experiment that measured the differences in driving performance and eye movements for drivers who performed a concurrent mobile phone task and those who did not (Balk et al., 2006). Eye movement data were collected while participants viewed a low-fidelity driving simulator on a desktop monitor. Data were collected using a Tobii eye tracker. Following each scenario, participants answered an SA question regarding the scenario events and gave a confidence rating of the accuracy of their responses. The results are a first step to determine if eye movements predict performance on SA measures.

Method

Participants

Sixteen Clemson University undergraduate students participated in the experiment. All participants had valid drivers' licenses and a minimum of two years driving experience (M = 3.5 years). One participant was removed from analysis because the eye-tracker was miscalibrated and the eye movement data were not accurate. The experimental session lasted about 25 minutes. Participants received course credit for their participation.

Apparatus

Eye movement data were collected using a non-invasive Tobii 1750 eye tracker, sampled at 50 hz with a latency between 25 and 35 ms (see Figure 2.1). The Tobii was chosen in part for its accuracy throughout longer experiments lasting more than 5 minutes. The average accuracy of the Tobii across a number of participants is 0.5 degrees, where one degree of accuracy equals an average error of one centimeter between the measured and the intended gaze (Tobii Eye Tracker, 2006).

Figure 2.1. The Tobii 1750 Eye Tracker.

The Tobii 1750 collects eye movement data passively; users sit in front of the monitor and data are captured. The Tobii hardware uses binocular eye tracking and all calculations are done automatically by the system. The camera field of view (FOV) is approximately 20 x 15 x 20 cm at a viewing distance of 60 cm from the computer screen. According to the user manual, this is large enough for any comfortable head position while sitting with normal posture in front of the monitor. This is in part because only one eye needs to be in the FOV, which increases the tolerance to 30 x 15 x 20 cm (Tobii Eye Tracker, 2006). Participants viewed the scenarios on a 17 in LCD computer monitor (1280 x 1040 screen). The low-fidelity driving simulator was developed using C++, OpenGL, and SDL.

Design

The experiment employed a 2 x 2 (phone condition x number of cars) mixed model design, with eight participants (3 males) in the no-phone (control) condition and eight participants (2 males) in the mobile phone (distraction) condition. All participants viewed 24 scenarios, 12 containing four cars other than the driver's car and 12 containing seven cars other than the driver's car.

Materials and Tasks

The driving simulator was designed so the viewpoint was from the cockpit of a car driving on a three lane road. The screen was sectioned into a car's windshield view, rearview mirror, and left and right mirrors (See Figure 2.2). The simulator and scenario designs were based on previous research by Gugerty (1997).

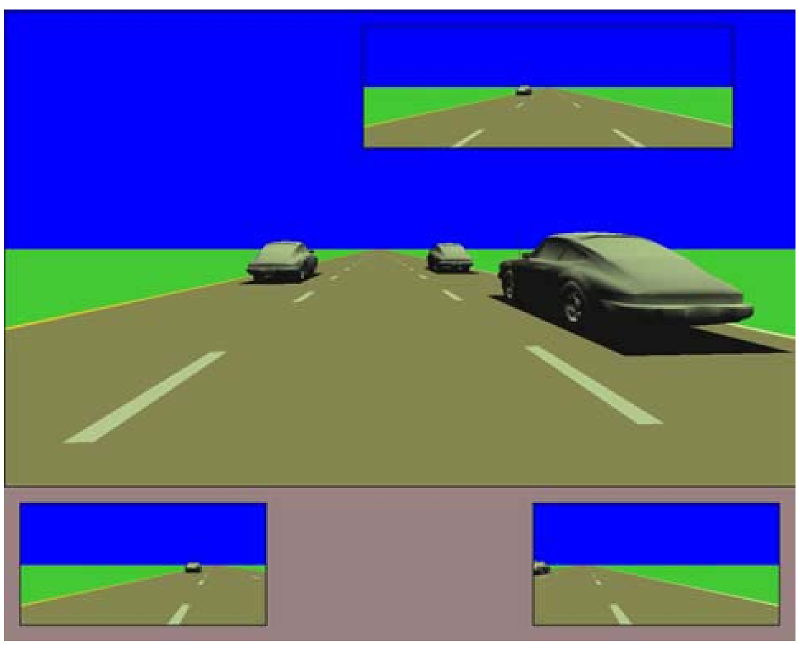


Figure 2.2. Screenshot of the monitor for a scenario containing four cars.

In 16 scenarios (eight each in the four and seven car conditions), a potentially hazardous event occurred during the trial. Potentially hazardous events consisted of a car changing lanes several times, a car changing speeds several times, two cars on a collision course, a car driving fast the entire time, a car driving slow the entire time, a car about to pass another car, a car tailgating behind another car, and a car weaving in and out of its lane. For the current study, scenarios will be described by the number of cars and the type of event; e.g., the four car scenario involving two cars on a collision course will be labeled as Collision Course (4) (See Table 2.1 for the names of events).

Table 2.1
Name and description of each event

Name	Description
Change Lanes	Car changing lanes several times
Change Speeds	Car changing speeds several times
Collision Course	Two cars on a collision course
Fast	A car driving faster than the other cars throughout the scenario
Slow	A car driving slower than the other cars throughout the scenario
Pass	A car about to pass another car at the end of the scenario
Tailgate	A car tailgating behind another car for a portion of the scenario
Weave	A car weaving in and out of its lane

All scenarios lasted 30 seconds. Following a scenario, a question about the event in the scenario was presented and the participant answered a multiple choice question about the event. Every question used a map view, showing the driver's car and each traffic car's ending position on the screen, and asked the participant to identify which car was involved in the event (See Figure 2.3). For the eight scenarios with no hazardous events, a question was asked about an event that did not actually occur. Participants were made aware that non-event scenarios would be included in the experiment so that they would not assume an event would occur in every scenario. All trials had a ‘no car’ option, which was the correct choice for the non-event scenarios. Confidence ratings were also recorded after each response to the SA questions. Ratings were based on a five point Likert scale where 1 indicated ‘not at all confident’ and 5 indicated ‘very confident.’

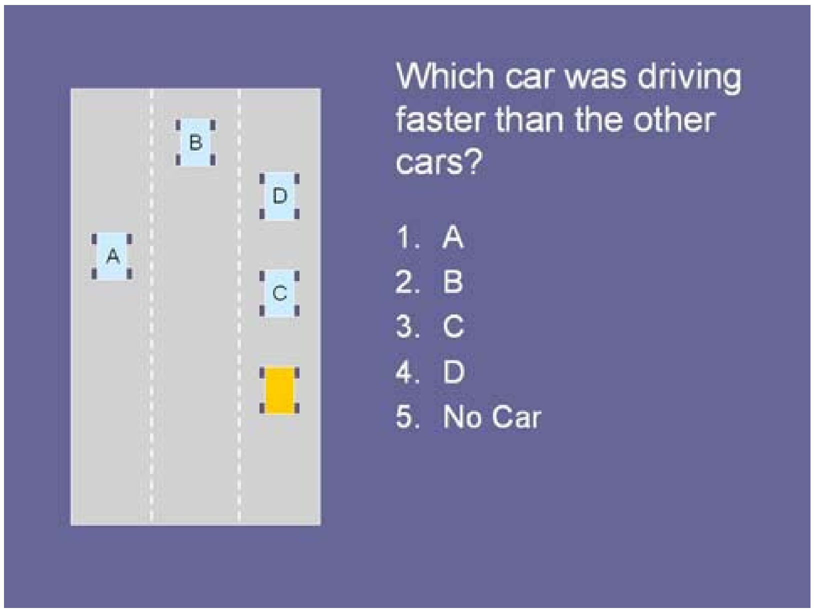


Figure 2.3. Screenshot of a question and response map presented after the completion of a scenario (Orange car is driver’s, blue cars are traffic cars).

The simulated mobile phone (distraction) task consisted of a foreign language learning compact disc synced to the start and stop times of each trial. The language learning task was selected because it was automatically paced and required participants to listen to the speaker, repeat phrases, and respond to questions throughout each scenario. Comprehension of the language learning task was tested at the end of the experiment by asking participants questions related to the information presented to them throughout the scenarios.

Procedure

Participants read and signed an informed consent form before beginning the experiment. All participants viewed several practice trials in order to familiarize themselves with the experimental design. Before the start of the actual trials, participants' eye movements were calibrated with the eye tracker. They were recalibrated halfway through the trials. Each participant completed the 24 trials in a different random order, although participants in the phone condition received the audio portion in a sequential order. Upon completion of the scenarios, participants in the mobile phone condition answered several questions measuring their comprehension of the language learning portion. After the completion of the trials, all participants completed a short questionnaire about their mobile phone habits, usage, and attitudes.

Results and Discussion

Data Collection and Preliminary Analysis

The eye tracker collected eye movement data throughout each of the 24 scenarios, though only the 16 with hazardous events were analyzed. Every 20 milliseconds the x and y coordinates for both the left and right eye were recorded, along with a validity code indicating the quality of the data. Areas of interest (AOIs) were established around the cars in each scenario that were involved in each event. Not all events occurred in front of the driver; in some cases, AOIs were defined for cars in the side and/or rearview mirrors. The raw data were filtered and invalid data points were removed from the data set. Once the points were removed, another filter determined which data points were saccades and which were fixations based on the locations of the two points recorded before and after each point.

Several participants' eye movement data sets contained a large proportion of invalid data points on certain scenarios. Participants' data for a specific scenario were removed from analysis if 33.3% or more of their eye movement data for that scenario were missing or invalid. This occurred in eight instances, all in the phone condition. Because analysis of the data occurred at the scenario level, participants who did not have enough eye movement data for a particular scenario were still included in other scenarios where they had an acceptable number of data points. One participant was removed completely from analysis because upon review of the participant's gaze replays for individual scenarios, it appeared the eye tracker was miscalibrated and the individual data points did not correspond to where the participant was actually looking.

For the remaining participant data sets, another preliminary data analysis issue concerned temporal gaps in the data created by invalid data points. Small gaps (of 40 ms or less) were ignored, but larger time gaps (greater than 40 ms) were flagged. The two data points prior to each large gap and the two data points immediately following the gap were marked and excluded from analysis. This was done because fixations were determined by comparing each data point to the two time intervals before and after it; and these five consecutive data points were not available when large time gaps existed. If a large time gap occurred, it was impossible to determine if the participants remained looking at the fixation point or if they had possibly looked away and then back within the missing time period. After removing data points around large gaps, fixations were defined as consecutive data points with velocities less than 130 degrees visual angle per second for at least 150 ms.

When calculating number of fixations and fixation duration, time gaps were ignored, likely leading to some overestimations of fixation duration and underestimation of number of fixations. Both of these variables were deemed to be relevant, even though they may not be a completely accurate picture of the actual fixations exhibited by participants. The other option would have been to stop every fixation before the start of a time gap and to restart it after the gap. If this were done, error would have been introduced in the other direction, with more fixations and shorter durations, which is also inaccurate. It was thought that erring on the side of longer fixations was likely more accurate because fixations occurred nearly 90% of the time throughout the event for each scenario. When a fixation continued through a time gap, it indicated that participants were looking at the same place before and after the gap took place. It is likely that in most instances, participants' eyes remained in that position rather than moving away and returning before the gap ended. Although this method of calculating fixations may have lead to some error, this error was not expected to differ across the different conditions in the data analysis. Therefore, number of fixations and fixation duration were considered useful variables to test our hypotheses.

Data Analysis Plan

For the purposes of the current research project, analysis of the Study 1 data focused on eye movements only during the time when a potentially hazardous event occurred within a scenario. Event durations ranged from 3 to 30 s. Level of phone use was used as a between-subjects independent variable in the analysis. Number of cars present in the scenario (4 or 7) was used as a within-subjects independent variable. Thus, the analysis focused on how eye movement variables differed depending on level of phone use and number of cars present. The binary dependent variable was SA accuracy on a single scenario (Incorrect = 0, Correct = 1); it was determined by participant response to the multiple choice question at the end of the scenario.

Three eye movement variables were investigated: percent time fixating on an event, number of fixations during an event, and mean fixation duration during an event. Percent time fixating on an event was calculated by dividing the number of data points labeled as fixations in the AOI during each event by the total number of data points during that event. Number of fixations during an event was determined from the raw data by isolating groups of data points during an event labeled as fixations between two data points labeled as saccades. Each group was labeled as an individual fixation and the duration of the group was also calculated. Mean fixation duration during an event was determined by dividing the total time spent fixating during an event by the total number of fixations during that event. Fixations *anywhere* on the driving scene, not just on the AOI defining the event, were included in the calculation of the "number of fixations during an event" and "mean fixation duration during an event" variables. Thus, these two variables assessed how participants allocated attention across the whole scene during a critical driving event. These three eye movement variables were considered as predictor variables; the analysis evaluated the eye movement variables' relationship to the dependent variable of SA accuracy.

In all scenarios besides Pass (4), participants in the control (no phone) condition were correct more often than participants in the distraction (phone) condition. Participants in the no phone condition were accurate 70% of the time, while participants in the phone condition were only accurate 36% of the time. Recarte and Nunes (2003) found that increased mental workload from auditory listening and verbal production tasks affected participants' abilities to detect visual stimuli in real driving conditions. The current result is similar to an almost 30% reduction in detection capabilities found on an actual driving tasks due to endogenous distraction from mental tasks. They explain that their result is "practically meaningful as an estimate of the increased risk of distraction errors hypothetically leading to traffic conflicts or accidents" (pg. 130).

Generalized estimated equations analysis

The actual data analysis focused on answering three main questions regarding eye movements and SA accuracy. First, determine what the overall effect of phone use and traffic level was on SA accuracy to provide context for interpreting the eye movement results. Second, determine how different eye movement variables affect SA accuracy (i.e., main effects). Third, determine how eye movement variables affect SA accuracy based on phone use and traffic level (i.e., interactions). To answer these questions, several generalized estimated equations (GEE) analyses were performed on the data set. One important function of examining Study 1 data was to develop a proper technique for analyzing the next study's data set. Due to the complex nature of the experimental design, common statistical analyses were not appropriate for the current study. The mixed model design with within-subjects variables and a binary dependent variable indicated that the appropriate analysis would be a GEE, a type of logistic regression. Logistic regressions are typically a better choice of analysis compared with multiple regressions or discriminant analysis because they do not require the data to be normally distributed, to be related linearly, or to have groups with equal variances. In addition, binary dependent variables are accepted (Mertler & Vannatta, 2005).

The first GEE examined only the effects of phone use (phone vs. no phone) and traffic level (4 vs. 7 cars) on SA accuracy. In order to build the model, the specified between subjects effect was participant and the within subjects effects were number of cars and scenario; the specified effects were separate from the IVs. For all GEEs in this analysis, the correlation matrix structure was exchangeable and the probability distribution was normal. The mean percentage correct for the SA queries broken down by phone group and traffic level are presented in Table 2.2. These data suggest that SA was more accurate in the no-phone and the low-traffic conditions.

Table 2.2
Percent correct on SA queries (with standard error in parentheses)

	Phone Use		
Traffic Level	*No phone*	*Phone*	*Mean*
Low (4)	77 (3)	41 (9)	61 (6)
High (7)	64 (6)	31 (6)	47 (5)
Mean	69 (2)	39 (8)	55 (3)

There was a significant main effect for phone group (p = .00, Wald = 31.33, β = 1.47) and the main effect for number of cars approached significance (p = .08, Wald =

3.17, β = 0.53). The group x number of cars interaction was not significant. The means for SA accuracy by group and traffic level are graphed in Figure 2.4.

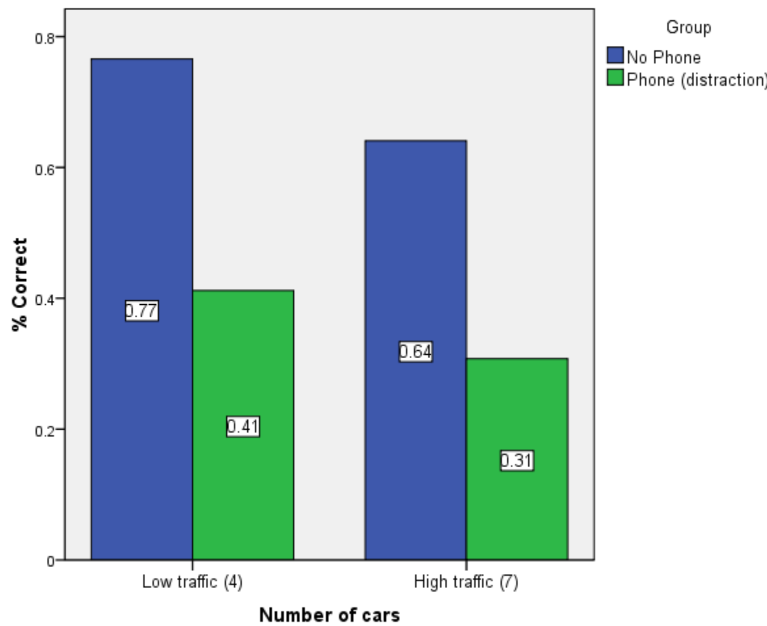


Figure 2.4. The percent correct for SA accuracy by group and number of cars.

Participants in the no phone condition answered SA questions accurately more often than participants in the phone condition. This trend occurred for both low traffic and high traffic scenarios. Participants in both conditions also tended to answer more SA questions correctly in low traffic scenarios than in high traffic scenarios, but this result was not significant.

The second GEE examined the main effects of the three eye movement variables (percent time fixating on events, number of fixations during the event, and mean fixation duration during the event) along with the main effects of group and traffic level on SA accuracy. This analysis focused on whether the eye movement variables had an overall effect on SA accuracy. A third set of GEEs looked at whether the effect of a single eye movement variable interacted with phone use and number of events. For example, for percent fixations on an event, this analysis included all the main effects of the second GEE plus the percent event fixations by phone use and the percent event fixations by traffic level interactions. Similar analyses were run for the other two eye movement variables. It should be noted that the mean fixation duration variable had large skew and kurtosis values. To fix this, a natural log transformation was performed on the variable and those values were used for the GEEs. Finally, a fourth set of GEEs looked at the three-way interaction between a single eye movement variable, phone use and traffic level. For example, for percent fixations on an event, this analysis included all the main effects and two-way interactions of the second and third GEEs plus the percent event fixations by phone use by traffic level interaction. Similar analyses were run for the other two eye movement variables.

When all of the predictor variables were included in the GEE model, only the main effects for group and % fixations in AOI were significant (Group: $p = .00$, Wald = 13.0, $\beta = 1.2$). Participants in the no phone condition were accurate 69% of the time (StdE = 2%) while participants in the phone condition were accurate 39% of the time (StdE = 8%). There were no significant main effects for number of cars ($p = .11$), number of fixations ($p = .62$), or mean fixation duration ($p = .99$). The significant main effect for % fixations in the AOI supported the hypothesis. It will be explored in more detail in the next section.

Percent time fixating on events

The hypothesis was that a higher percentage of time fixating on an event would be associated with more accurate SA for that event. Table 2.3 shows the overall effect of percent time fixating on events on SA accuracy, as well as how this effect varied with phone use and traffic level. The overall percent time fixating was 44.3% (*SE* = 2.2%) when SA was accurate and 27.5% (*SE* = 2.4%) when SA was inaccurate; and the main effect of percent time fixating showed that this difference was significant (p = .00, Wald = 10.1, β = 2.2). This supports the hypothesis that more time fixating an event would lead to better SA.

Table 2.3
Percent time fixating events for correct and incorrect SA responses (with standard error in parentheses)

		Phone Use					
		No phone		*Phone*		*Mean*	
Traffic Level	*SA accuracy*						
Low (4)	Correct	49*	(3)	29	(6)	30	(3)
	Incorrect	27	(5)	31	(4)	32	(3)
High (7)	Correct	50**	(4)	37**	(6)	46	(3)
	Incorrect	29	(6)	24	(4)	26	(4)
Mean	Correct	49**	(2)	32	(4)	44**	(2)
	Incorrect	28	(4)	27	(3)	27	(2)

Significance levels: p < .05, p < .01***

However, the data in Table 2.3 suggest that the benefit of fixating longer on an event was stronger in the no-phone condition (49% fixating for accurate SA; 28 % for inaccurate SA) than in the phone condition. This conclusion was supported by a significant interaction of percent time fixating and phone use (p = .00, Wald = 15.9, β = -0.615), and by a significant simple effect of percent time fixating on SA accuracy within the no-phone condition (p = .00, Wald = 35.04, β = 4.3), but not within the phone condition (p = .17). Thus, fixating longer on an event was only associated with better SA when participants were free from dual-task distraction.

Finally, the interaction of percent time fixating and traffic level (p = .69, Wald = .157, B = -0.615) and the three way interaction (p = .13, Wald = 2.3, β = 4.9) were not significant. The data were then split into four conditions based on phone use and traffic level (No Phone (4), No Phone (7), Phone (4) and Phone (7)) to further inspect the simple effects. Individual GEEs were run for each condition examining the relationship between the three eye movement variables and SA accuracy. When all three variables were included in the GEE model, percent time fixating on events was significant for all conditions except Phone (4).

To put these associations between percent time fixating on events and SA accuracy in context, it helps to first recall the main effect of phone use on SA accuracy, i.e., SA was significantly higher in the no-phone condition (69%) than in the phone condition (39%). Thus, for the no phone condition, where SA accuracy was high, the amount of time fixating on an event showed a strong positive association with SA accuracy. In contrast, for the phone condition, where SA accuracy was low, time fixating on an event was not significantly associated with SA accuracy. Thus the hypothesis that more time fixating on an event is associated with higher SA accuracy for that event was supported when SA accuracy was high, but not when SA accuracy was low. Interestingly, participants in the no-phone group who answered the SA query correctly fixated on the event for almost half of its duration. In contrast, participants in the phone group, who were distracted from the driving task, fixated less on the important information in the scene and, perhaps consequently, showed lower SA.

Number of fixations during an event

Number of fixations *during* an event assesses fixations anywhere on the driving scene during a critical driving event, not only fixations on the event itself. The hypothesis was that fewer fixations during an event would be associated with more accurate SA for the event. As Table 2.4 shows, there was no main effect of number of fixations on SA accuracy (p = .25). However, the data in the table suggests that during low traffic, fewer fixations was associated with more accurate SA; while during high traffic, more fixations was associated with more accurate SA. This conclusion was supported by a significant interaction between traffic level and number of fixations (p = .00, Wald = 21.48, β = -0.122). Further analysis of the interaction using simple effects tests revealed that, in the four car scenarios, those who answered correctly had significantly fewer fixations than those who answered incorrectly (p = .03, Wald = 4.867, β = -.039); in contrast, in the seven car scenarios, those who answered correctly had significantly more fixations (p = .001, Wald = 11.61, β = .093). The interaction of number of fixations and phone use (p = .998) and the three way interaction (p = .441) were not significant.

Table 2.4

Number of fixations during an event for correct and incorrect SA responses (with standard error in parentheses)

		Phone Use					
		No phone		*Phone*		*Mean*	
Traffic Level	*SA accuracy*						
Low (4)	Correct	13.1	(1.7)	12.4**	(2.1)	12.9*	(1.3)
	Incorrect	21.9	(3.8)	19.5	(2.8)	20.3	(2.2)
High (7)	Correct	9.5 **	(1.2)	14.0*	(3.0)	10.7 **	(1.2)
	Incorrect	7.0	(0.9)	9.2	(1.0)	8.4	(0.7)
Mean	Correct	11.4	(1.1)	13.1	(1.7)	12.0	(0.9)
	Incorrect	12.9	(2.0)	13.9	(1.5)	13.5	(1.2)

Significance levels: p = .05, .01***

Individual GEEs run for each condition examining the relationship between SA accuracy and number of fixations revealed significant simple effects for number of fixations for all groups except No Phone (4). The means for number of fixations during an event are presented in Table 2.4. The associations between traffic level, number of fixations and SA were again seen in the four conditions. Compared to those who answered incorrectly, participants who answered correctly showed fewer fixations in one of the four-car conditions, and more fixations in both seven-car conditions.

To put these findings in context, recall that in the four car scenarios participants answered 61% of the SA questions accurately, compared with 47% in the seven car scenarios; and this effect approached significance (p = .12). Thus, for the low traffic condition (where SA accuracy was high) accurate SA for an event was associated with fewer fixations on the overall scene, while inaccurate SA was associated with more fixations. In contrast, for the high traffic condition (where SA accuracy was lower) accurate SA for an event was associated with more fixations than inaccurate SA. Thus the hypothesis of fewer scene fixations being associated with higher SA for events was supported when SA was high and not supported when SA was low. Though the effect was significant for the high traffic condition, it should be noted that both low SA and high SA mean number of fixations were lower than the high SA mean number of fixations in the low traffic condition.

One explanation for the inverse relationship between scene fixations and SA accuracy in the low traffic condition is that people with inaccurate SA had a high number of scene fixations because they had not noticed the critical event and were scanning the whole driving scene, whereas people with accurate SA had fewer fixations because they had noticed the event and were focusing their attention on it.

Mean fixation duration during an event

Recall that this variable assesses fixations anywhere on the driving scene during a critical driving event. The hypothesis was that shorter fixations during an event would be associated with greater SA for that event. The means for number of fixations during an event are presented in Table 2.5. Recall that the overall GEE with all predictors in the model revealed no main effect for mean fixation duration (p = .99). The GEE examining the two way interactions of mean fixation duration during an event with phone use (p = .90, Wald = 0.015, β = 0.66) and with traffic level (p = .14, Wald = 2.16, β = 0.66) revealed no significant interactions. The three way interaction was not significant (p = .64). Individual GEEs run for each group examining the relationship between SA accuracy and mean fixation duration revealed no significant effects. Thus the hypothesis regarding mean fixation duration was not supported.

Table 2.5
Mean fixation duration during an event for accurate and inaccurate SA (with standard error in parentheses)

		Phone Use		
		No phone	*Phone*	*Mean*
Traffic Level	*SA accuracy*			
Low (4)	Correct	1271 (194)	907 (113)	1162 (141)
	Incorrect	853 (137)	771 (72)	798 (65)
High (7)	Correct	1171 (152)	756 (80)	1055 (114)
	Incorrect	1348 (317)	787 (91)	1006 (141)
Mean	Correct	1225 (126)	842 (73)	1114 (93)
	Incorrect	1153 (201)	779 (59)	916 (84)

For the low traffic condition, there was non-significant trend whereby accurate SA for events was associated with longer fixations and inaccurate SA with shorter fixations. This was the opposite of our prediction, which was based on the assumption that long fixation durations indicated difficulty in processing information (Jacob & Karn, 2003). However, in dynamic scenarios, longer fixation durations may indicate better performance because those with high levels of SA might notice a hazardous event and focus in on it while those with lower SA levels might continue to scan their environment. A previous study supports this idea. Chapman and Underwood (1998) found that participants had longer fixation durations on hazardous events compared with the rest of the non-hazardous information in test scenarios.

This alternative explanation for long fixations also fits with the data for number of scene fixations in the low traffic condition, where high SA was significantly associated with few fixations, and low SA with more fixations. Taken together, these two variables suggest that in low traffic conditions, participants with high SA were making fewer but longer fixations, mostly on the critical event; while participants with low SA were making more but shorter fixations as they scanned the entire scene.

Study 1 General Discussion

Study 1 analyses revealed that both distraction level and number of cars affect SA performance. Not surprisingly, performance is better when participants aren't distracted from the primary task and when there are fewer objects to attend to in the environment. Some of the hypotheses about relationships between eye movements and SA were supported. A higher percent of time fixating on an event was associated with greater SA when participants were not overloaded with the phone task; but this association was weak or not present when participants were distracted in the phone condition. Fewer fixations were associated with greater SA when participants had only four cars to track; but the opposite was true for seven cars. Thus, a surprising finding of this study was that more effective eye movements were only associated with more effective SA when extrinsic or intrinsic workload was low.

One goal of Study 2 is to further understand how different patterns of eye movements lead to higher or lower SA when SA is measured by participants' responses to additional queries related to a dynamic scenario. Another goal of Study 2 is to investigate whether eye movement variables can predict current and future SA. This is in contrast to Study 1, which mainly focused on past events. This distinction will allow for more detailed analysis and a clearer understanding of the processes underlying the development of SA.

CHAPTER THREE

STUDY 2

The purpose of Study 2 was to determine if fixations and eye movement patterns predict performance on direct measures of SA. Study 1 revealed associations between eye movement variables and SA that were dependent on extrinsic, side-task and intrinsic, within-task workload. However, the task, though dynamic, did not require ongoing control input from participants. In addition, the scenarios lasted only 30 seconds, which likely affected the amount of SA that was obtained. Study 1 provided evidence that in several situations, amount of time spent fixating on an event and number of fixations during an event predict SA performance, but other eye movement measures not examined in the previous study may also play an important role. In addition, it is important to determine which aspects of the eye movement results from the first study are seen in a more dynamic, user-controlled task that requires users to maintain SA to successful complete it.

In Study 2, trained air traffic controllers completed three scenarios using a low fidelity Terminal Radar Approach Control (TRACON) simulator. TRACON controllers typically manage the airspace surrounding a major airport and several satellite airports. Their responsibilities include directing air traffic departing from the airports and accepting aircraft from adjacent sectors. Accepted aircraft will either be directed to an airport for arrival procedures or handed off to an adjacent sector once an appropriate altitude has been reached. Air traffic control (ATC) is a popular research area because SA is an integral part of a controller's job; they must build and maintain SA throughout their shift in order to have a high performance level and no errors. In testimony to the House of Representatives Subcommittee on Aviation, Committee on Transportation and Infrastructure (GAO-08-481T, 2008), it was stated that, "The primary causes of incursions, as cited by experts we surveyed and some airport officials, include human factors issues, such as miscommunication between air traffic controllers and pilots, a lack of situational awareness on the airfield by pilots, and performance and judgment errors by air traffic controllers and pilots" (pg. 8). Air traffic controller errors accounted for 28% of the incursions during FY07 (GAO-08-481T, 2008).

Researchers have often used ATC to study SA (e.g., Durso et al., 1998b; Durso, Bleckley, & Dattel, 2006; Endsley, 2000; Endsley & Jones, 1995; Hauland, 2008). As previously discussed, Endsley & Jones (1995) performed a GDTA to determine the requirements at each level of SA for TRACON controllers. SAGAT queries were developed based on the results of this analysis and those queries have been used in a variety of ATC SA measurement tasks (e.g., Endsley, 2000; Endsley & Rodgers, 1996; Kaber, Perry, Segall, McClernon, & Prinzel III, 2006).

A common SAGAT analysis strategy involves analyzing the queries measuring each level of SA separately. Wickens et al. (2005) point out that, "In dynamic systems, there is a fuzzy boundary between Stage 2 (understanding) and Stage 3 (prediction) because the understanding of the present usually has direct implications for the future, and both are equally relevant for the task" (pg. 2). While SAGAT queries are typically divided into the three levels of SA, it may be more applicable to instead focus on current and future states. Durso et al.'s (2006) SPAM queries were focused on past, present and future events (see also Durso, Bleckley, and Dattel, 2006). In another experiment, Durso et al. (1998a) asked six queries using both SPAM and SAGAT methods; three questions regarding the current state of the TRACON airspace and three regarding the future state. Interestingly, results indicated that controllers who were very accurate on current-oriented queries were less effective at the ATC task than those who were very accurate on future-oriented queries. Durso et al. (1998a) found that the study, "…also supplied evidence that comprehension of the current situation and projection into the future are distinguishable and important components in the SA of air traffic controllers" (pg. 17).

Even within the SA research using direct query measures, there are numerous methodologies that have been employed by researchers. Based on the objectives of the current experiment and the analysis techniques developed from Study 1, Study 2 SA queries focused on current and future states of the TRACON airspace. Current state queries included questions regarding aircraft groundspeed, altitude and heading. Future state queries included questions regarding aircraft arrival and departure points and altitudes.

The focus on current and future SA, as opposed to SA Levels I, II, & III, was selected because this research is a first step in determining the processes that occur during real world tasks. While distinguishing between the three levels of SA may be relevant at the theoretical level, the purpose of this research is to determine which aspects of eye movements contribute to the development and maintenance of SA. Current state queries encompass both perception and comprehension, and future state queries encompass perception, comprehension, and projection. It was expected that participant eye movements would predict performance on SA queries in a number of ways. The analysis was broken down by overall SA performance, current state SA performance and future state SA performance to determine if SA for different types of events was predicted by eye movements across time.

Analysis of overall SA was based on eye movement variables that assessed how much a person attended to an individual aircraft's AOIs, including both the aircraft icon on the radarscope and the corresponding flightstrip. It was hypothesized that percent of time spent fixating on an aircraft's AOIs would be positively associated with SA accuracy on the ten questions for that aircraft. This hypothesis was based on the Study 1 finding that percent time fixating on a task-relevant event was positively associated with SA for that event.

Number of fixations within the aircraft AOIs was also examined as a predictor for overall SA in Study 2. This measure differed from Study 1, which examined total number of fixations across an entire scene during an event. It was thought that more time fixating on an AOI is a combination of both a higher number of fixations and longer fixation durations. Thus, it was hypothesized that number of fixations in an AOI would be positively associated with SA accuracy for that AOI. This hypothesis was based on the finding in Study 1 that more time fixating on a task-relevant event was associated with higher SA for that event.

Mean duration of fixations on an AOI was also examined in Study 2. This measure also differed from Study 1, which examined mean duration of fixations across an entire scene during an event. It was hypothesized that mean duration of fixations on an

AOI would be greater for participants with higher SA scores. This hypothesis was also based on the finding in Study 1 that more time fixating on a task-relevant event was associated with higher SA for that event.

Additional eye movement measures were included in the current study's analyses that were not in Study 1. The total overall number of fixations over a specified amount of time was included as a variable in the analysis. This measure was included because Jacobs & Karn (2003) found that it was the most common measure used in their review of usability studies using eye tracking. In usability studies, total number of fixations is typically negatively associated with search efficiency. The current study is not a search task, so no true hypotheses are stated. The total number of fixations measure is similar to Study 1's measure of number of fixations during an event. In Study 1, the effect of number of fixations was dependent on the level of traffic. The ATC task in Study 2 will have a varying number of aircraft across the scenario, making predictions based on Study 1's results complicated.

One other included measure that has not yet been associated with SA is the Nearest Neighbor Index (NNI). A Simple Tool for Examining Eye Fixations (ASTEF) was developed to analyze fixation distributions in time series eye movement data such as in the current experiment (Camilli, Nacchia, Terenzi, & Di Nocera, 2008). The NNI is used as a spatial measure to determine distance between gaze points, regardless of direction, and is an estimate of whether fixations are randomly dispersed or more aggregated (Camilli et al., 2008). Initial use of the Nearest Neighbor Index has been to measure workload differences based on fixation distributions. Camilli et al. (2008)

believe that, “(NNI) might reflect the use of different visual scanning strategies… in complex and more demanding task situations, a wider fixation pattern (i.e., random or near random) might be used to optimize prompt attending to incoming information” (pg. 374). Since the Nearest Neighbor Index has only been compared to workload measures in the past, there were no hypotheses for the relationship between it and SA performance.

The overall SA variable was also divided into current SA and future SA for further analysis. The five predictor variables discussed above were included in the analyses for each of the three dependent variables (overall SA score, current SA score, and future SA score). Of the ten questions, seven measured current state SA while three measured future state SA. Durso et al. (1998a) found differences between current- and future- oriented controller performance on an ATC task. It is likely that some significant relationships will be due to either current or future SA performance levels, independent from the other.

All of the previous measures are based on eye movement data predicting SA. Thus, only eye movements made before the SA queries were answered were included in the analysis. Overall performance measures were also included to determine if eye movements predicted SA and performance differently. The overall eye movements throughout the scenario were examined to determine if the relationship between the performance measures and eye movement variables can be determined using overall measures. The performance measures included number of actions remaining 15 minutes into the scenario, and number of errors that occurred up to 15 minutes in the scenario. The following general eye movement measures were considered as possible predictors of

SA as well as overall performance: percent time fixating in the communication box overall, percent time fixating on aircraft and relevant fixes overall, percent time fixating on flightstrips overall, and percent time fixating on aircraft icons overall.

Multiple eye movement data files were used for the current analysis. Predictor variables that were determined using the aircraft level fixation data (including percent time fixating on aircraft AOIs, number of fixations on aircraft AOIs, mean fixation durations on aircraft AOIs, total number of fixations overall, and NNI values) were calculated over two time periods. The first included all data from the start of the scenario up to the query break. The second included data for the 60 seconds leading up to the query break. Examining the data set both ways allowed the experimenter to determine if SA at a particular moment is better predicted over a span of time, or is instead predicted by eye movements in a short span immediately preceding the break. Predictor variables determined using the scenario level fixations (including percent time fixating on aircraft, flightstrips, communications box, and airports and fixes variables) were also calculated over two time periods. The first included eye movements from the start of the scenario up to the query break, and the second included eye movements for the entire scenario.

Case studies of eye movements over time were also examined as a first step to determine if controllers' scan paths can reveal their level of SA. Several instances involving aircraft which could potentially or actually did conflict were examined. It was hypothesized that for events where future states needed to be anticipated, such as potential separation conflicts, scan paths would indicate planning by controllers. For example, if controllers focused on two aircraft which may conflict consecutively and repeatedly, they would take the appropriate actions to avoid conflict. Controllers who did not exhibit these eye movement patterns were not expected to take the appropriate corrective actions.

If the relationship between performance variables and eye movement variables can be determined using general eye movement measures, the more labor intensive analyses may not be necessary in future research. If instead, the precise eye movement analyses are better predictors of SA accuracy, future research should focus on the specific process measures.

Method

Participants

Sixteen certified air traffic controllers participated in this study. Due to problems with the eye tracking equipment and laptop, five participants' eye movement data were not recorded properly. The remaining 11 controllers' data sets were used in the analysis. Two participants were not able to complete the third scenario due to time constraints. Thus, 11 participants were included in analysis for the first two scenarios, and nine participants were included for the third scenario. Participants were recruited by the experimenter from ATC centers in South Carolina and Georgia including the cities of Greenville, Charleston, Myrtle Beach and Atlanta. Their experience levels ranged from1.5 to 26 years experience with an average of 7.27 years. Six were en route controllers and five were TRACON controllers. All had at least 0.5 years experience using radar with an average of 6.6 years. The experiment lasted two hours. Participants received $50 upon completion of the experiment.

Materials

As in Study 1, eye movement data were collected using a non-invasive Tobii 1750 eye tracker, sampled at 50 hz with a latency between 25 – 35 ms. Participants viewed the scenarios on a 17 in Tobii LCD computer monitor (800 x 600 resolution screen). The experimenter travelled to the participants' locations. All data collection occurred in small rooms where only the experimenter and participant were located, with no other observers present.

ClearView software. ClearView analysis software ran the Tobii and combined the eye movement data with keystrokes and recordings of what was occurring on the computer screen (Tobii Eye Tracker, 2006). At the start of the experimental scenarios, the researcher calibrated the eye tracker using the ClearView calibrator. During the experiment, the software recorded participant eye movement data along with the on-screen stimulus and mouse movements. Data were collected every 20 ms throughout each scenario.

TRACON ATC Simulator software. The TRACON II Air Traffic Control Simulator by Wesson International (1990) was used for the experiment. The simulator allows a controller to direct air traffic in an airspace around one of five large cities. The airspace includes a major airport and its associated satellite airports (TRACON II, 1990). For the current experiment, the Los Angeles sector was used, which includes the Los Angeles airport (LAX) and four satellite airports. The simulation screen is divided into four main sections: the radarscope, active and pending flight strips, and the communications box (See Figure 3.1).

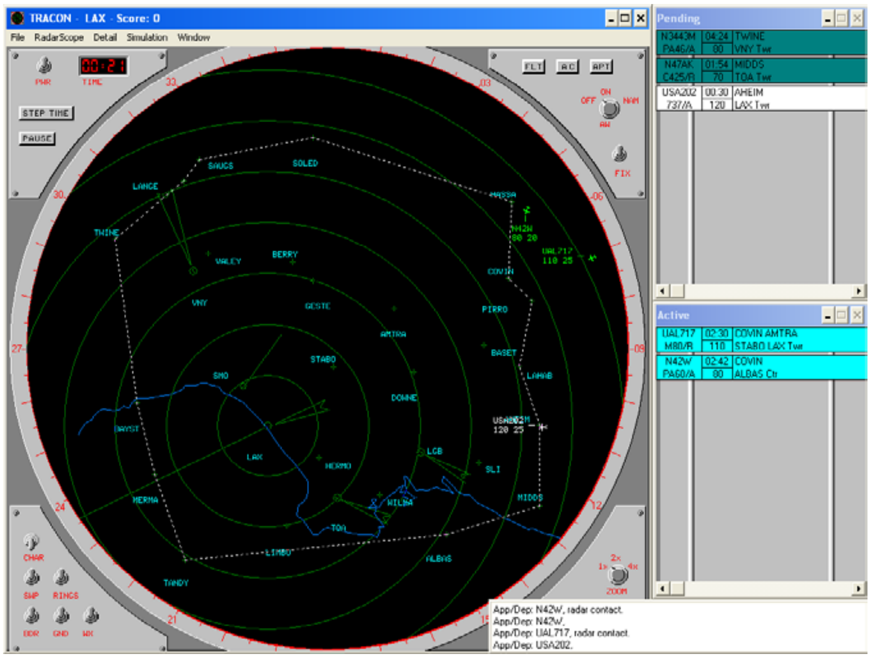


Figure 3.1. Screenshot of the TRACON II ATC Simulator. The radarscope takes up the majority of the screen, with the pending and active flightstrips and the communications box placed along the right-hand side.

Airports are marked with circles while intersections and radio beacons are marked with plus signs. Aircraft are marked with an aircraft icon and information about the aircraft is located in the datatag next to it. When controllers direct an aircraft, they are able to vector (turn) it, change its altitude or groundspeed, or ask it to hold at or move directly towards a particular fix (waypoint). The flightstrips inform controllers of the flightplans of aircraft entering or currently in their sector. The flightstrips describe the aircraft type, location, altitude, speed, and requested route. The communications box allows controllers to see all of the exchanges between themselves and the aircraft in their sector, as well as communication between themselves and tower and en route controllers. There are three types of flights in the simulation: overflights, departures, and arrivals.

Overflights are the simplest; they only require controllers to accept a handoff when a flight enters the sector and hand it off when it approaches the next sector's boundary. Departures must first be released by the controller before taking off, which requires them to pay attention to the communications box to ensure they do not miss a release request. Once released, an aircraft typically takes several minutes to reach the appropriate altitude to appear on the radarscope. Departures also need to be handed off to the next sector when they approach the sector boundary. Arrivals tend to be the most difficult; controllers must direct an arrival to the appropriate fix point and adjust its altitude for a proper landing. Arrivals must be within 300 ft of the appropriate landing altitude and 30 degrees of the airport's specified heading before reaching the final approach fix point in order to not miss their approach (TRACON II, 1990). The simulator also has several additional options for customizing scenarios, including pilot ability and probability for potential problems. Perfect pilots were selected, meaning no read back or execution errors occurred; there were also no weather problems or emergencies in the scenarios.

Scenarios. Participants completed two training scenarios before the start of the actual experiment. The first training scenario focused mainly on arrivals, because they typically require the most input from the controllers. The scenario contained four arrivals to various airports in the airspace, along with one departure and one overflight. The second training scenario contained six aircraft: two arrivals, two departures, and two overflights. During the second training scenario, the experimenter asked participants to look away from the computer screen 4.5 minutes into the scenario. Aircraft positions were marked on a paper printout of the radarscope and participants answered questions related to the current and future states of the aircraft in their sector. Once the training was complete, participants completed the experimental scenarios. The first experimental scenario contained ten aircraft: four arrivals, three departures, and three overflights. The second experimental scenario contained 12 aircraft: four arrivals, four departures, three overflights, and one practice flight, which took off from an airport and landed at another in the sector. The third experimental scenario contained 11 aircraft: five arrivals, five departures, and one overflight. The actual experimental scenarios each contained several instances where if no action was taken, a separation conflict between two aircraft would occur. Participant issued commands to avoid conflicts indicated adequate SA. The potential conflicts were implicit performance measures, thus it cannot be assumed that SA drove participant inputs. Prior to actual data collection, the training session and scenarios were pilot tested for realism and difficulty by five experienced controllers.

Performance measures. When a scenario was completed, an overall performance score was generated by the simulator. Points were deducted for separation conflicts, missed approaches, and handoff errors. These were combined with the number of commands issued to determine the participant's score, based out of a total possible score for each scenario. Due to the length of the scenarios, participants were typically unable to see them to completion and the resulting scores were not accurate or comparable across participants. Because of this, number of actions remaining and number of errors up to the 15 minute point in a scenario were used as indicators of performance in lieu of the computer generated score.

Situation awareness measure. The experimental scenarios were each stopped at a predetermined time, at which point participants answered a number of queries (Appendix A) in order to determine their level of SA. SA queries were presented to participants between eight and nine minutes into each scenario. Questions were developed using the SAGAT queries used by both Endsley (2000) and Endsley and Rodgers (1996) and current and future questions from Durso et al. (1998a, 2006). Personal correspondence with and feedback from Durso allowed the experimenter to further refine the query set.

For each query break, the experimenter asked participants to turn away from the monitor. The experimenter then paused the scenario and marked the locations of all aircraft on the radarscope by numbering them 1 through X on a paper print-out of the radarscope. Participants were queried about three pre-selected aircraft, which were the same for all participants within each scenario. The aircraft were chosen based on relevance to the scenario. Aircraft priority levels were determined by participants during pilot testing. Aircraft with medium or high priority levels by a majority of pilot test participants were chosen for the experimental queries. Comparison aircraft for queries about differences in altitude level or groundspeed between two aircraft were also pre-selected based on proximity and relevance to the queried aircraft.

Design

The study employed a mixed model design. All participants completed at least two scenarios, with nine of eleven participants completing all three. During each scenario, participants were stopped one time and asked to complete the set of SA queries for three aircraft. Their eye movements were tracked throughout each scenario. SA measures (overall, current, and future query accuracy) and performance measures (number of actions remaining and number of errors committed at 15 minutes) were the dependent variables. Eye movement measures (percent fixations on AOIs, number of fixations on AOIs, mean fixation duration on AOIs, total number of fixations overall, and the Nearest Neighbor Index measures) were the predictor variables (covariates). Analyses focused on the relationship between predictor and dependent variables, with a single case consisting of an individual participant's score on predictor and dependent variables for one scenario, or for particular aircraft within a scenario.

Procedure

Participants read and signed an informed consent form before beginning the experiment. They then filled out a demographic questionnaire detailing their years of ATC experience, as well as what areas of ATC (TRACON, tower, center) they have worked in and their previous ATC simulator experience (See Appendix B). Once the questionnaire was completed, they went through a self-paced TRACON II simulator training PowerPoint presentation developed by the experimenter and tested by the pilot participants. The training took around 30 minutes on average to complete. The presentation described the basic functionality of the simulator and its controls. It also highlighted discrepancies from typical ATC operations and possible issues that may arise when using the simulator controls. Throughout the training, the experimenter answered any questions from the participants.

Participants then completed two training scenarios on the TRACON II simulator. Once the training scenarios were completed and before beginning the experimental scenarios, the Tobii eye tracker was calibrated. Then participants completed three experimental scenarios. Any questions participants had about the experiment were answered once the experiment was completed. Participants were thanked for their time and given a thank you card with $50 enclosed for their participation.

Results and Discussion

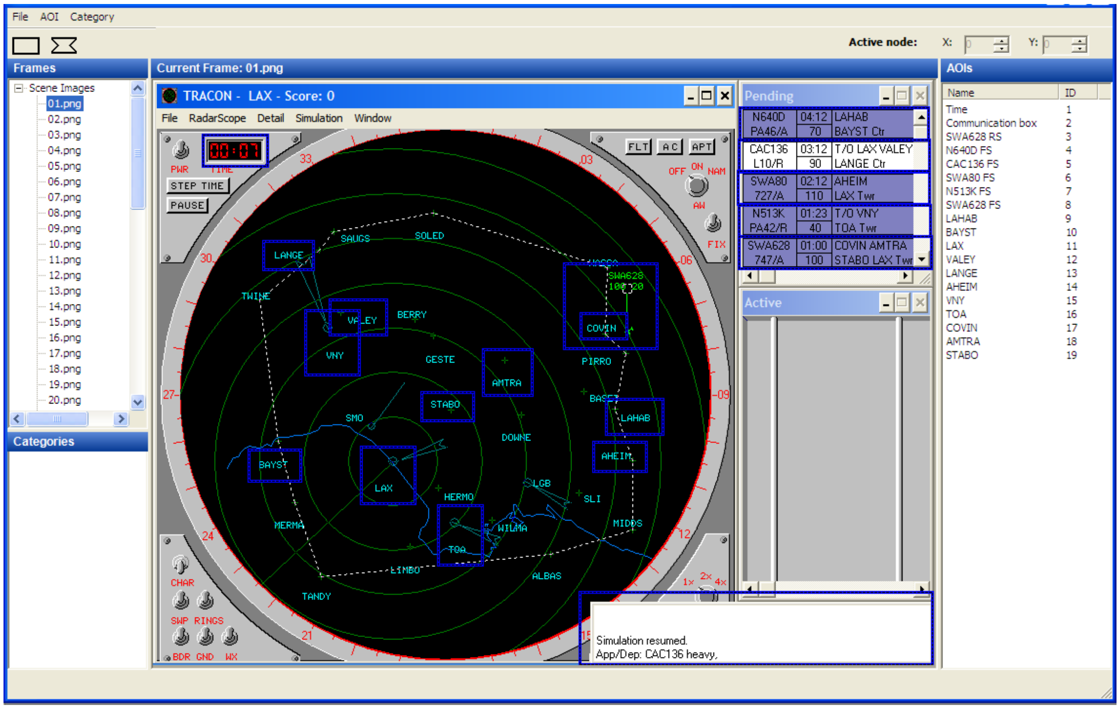
Data collection and scoring of eye movement variables

The ClearView software has an analysis function to allow for further examination of eye tracking data. Due to the dynamic nature of the scenarios, the experimenter had to manually define AOIs for each participant and each scenario. Given that participants issued aircraft control commands that, when followed, change the rest of that scenario, each participant experienced a different sequence of events during each scenario. Thus, the process of defining AOIs within each scenario had to be repeated for every participant. AOIs included objects that remained fixed (time box, communications box, airport locations, and relevant fixes) and objects that changed position during a scenario (flightstrips, aircraft icons and datatags). Aircraft icons and their datatags changed frequently during scenarios, while flightstrips changed position occasionally.

Before the experimenter was able to define AOIs, each scenario had to be broken down into scenes. Because the flightstrip for each aircraft was a relevant AOI that changed position only occasionally, scenes were defined within each scenario (and for each participant) whenever a new flightstrip appeared in the pending flightstrips section, a flightstrip moved from the pending to the active section, or a flightstrip was removed from the active section. In addition, if one of these actions did not occur for over 15 to

20 seconds, a new scene was marked. Number of scenes varied for each participant and scenario depending on the timing of flightstrip changes. Number of scenes ranged between 41 and 73 across participants and scenarios. The scenario with the fewest number of scenes averaged 20 seconds per scene; so, on average, scenes were less than 20 seconds long.

Once scenes for each scenario were established, AOIs were defined. For each scene, the experimenter defined all relevant AOIs, as described above. The aircraft on the radarscope moved once every seven seconds as the radar refreshed. At each scene change, all AOI positions were updated by the experimenter. AOI boxes around the aircraft were large enough to encase all movement of the aircraft and its leaderline and datatag for the time elapsed for the entire duration of that scene. As each scenario advanced, additional AOIs were added when aircraft or flightstrips appeared on the screen (See Figure 3.2).



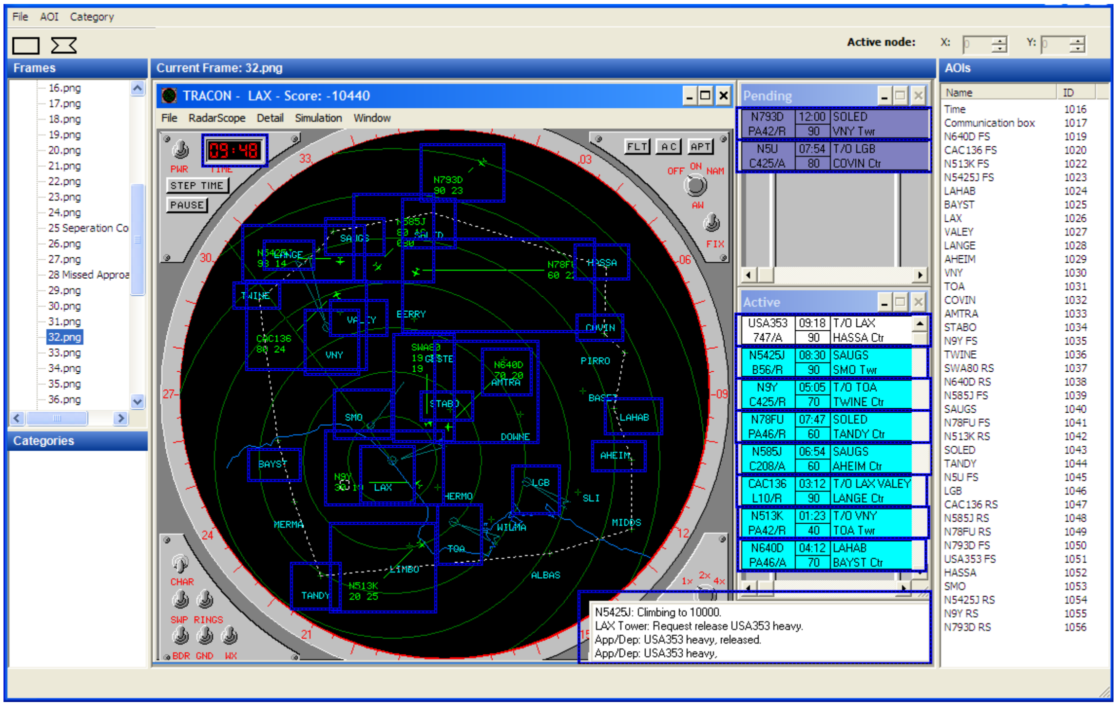


Figure 3.2. Screenshots illustrating how AOIs were defined. The top screenshot shows AOIs for the first scene in an example scenario while the bottom screenshot shows AOIs for a later scene in the same scenario.

The ClearView output for each scenario included a file of the AOI fixation output, which included the timestamp for each fixation, duration of fixation, and AOI name for the fixation (if an AOI was fixated upon). For Study 2 analysis, a fixation was defined as a mean fixation duration of 100 ms with a fixation radius of 30 pixels. The ClearView study settings recommend those settings for stimuli with mixed content (both pictures and reading). The experimenter combined all participants' data files into an aggregate file which included all AOI information for every participant for each scenario. Once combined, the experimenter converted the named AOIs into numbered AOIs for analysis purposes. For the first set of analyses, only the eye movement data leading up to a query stop was included. This was done because the analysis focused on how eye movements prior to a query break predicted SA score during the break.

It should also be noted that the initial fixation data files included many instances where more than one AOI was defined for a particular timestamp. These duplicate AOI lines occurred when several AOIs covered one another as the scenario progressed (See Figure 3.1 for an example; as the aircraft moved across the display, several bounding aircraft AOI boxes overlapped a fix, airport, or another aircraft's AOI box). All of the data lines, including duplicates, were included in the initial data file in order to ensure that all instances of fixations on a particular AOI were included. To calculate the variables of total number of fixations in an AOI, mean fixation duration in an AOI, and total amount of time spent fixating in an AOI, all of the fixation lines, including duplicate fixations, were included. To calculate the variables of total number and duration of fixations, the duplicates were first removed so there was only a single fixation for each timestamp. Then, the total duration of fixations in an AOI was divided by the total duration of fixations with duplicates removed to determine the percentage of fixations which were in a particular AOI.

The Nearest Neighbor Index (NNI) was determined using A Simple Tool for Eye Fixations (ASTEF) for the time period between the start of a scenario and the query break. There are two separate NNI number outputs based on different algorithms. Both the Convex Hull and Smallest Rectangle were included as independent variables. The Convex Hull algorithm "creates a temporary hull from the first 3 points, and then adds other triangles for each outer point," while the Smallest Rectangle algorithm "creates a bounding box for defining the rectangle having the smallest area comprising all the examined points" (Camilli et al., 2006, pg. 4). According to Camilli et al. (2006), an NNI score is the ratio of the mean distances between pairs of fixations in a set of actual fixations to the expected inter-point distances based on chance or random dispersion. When NNI is smaller than 1, fixations are more aggregated; when NNI is larger than 1, fixations are more dispersed in a regular pattern; and when NNI is close to 1, fixations are randomly dispersed (Camilli et al., 2006).

SA scoring

For each scenario, participants answered ten questions about each of three pre-determined aircraft in the scenario. Current state queries included seven questions related to each aircraft. Participant responses were scored as either correct or incorrect within a pre-specified tolerance level (e.g., altitude within 1000 ft.), which resulted in a number correct for current SA queries for each aircraft ranging between 0 and 7. Future state queries involved three questions regarding the future status of the three pre-determined aircraft (e.g., where will the aircraft be landing). Participant responses were scored as either correct or incorrect, which resulted in a future SA number correct for those aircraft ranging between 0 and 3.

Determining final eye movement predictor variables

The calculations just described yielded five eye movement predictor variables. Three of these were aircraft specific variables: the percent of time spent fixating in a specific aircraft's AOIs (calculated by dividing the amount of time spent fixating in an aircraft's AOIs divided by the total amount of time spent fixating), the number of fixations in an aircraft's AOIs, and the mean fixation duration in an aircraft's AOIs. Two of the predictor variables were based on eye movements across the whole scene (not specific AOIs): the total number of scene fixations and the NNI measure, calculated using both the convex hull and smallest rectangle algorithms. Due to the natural relationships between variables, regressions were run to determine if all predictor variables should be included in the analysis. In regression analysis, smaller tolerance values indicate that a predictor variable is highly correlated with other variables. When all six predictor variables were included in the regression (percent time fixating in the AOI, number of fixations in the AOI, mean fixation duration in the AOI, total number of fixations, NNI convex hull, and NNI smallest rectangle) both percent time fixating in the AOI and number of fixations in the AOI had tolerance values less than 0.1 (.071 and .062 respectively). When the regression was rerun with number of fixations in the AOI removed, percent time fixating in the AOI tolerance increased to .989. Due to the high correlation between the two variables, number of fixations in the AOI was removed from analysis. It was also determined that only one measure of NNI should be used. The tolerance levels for the NNI convex hull algorithm were slightly lower than the NNI smallest rectangle algorithm, so the NNI convex hull variable was removed from analysis. Once the two variables were removed, the remaining four predictor variables' tolerance statistics were all .9 or greater. Thus, two AOI specific variables (percent time fixating in an AOI, mean fixation duration in an AOI) and two scene-level variables (total number of fixations, and NNI smallest rectangle) were the eye movement predictor variables for the remaining analyses.

Scoring of ATC performance variables

The number of actions remaining was used as the performance measure for each scenario, similar to Durso et al. (1998a). Actions remaining were determined by the experimenter at the 15 minute point in each scenario. The fewest number of actions remaining based on the current position and altitude of the aircraft on the radarscope (and fewest actions required if pending aircraft remained) was determined. For example, if an aircraft was an arrival that had not yet landed, the experimenter determined the number of actions required to maneuver the aircraft to a particular heading and altitude, and then the handoff to tower was added as an additional action. Previous researchers have used this method by employing subject matter experts to determine the number of actions remaining (Vortac, Edwards, Fuller, & Manning, 1993). They explain that the actions remaining measure is, "a quasi-objective measure because, for a subject matter expert, there is little uncertainty about what actions are required for a given aircraft before it is handed over to an adjacent facility" (pg. 639). While actual air traffic control requires many more considerations when directing air traffic, the simulator allows for only a small number of discrete actions in order to successfully maneuver the aircraft. Although the experimenter is not an air-traffic controller, she has a high level of experience with the simulator and a basic understanding of the requirements for efficient maneuvering of the aircraft in a sector. The number of actions remaining for a particular aircraft ranged from 1 to 5. The total number of actions remaining for all aircraft still in the scenario at the 15 minute point made up the dependent variable.

The second performance variable was the number of errors committed by the controller during the scenario. The simulator generated error messages for missed arrival approaches, handoff to the next sector errors, and separation conflicts between aircraft. These errors were summed for each scenario and made up the number of errors performance measure. This approach was taken because the simulator generated a performance score which was made up of these error types, as well as additional factors (such as number of commands issued and number of aircraft landed) at the end of each scenario. Because participants stopped the scenarios at different times the performance scores were not consistent, number of errors up to 15 minutes in the scenario was thought to be a more accurate measure of performance.

Overall descriptive statistics

Descriptive statistics were calculated to ensure that there were no outliers in participants' scores overall and within each scenario. Overall, participants were 77% accurate on the SA queries (73% accurate for current queries and 86% accurate for future queries). The overall high SA scores were expected. However the 86% accuracy on future SA could be due either to very high future SA in these participants or relatively easy future SA questions. Individual participants' overall SA accuracy percentages ranged from between 60% to 97% for the individual scenarios. On average, participants fixated on the queried aircraft AOIs for 8% of the overall time between the start of the scenario and the query break, with a range from 4 to 13% for the individual scenarios. The mean number of fixations overall was 984, with a range from 696 to 1245 for the individual scenarios (see Table 3.1).

Table 3.1
Descriptive statistics for SA, performance, and eye movement variables for each participant.

	Mean	**SE**	**Min**	**Max**
SA overall (% correct)	76.9	2.0	71.2	92.2
SA current (% correct)	73.2	2.0	64.3	88.9
SA future (% correct)	85.8	3.0	70.8	100.0
Actions remaining	9.6	1.0	6.3	16.8
% time fix on queried aircraft	7.8	0.4	5.0	9.5
Mean fix dur. on queried aircraft (ms)	442	17.0	357	529
Total # scene fixations	984	32.9	741	1145
NNI smallest rectangle	0.65	0.01	0.61	0.68

The NNI smallest rectangle scores for the time frame before the query break ranged between 0.60 and 0.72, indicating that all participants had aggregated fixations (See Figure 3.3). The overall descriptive statistics indicated that all participants performed at a satisfactory level and no extreme variability in performance occurred.

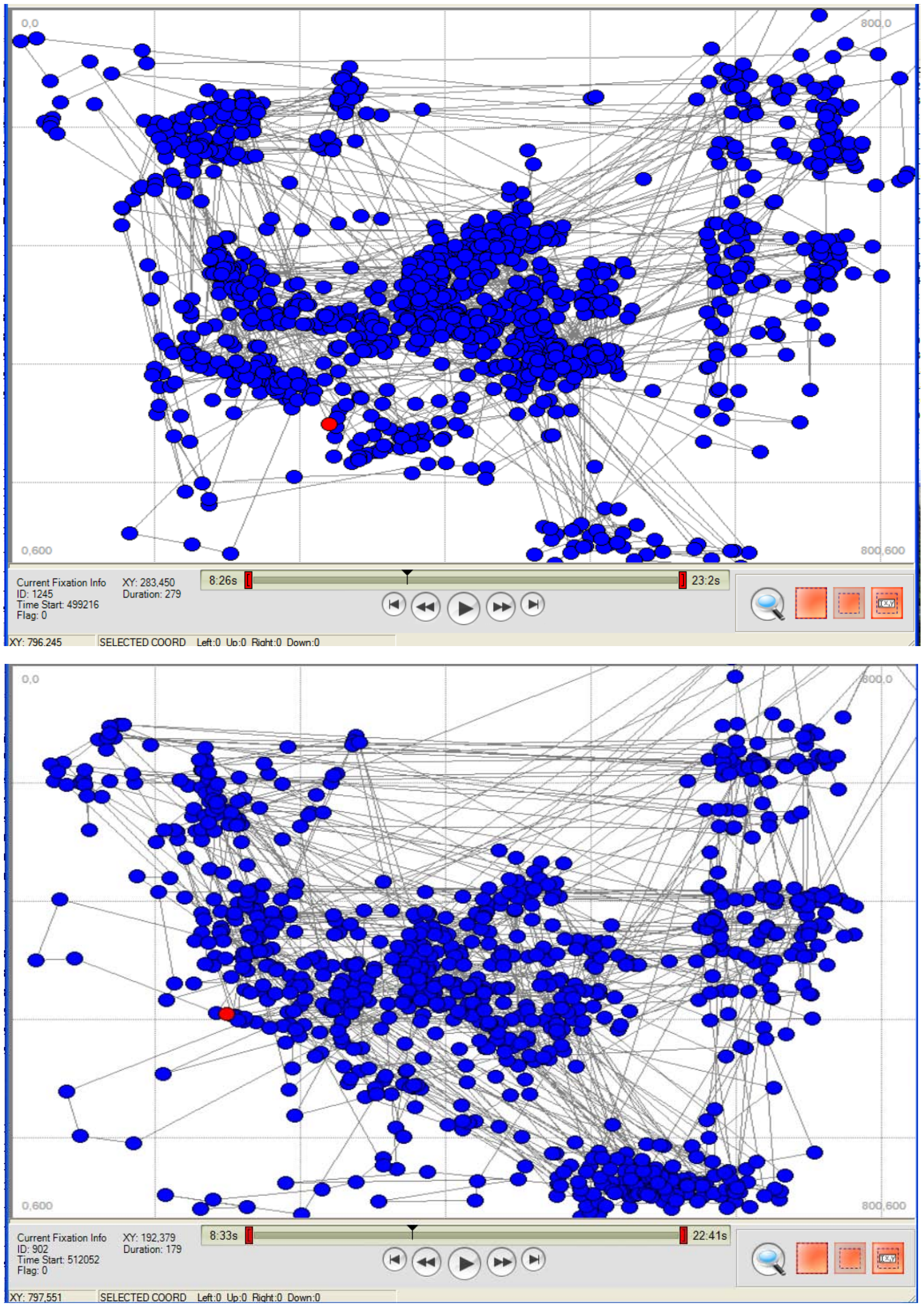


Figure 3.3. Lowest and highest participant NNI values. The top image shows fixations for a participant whose NNI smallest rectangle value equaled 0.6, while the bottom image shows fixations for a participant whose value equaled 0.72. According to Camilli et al. (2006), the top image (NNI = 0.60) has more aggregated fixations while the bottom image's fixations (NNI = 0.72) are slightly less aggregated and closer to random dispersion (NNI = 1.0).

Effects of eye movements on SA

It was hypothesized that percent of time spent fixating on an aircraft's AOIs (aircraft icon and flightstrip) would be positively associated with SA accuracy on the ten questions for that aircraft. It was also hypothesized that mean fixation duration on an AOI would be positively associated with SA accuracy for that AOI. In addition, it was hypothesized that less random, more aggregated eye fixations would be associated with higher SA; thus it was predicted that as the NNI score decreases, SA scores increase. Total number of fixations was included as a predictor variable because its relationship to performance has been measured in previous studies. A review of usability study research showed that total number of fixations is negatively correlated with search efficiency (Jacob & Karn, 2003). Because the current research is not based on search efficiency, no true hypothesis was stated; instead it was determined to measure number of fixations and see if it had an effect on performance in an ATC task. In addition, the analysis was divided into current and future SA scores in order to determine if particular eye movement variables predicted performance on one or the other category of SA in varying ways. Previous research has analyzed the three levels of SA, as well as current and future SA separately from overall measures (Durso et al., 1996; Endsley et al., 1998).

These hypotheses were tested using hierarchical linear modeling. Hierarchical linear modeling (HLM) allows for both fixed and random model effects, and accounts for repeated measures within the data set. Common regression analysis techniques do not account for multiple observations within a particular hierarchy and a primary assumption for analysis is that observations are independent from one another. On the other hand, HLM accounts for nesting and hierarchical variables (Osborne, 2000). The mixed model analysis procedures in SPSS 16.0 were used to run the hierarchical linear models. The models for the current study all used a random intercept. All predictor variables were fixed effects in the model, while participant was a random effect. In addition, mixed models calculate and output residuals, which were used to calculate effect sizes.

Defining variables for hierarchical linear models. Performance data for each participant was aggregated into an overall data file that included a separate case (or line) for each aircraft within each scenario for each participant (11 participants, 2 to 3 scenarios each, and 3 aircraft per scenario). In 5 cases (out of 93 separate aircraft queries), the participants' SA query responses indicated that they were responding based on their SA of another aircraft on the radarscope. These instances were identified by the experimenter when responses corresponded to a specific plane other than the one queried. These 5 cases were removed from analysis, leaving 88 cases in the aggregate data file. The dependent variables included the overall SA score for each aircraft (number correct out of 10), the current SA score (number correct out of 7) and the future SA score (number correct out of 3). Due to the high level of SA exhibited by the participants (See Table 3.2 for dependent variable descriptive statistics), the SA scores tended to have high kurtosis and negative skewness. In order to correct this, each SA score was converted to a proportion correct and then transformed via an arcsine-root transformation, as follows :

$$\text{Transformed SA score} = \text{arcsine} \sqrt{(\text{proportion of SA queries correct})}.$$

This transformation was selected because it is recommended for proportion correct data; and it resulted in acceptable skew and kurtosis values and a better model fit than with untransformed scores (Wheater & Cook, 2000).

Effect size was calculated using the within-subjects residuals from the mixed models output. For each mixed models analysis, the unique effect size for each predictor variable was calculated along with the total effect size. For example, in models with two predictor variables, unique variance in a dependent variable accounted for by predictor variable A (expressed in terms of R^2) was calculated as follows:

$$\text{Unique } R^2(A) = (B - AB) / \text{Intercept}.$$

Where B is the within-subjects residual for the model when only B is included as a predictor, AB is the within-subjects residual for the model when both A and B are included as predictors, and the Intercept is the within-subjects residual when no predictor variables are included.

A separate aggregate data file was created using the raw eye movement data for the 60 seconds leading up to the query break in each scenario. The same eye movement variables were calculated from this file as were used for the main analysis of eye movements for the entire time up to the break. The same mixed model analyses were run to determine if eye movement data for the 60 seconds leading up to the query break predicted performance on the SA queries differently than the overall time frame before the break eye movement data. It was thought that the 60 second data may predict performance better than the overall data because it might better capture the recently fixated upon AOIs; if the queried aircraft were recently fixated on, SA scores for those aircraft may be higher.

Table 3.2 shows the descriptive statistics, averaged across participants at the level of individual aircraft or individual scenarios, for the variables used in the hierarchical linear modeling analyses of the effects of eye movements on SA. The four main eye movement predictors are shown both for the entire time up to the query break (pre-break) and for the 60 seconds prior to the query break. The range of scores on the NNI smallest rectangle measure increased from 0.60 to 0.72 for the pre-break data to from 0.45 to 0.72 for the 60-second data, indicating that fixations in the 60 second time frame were more aggregated compared with the entire time before the break.

Table 3.2
Predictor and dependent variable descriptive statistics for aircraft and scenario for analysis of effect of eye movement variables on SA

	Min	Max	Mean	SE	Min	Max	Mean	SE
	Entire time before break				*60 seconds before break*			
Aircraft Level								
Predictor Variables								
% time fix on AOI	0	27	0.8	0.6	0	40	11	1
Mean fix dur on AOI	199	832	442	13	0	795	403	16
Total # scene fix	696	1245	981	14	78	159	116	2
NNI smallest rec	0.60	0.72	0.65	0.004	0.45	0.72	0.54	0.006
Dependent Variables								
Overall SA (max 10)	1	10	7.72	0.19				
Current SA (max 7)	1	7	5.14	0.14				
Future SA (max 3)	0	3	2.58	0.09				
Scenario Level								
Predictor Variables								
% time Comm box	2	17	7.9	0.6				
% time Airport/Fix	9	30	17.8	1.0				
% time Aircraft	29	87	63.0	2.0				
% time Flightstrip	10	25	14.5	0.8				
AC % Fix std. dev.	7	12	8.6	0.0				
Dependent Variables								
Overall % SA correct	60	97	77	2				
Current % SA correct	48	95	73	2				
Future % SA correct	56	100	86	3				

Effects of aircraft-specific and scene level eye movement variables on SA. The results of the mixed models testing effects of the two aircraft-specific and two scene-general eye movement variables on SA are in Table 3.2. The hypothesis that percent time fixating on an aircraft's AOIs would be positively associated with SA accuracy was supported for the time frame before the query break. Interestingly, percent time fixating on the AOI was the best predictor of overall SA score (R^2=.093); and there was also a significant main effect of percent time fixating on the current and future SA scores (R^2=.074 and R^2=.037 respectively). This result indicates that fixating on relevant aircraft is important to both current and future SA accuracy. There were no significant main effects for the other predictor variables for the time frame leading up to the query break; thus those hypotheses were not supported.

When the time frame included only the 60 seconds leading up to the query break, there were no significant effects of either aircraft-specific or scene-general eye movements on SA (see Table 3.3). Comparing the effects of eye movements on SA for the longer and shorter time frames supports the emphasis on building SA throughout a task, and indicates that a 'snapshot' is not enough. The 60 second time frame was an exploratory choice made to capture a lesser amount of time than the entire time before the break. Other time periods should be tested to better understand the building and maintenance of SA throughout a dynamic task.

Table 3.3.
Mixed model results for eye movement predictor variables and SA dependent variables.

	p	β	Unique R^2	*p*	β	Unique R^2
	Entire time before break			*1 min before break*		
Aircraft Level						
Overall SA						
% time fix on AOI	.01**	1.26	.093	0.98	-0.01	.000
Mean fix dur on AOI	.61	-0.00	.000	0.78	0.00	.000
Total # fix	.69	0.00	.000	0.38	0.00	.000
NNI smallest rec	.48	0.61	.000	0.76	-1.44	.000
Current SA						
% time fix on AOI	.02*	1.19	.074	0.83	-0.09	.000
Mean fix dur on AOI	.61	-0.00	.000	0.95	0.00	.000
Total # fix	.67	0.00	.000	0.26	0.00	.000
NNI smallest rec	.32	0.88	.000	0.86	-0.09	.000
Future SA						
% time fix on AOI	.05*	1.89	.037	0.90	0.09	.000
Mean fix dur on AOI	.72	-0.00	.000	0.39	0.00	.000
Total # fix	.54	-0.00	.000	0.92	0.00	.000
NNI smallest rec	.98	-0.03	.000	0.96	0.04	.000
Scenario Level						
Overall % SA						
% time Comm box	.41	0.44	.000			
% time Airport/Fix	.36	0.29	.000			
% time Flightstrip	.02*	1.39	.122			
% time Aircraft	.23	0.22	.019			
Overall % SA						
AC % Fix Std. Dev.	.04*	-0.25	.057			
Current % SA						
% time Comm box	.28	0.62	.000			
% time Airport/Fix	.21	0.42	.022			
% time Flightstrip	.04*	1.39	.070			
% time Aircraft	.08	0.35	.081			
Current % SA						
AC % Fix Std. Dev.	.50	-0.91	.000			
Future % SA						
% time Comm box	.92	0.08	.000			
% time Airport/Fix	.91	-0.05	.000			
% time Flightstrip	.12	1.39	.053			
% time Aircraft	.76	-0.08	.000			
Future % SA						
AC % Fix Std. Dev.	.00**	-5.93	.335			

$*p \leq .05$, $**p \leq .01$

Effect of fixating on general AOIs on SA. The preceding analyses tested how SA was affected by fixating in small AOIs (representing relevant aircraft) or by general patterns of fixating the whole scene. In this set of analyses, general AOI predictor variables were calculated in addition to the above predictor variables. The general AOI analysis was exploratory in nature; it was done to determine if examining the eye movement data in larger groups of AOIs predicted SA. The four general AOI groups included in this analysis were the communication box, all airports and relevant fixes, all flightstrips, and all aircraft icons. The percentage of time fixating on each group was calculated by determining the total duration of fixations on all AOIs in the group and then dividing this total duration by the total duration of fixations overall once duplicate fixations were removed. Due to overlapping AOIs on the radarscope, percent time fixating on aircraft and percent time fixating on airports and fixes are likely inflated for a majority of participants. Not surprisingly, participants tended to fixate for the majority of the time on the aircraft icons (M=63%), followed by the additional information on the radarscope including airports and fixes (M=18%), flightstrips (M=15%) and finally the communication box (M=8%) (See the descriptive statistics in Table 3.2). Diagnostics revealed no multicollinearity problems when these four predictor variables were included in a regression. Since the general AOIs were calculated at the scenario level (as opposed to the aircraft level), the percentage correct SA for each scenario was the dependent variable. The percentage correct variables had acceptable skewness, kurtosis and model fit values; therefore they were used in the analysis without transformation.

Mixed model analyses revealed a significant main effect of percentage time fixating on the flightstrips on overall SA accuracy (see Table 3.3). As percentage of time fixating on flightstrips increased, overall SA increased (R^2=.122). The main effect of percentage time fixating on flightstrips was also significant for the current SA accuracy DV, though the effect size was smaller than the overall SA effect (R^2=.07). Although not significant, the effect size for percentage time fixating on a flightstrip on future SA accuracy was also worth noting (R^2=.053).

Interestingly, one other general AOI predictor variable also had notable effect size, though the results were not significant. The percentage time fixating on aircraft AOIs showed an effect size of .06 on overall SA accuracy and an effect size of .08 on current SA accuracy (both small effect sizes).

These results suggest that the percentage time fixating on larger groups of AOIs have an effect on controllers' SA. The most notable finding was that spending more time fixating on flightstrips led to a significant improvement in overall and current SA accuracy. Attending to flightstrips implies planning by the controllers. The flightstrips describe where aircraft will be entering and exiting the airspace, as well as their requested altitude. The aircraft icons on the radarscope inform controllers of where the aircraft are located at the current time, but the flightstrips inform controllers of the planned future movements of these aircraft as well as the planned future movements of aircraft not yet on the radarscope. It should also be noted that although not significant, effect sizes suggested a tendency for overall and current SA to increase with percent time fixating on aircraft. This trend should be explored further in future research. It is especially interesting that both percent time fixating on flightstrips and on aircraft had considerable effect sizes, considering that there was a significant negative correlation (-.64) between flightstrip fixations and aircraft fixations ($p < .000$).

Effect of focusing or distributing attention to aircraft on SA. Another exploratory analysis was based on the observations that participants typically appeared to distribute attention equally to all of the aircraft in the environment, but when they focused too long on a particular aircraft they were more likely to miss important information in the airspace. To capture whether participants focused attention mainly on a few aircraft or distributed attention equally to all aircraft, the standard deviation of the percent time fixating on each aircraft was calculated. This variable was calculated for each scenario and each participant. Lower standard deviations would be associated with more equal distribution of attention across aircraft, and higher standard deviations with focusing attention. Also, controllers with lower standard deviations would be expected to have higher SA for the airspace in a particular scenario.

The standard deviation was determined by first calculating the percent time fixating on each aircraft icon in a scenario, which was determined by dividing the total duration of fixations on a specific aircraft icon by the total time spent fixating on all aircraft icons in a scenario. Once the percentage fixation times were calculated for each aircraft, the standard deviation of these percentages was calculated. If an aircraft was on the radarscope it was included in analysis, even if it was not fixated upon. The standard deviation variable was included as the only predictor in separate mixed model analyses with overall, current and future SA percent correct as the dependent variables.

Surprisingly, although there was very little difference between the minimum and the maximum standard deviations (7% to 12%, see Table 3.2 for descriptive statistics), there were significant main effects for aircraft standard deviations for both overall SA and future SA percent correct (R^2=.057 and R^2=.335 respectively, see Table 3.3). Since the aircraft standard deviations were significant with a large effect size for future SA and not significant with no effect size for current SA, it is assumed that the future SA effect size is driving the overall SA effect. In this effect, as participants' standard deviations of percent time fixating on specific aircraft in a scenario decreased (i.e., as they distributed their attention more equally across all aircraft), future SA score percent correct increased (See the scatter plot in Figure 3.4). This was a surprising finding, especially considering the large effect size.

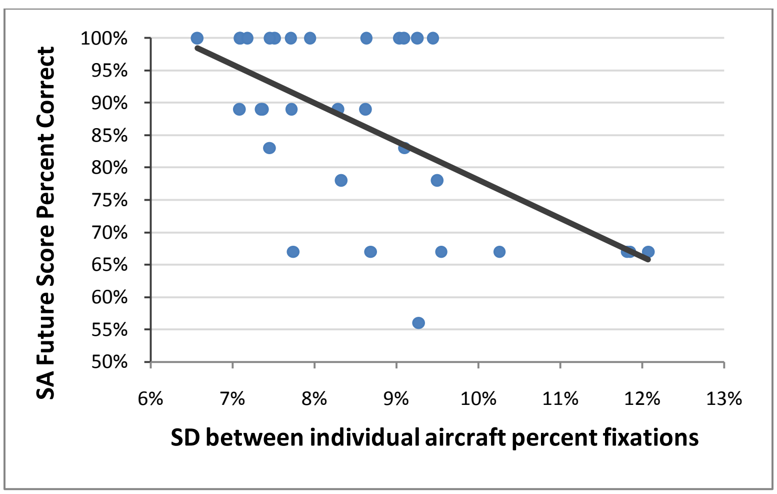


Figure 3.4. Standard deviation between percent fixations upon individual aircraft by SA future queries percent correct

Effects of eye movements on performance

No specific hypotheses were advanced regarding relationships between eye movements and ATC performance variables. Researchers have previously found positive correlations between SA and performance measures (e.g., Strybel, Vu, Kraft, & Minakata, 2008). Prince, Ellis, Brannick, & Salas (2007) found a positive correlation between team SA knowledge in a low-fidelity flight simulator and performance in a high-fidelity flight simulator (r=.41). Salmon et al., (2008a) found significant positive correlations between overall SAGAT SA scores (r=.662) and Level 2 SAGAT scores (r=.691) and performance on a military planning task. Because of this, it was expected that the same eye movement predictor variables used in the preceding SA analyses would also be positively associated with overall performance variables.

Performance dependent variables. Two measures of performance were used in the current analyses, number of ATC control actions remaining at the 15 minute point in a scenario and the total number of errors made during a scenario. The number of actions remaining has been used previously as a performance measure in SA research (Durso et al., 1998a; Vortac et al., 1993). Control actions included in the number of actions remaining variable consisted of releasing any remaining aircraft for takeoff, maneuvering the aircraft to the appropriate flight level and direction for landing or handoff, and handing off the aircraft to either tower or en route controllers. The number of errors in a particular scenario was the sum of handoff errors, missed approaches, and separation conflicts occurring up to the 15 minute point in each scenario. The experimenter viewed each participant’s gaze replay and tallied each error message generated by the simulator.

The dependent variable was the sum of the three error types. Descriptive statistics on these performance variables are shown in Table 3.4). The number of errors in a scenario mean was low, likely due to the controllers' expertise in ATC.

Table 3.4
Eye movement predictor variables and performance dependent variables descriptive statistics

	Min	Max	Mean	SE
	Entire time before break			
Aircraft Level				
Predictor Variables				
% time fix on AOI	0	27	8	0.6
Mean fix dur on AOI	199	832	442	13
Total # scene fix	696	1245	981	14
NNI smallest rec	0.60	0.72	0.65	0.004
Scenario Level				
Predictor Variables				
% time Comm box	2	15	7	0.5
% time Airport/Fix	10	29	18	0.9
% time Aircraft	41	100	77	2.5
% time Flightstrip	7	25	11	0.7
AC % Fix std. dev.	3	6	4	0.1
Dependent Variables				
# Actions Remaining	4	22	9.23	0.43
# Errors in scenario	0	2	0.69	0.09

Mixed model analysis. Hierarchical linear models were again used for the analysis. The first analysis measured the effect of the four previously discussed eye movement variables (the aircraft-specific and scene-level variables) on performance measures. The four eye movement predictor variables were calculated from the beginning of the scenario up to the query break. Thus, the analysis measured whether eye movement variables from the first half of the scenario predicted performance over the whole scenario. All four variables predicted performance (see Table 3.5).

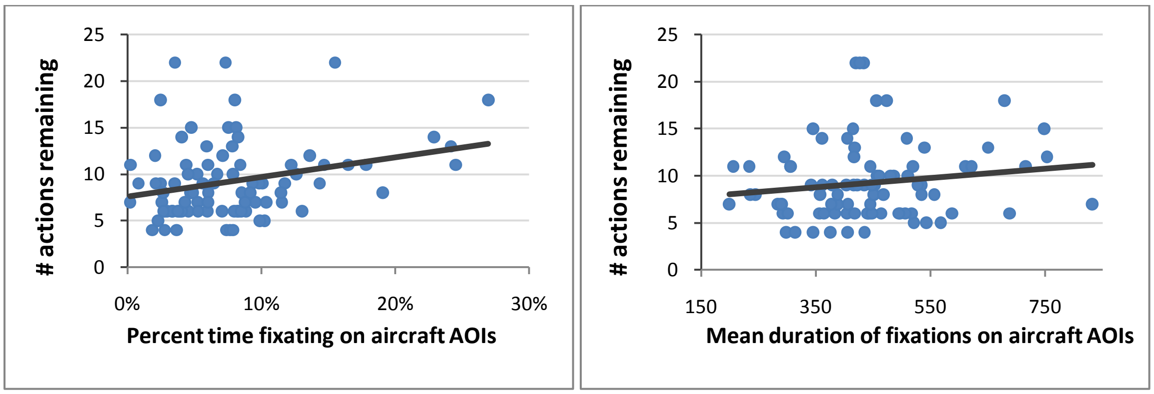
Table 3.5

Mixed mode results for eye movement predictor variables and performance dependent variables

	p	B	Unique R^2
	Entire time before break		
# Actions remain			
% time fix on AOI	.00**	0.92	.114
Mean fix dur on AOI	.01*	0.0004	.067
Total # fix	.93	-0.00002	.000
NNI smallest rec	.35	-0.54	.000
# Errors in scenario			
% time fix on AOI	.24	-0.76	.000
Mean fix dur on AOI	.91	0.00003	.000
Total # fix	.02*	-0.001	.047
NNI smallest rec	.02*	3.25	.058
	Whole Scenario		
# Actions remain			
% time Comm box	.68	-1.24	.000
% time Airport/Fix	.01*	3.58	.284
% time Flightstrip	.67	-0.99	.000
% time Aircraft	.74	0.22	.010
# Actions remain			
AC % Fix Std. Dev.	.32	-1.06	.110
# Errors in scenario			
% time Comm box	.51	2.55	.000
% time Airport/Fix	.90	0.18	.000
% time Flightstrip	.58	-1.70	.000
% time Aircraft	.20	-1.13	.000
# Errors in scenario			
AC % Fix Std. Dev.	.37	-1.03	.028

$*p \leq .05$, $**p \leq .01$

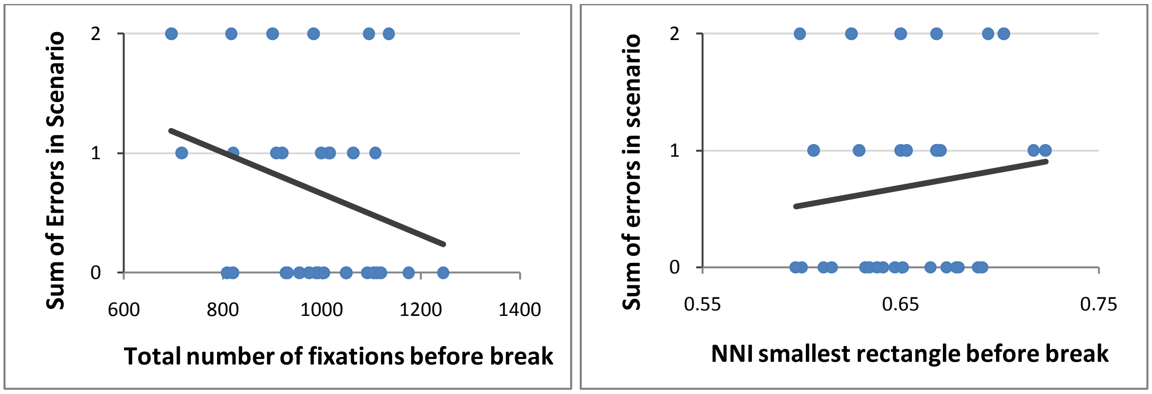
There were significant effects for the two aircraft specific measures, percent time fixating on an aircraft's AOIs and the mean fixation duration on an aircraft's AOIs, on number of actions remaining (R^2=.114 and R^2=.067, respectively). As the percent time fixating on aircraft AOIs or the mean fixation duration on aircraft AOIs increased, the number of actions remaining increased. Though this appears to be in the unexpected direction (with more focusing on important aircraft associated with poorer performance), it is important to understand that these two predictor variables measured fixation data for only the three aircraft queried. The results indicate that participants who fixated longer and for more time on those three aircraft had more actions remaining later in the scenario (See Figure 3.5a and 3.5b). This result will be explored further in a later analysis.



Figures 3.5a and 3.5b. (a) Percent time fixating on aircraft AOIs by number of actions remaining and (b) Mean duration of fixations on aircraft AOIs by number of actions remaining

There were also significant effects for the two scene-level eye movement measures, total number of scene fixations and NNI smallest rectangle, on the number of errors in a scenario (R^2 = .047 and R^2 = .058, respectively) (see Table 3.5 and Figures 3.6a and 3.6b). An increase in errors was associated with more scene fixations and a higher NNI smallest rectangle value. Recall that as NNI means increase towards 1, fixations are less aggregated and more randomly dispersed. Thus, these results show that errors increased when participants made more fixations and distributed their fixations more randomly. Camilli et al. (2008) provided evidence that more demanding task situations would be associated with NNI values closer to 1. When the NNI is examined as a measure of workload, as NNI values increased the individual's workload increased.

Thus, as the workload of participants increased the number of errors in the scenario also increased. This is interesting considering all participants experience the same scenarios, but the NNI scores indicate perceived workload levels may have varied.



Figures 3.6a and 3.6b. (a) Total number of fixations before the query break by sum of errors in the scenario and (b) NNI smallest rectangle value calculated using fixations up to break by sum of errors in the scenario.

Effect of fixating on general AOIs on performance. Another analysis was conducted to test whether the general AOI variables used in the previous analyses (i.e., percentage fixations in large groups of AOIs, e.g., flightstrips) predicted ATC performance (actions remaining and errors). Since the performance variables were measured over the entire scenario, the general AOI variables were also calculated over the entire scenario. The only significant effect was for percent of time spent fixating on airports and fixes on actions remaining ($R^2 = .284$). As percent time fixating on the airports and fixes on the radarscope increased, the number of actions remaining increased (See the scatter plot in Figure 3.7).   A possible explanation of this result is that allocating too much attention to information that was not as important as the aircraft and flightstrips prevented attention from being allocated to other tasks, thereby leading to additional actions remaining. Fixating on aircraft and relevant fixes was originally included because these AOIs were thought to be associated with future planning. Fixations on those areas of the radarscope typically occur when determining where aircraft will be heading. Another possible explanation is that this result is an artifact of how the AOIs were defined and the increased fixations on airports and fixes is due to overlapping AOIs.

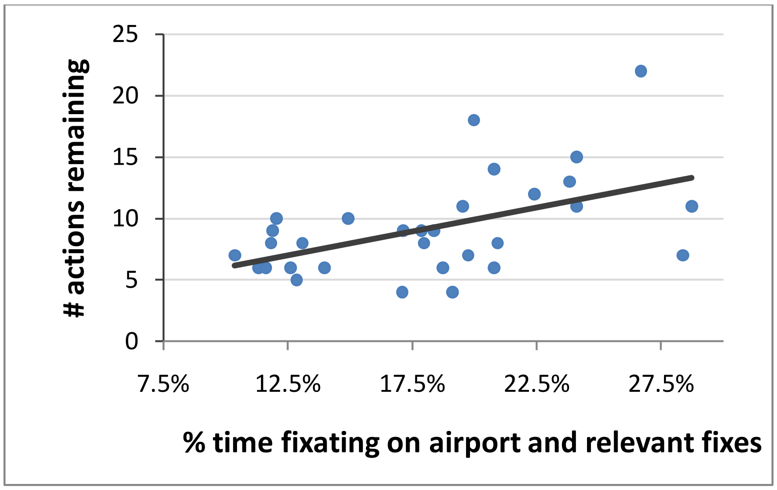


Figure 3.7. Percent of time fixating on airports and relevant fixes by number of actions remaining at the end of the scenario.

Effect of focusing or distributing attention to aircraft on performance. The standard deviation of percent time fixating on individual aircraft icons was calculated for the whole scenario (as opposed to up to the query break). It is interesting that when the standard deviation variable was calculated over the entire scenario, there was only a small amount of variance in the standard deviation of percentage fixations, suggesting that over the whole scenario, controllers were very similar in how they allocated attention to the aircraft icons (See Table 3.4 for descriptive statistics). The hierarchical linear model showed no significant effects of the standard deviation on either performance variable (see Table 3.5).

Effects of SA on performance

Recall that SA and performance are not directly related; instead, high SA typically leads to higher levels of performance, but that is not always the case (Endsley, 1995c; Wickens, 2008). While the relationship between SA and performance is not 1 to 1, many researchers agree that an individual's understanding and awareness of his immediate surroundings directly affects and can predict his performance (Durso, Bleckley & Dattel, 2006; Durso & Sethumadhaven, 2008). Implicit SA performance measures assume a direct link between operators' SA and their performance (Sarter & Woods, 1995). Numerous studies have found a significant correlation between SA and performance on a task (e.g., Endsley et al., 1998; Prince, Ellis, Brannick, & Salas, 2007; Salmon et al., 2008a). One recent study found a significant effect of SA on performance, measured using SAGAT, using the TRACON II simulator used in the current experiment (O'Brien & O'Hare, 2007). Based on previous research, it was expected that as SA increased, performance on the ATC task would increase.

As in the previous analyses, the two performance measures were number of actions remaining at the end of the scenario and number of errors during the scenario. The number of errors in each scenario were very low (from 0 to 2); though this is not surprising considering the participants were all expert air traffic controllers. (See descriptive statistics in Tables 3.2 and 3.4).

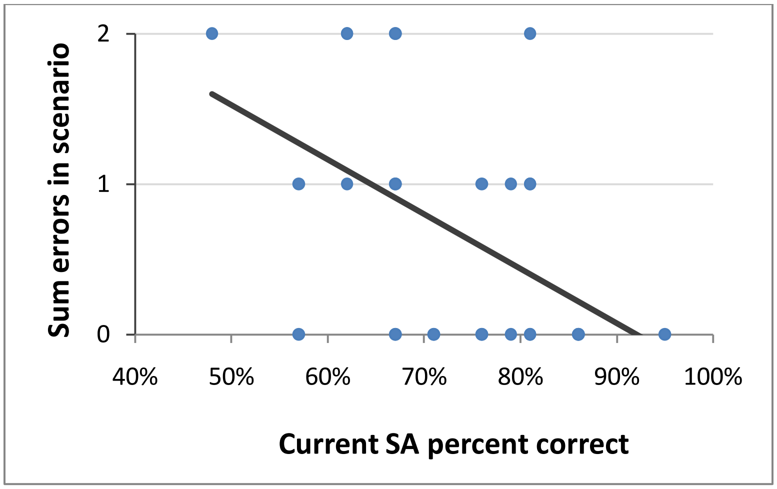
Mixed model analysis. The analysis examined the effect of SA query response accuracy (overall SA and current and future SA) on the performance measures (actions remain and number of errors in a scenario). When overall SA percent correct was included in the model there was no significant effect for number of actions remaining. When current and future SA percent correct were included as the predictors in the model; again there were no significant results for the number of actions remaining.

The mixed model analyses for overall SA and for current and future SA were again run with number of errors as the dependent variable. Results revealed that when overall SA percent correct was included in the model, there was a significant effect for the number of errors (R^2 = .244) (see Table 3.6). When current and future SA percent correct were included, there was no effect for future SA percent correct but there was a significant effect for current SA percent correct (R^2 = .244) (see Table 3.8). The results indicate that the percent correct on current SA queries is driving the overall SA percent correct effect. Thus, as current SA increased, the number of errors decreased (See the scatter plot in Figure 3.8).

Table 3.6.
Mixed model results for SA predictor variables and performance dependent variables.

	p	β	Unique R^2
# Actions remaining			
Overall SA % Correct	.27	-0.855	.000
# Actions remaining			
Current SA % Correct	.60	-0.401	.000
Future SA % Correct	.40	-0.465	.000
# Errors in scenario			
Overall SA % Correct	.00**	-2.330	.244
# Errors in scenario			
Current SA % Correct	.01**	-1.859	.244
Future SA % Correct	.36	-0.456	.000

$*p \leq .05$, $**p \leq .01$



Figures 3.8. Current SA percent correct by sum of errors in the scenario.

Study 2 general discussion

The results of Study 2 revealed novel findings about the relationship between eye movement variables and SA. The results underscore the importance of attention allocation in dynamic tasks. Tasks such as air traffic control require both focused and distributed attention (see Hauland, 2008). Recall that focused and distributed attention are different than the focal and ambient visual channels. Focused and distributed attention are both components of the focal visual channel; whereas the ambient channel is separate and involves the periphery of the visual field. In order to have a high level of SA and a high performance level in ATC, controllers must continually sample the airspace (typically employing a distributed attention strategy) to ensure that no separation conflicts occur. When an aircraft makes a request, or when controllers need to give instructions to avoid conflict, focused attention is required to make sure that whatever situation that required their attention is carried out properly. Even while they're focusing on a specific portion of the radarscope, controllers must continue to have a high level of SA for what is occurring around their sector to determine whether other aircraft require their attention. Consequently, controllers need to utilize both focused and distributed attention strategies; too much of one or the other lead to low SA and in turn, may lead to lower performance levels. The results of Study 2 support the idea of appropriate allocation of focused and distributed attention for high SA and performance.

As hypothesized, spending more time fixating on aircraft AOIs pre-selected by the experimenter to be high priority was associated with higher SA for those aircraft. This effect was seen when the analysis was run using overall SA score as well as both current and future SA scores. The effect size was larger for the overall SA score (9% of variance) compared with the individual current (7%) and future (4%) components, indicating that fixating on important aircraft AOIs increases awareness of both current and future events related to those aircraft. Therefore, focusing attention on important components of the scene increased SA. This result was not seen when the data set included only the 60 seconds leading up to the query break; highlighting the importance of building SA for events over time. Mean fixation duration on the pre-selected aircraft AOIs was another variable that measured attention to important information, but this variable did not predict SA.

When the data were examined using scene general AOIs, percent time fixating on flightstrips significantly predicted percent correct on the SA queries; that is, as percent time fixating on flightstrips increased, both overall and current SA increased. This result again illustrates that focusing attention on high-priority scene components can increase SA. It is somewhat surprising that fixating on flightstrips significantly predicted current SA, but not future SA. The future SA variable likely experienced a ceiling effect, because there were only three queries for each aircraft and performance was very high (M = 86% correct). The significant result for current SA was surprising because the flightstrips are mainly used in planning; they notify controllers of the intentions of the aircraft currently in the sector and the aircraft that will be entering the sector in the near future. The flightstrips inform controllers of what fix or airport aircraft will be entering and exiting the sector from, as well as the altitude at which aircraft need to be when exiting the sector if the flight is a departure or an overflight, or the altitude at which they will be entering if the flight is an arrival. The information on a flightstrip is static; it does not change as aircraft move through the airspace. Given the future-focused nature of flightstrips, the finding in this study that controllers who fixated more on flightstrips had higher SA seems to support research by Durso et al. (1998a), who found that future focused controllers (controllers who answered more future SA questions correct than current SA questions) had fewer actions remaining at the end of a scenario than current focused controllers (controllers who answered more current SA questions correct than future SA questions).

I also found that participants' future SA scores were predicted by the standard deviation of percent time fixating on individual aircraft in a scenario. In particular, as the standard deviation increased, future SA decreased; and this effect showed a large effect size. This predictor variable was initially chosen for analysis because informal observations during testing suggested that participants distributed their focal attention widely during routine operations and more narrowly for high priority events. An increase in the standard deviation of fixations on individual aircraft indicates a shift from a more distributed to a narrower allocation of focal attention on aircraft. Thus, this result suggests that overfocusing on a few individual aircraft impaired future SA scores, and that distributing attention more widely among aircraft is improved future SA.

The next set of analyses investigated how eye movement variables predicted ATC performance. The performance measures were based on two different aspects of successful air traffic control; the number of errors measured accuracy, whereas the number of actions remaining measured efficiency. Arguably the more important of the two performance variable is errors. Two eye movement variables reflecting how controllers scanned the entire scene, number of fixations and NNI, significantly predicted the number of errors that occurred throughout a scenario. First, as number of fixations increased, number of errors decreased. More fixations on the scene may indicate a more distributed allocation of focal attention. Thus this result suggests that as attention is distributed more widely, number of errors decreased; and conversely, as attention narrowed, number of errors increased. The NNI result appears to be an indicator of participant workload; where participants with NNI's nearer to 1, who likely were experiencing higher workload, had more errors in the scenario (Camilli et al., 2008).

The number of control actions remaining at the end of a scenario is a variable that reflects the efficiency of controllers' performance. Thus, it seems less critical than avoiding errors. There were significant effects of percent time fixating on aircraft AOIs and mean fixation duration on aircraft AOIs on actions remaining. This was somewhat surprising because as percent time fixating on aircraft AOIs and mean fixation duration on aircraft AOIs increased, the number of actions remaining increased. One explanation for this result is that participants who focus attention narrowly on important aircraft (by fixating more and for longer durations) are not distributing enough attention to other key scene events, resulting in more actions remaining than participants who focus on all aircraft for similar amounts of time.

In addition, it was found that as percent time fixating on airports and fixes increased, number of actions remaining increased. This result shows that narrowly focusing on less important AOIs negatively affected number of actions remaining. Participants who fixated more on airports and fixes may have done so because they were having a more difficult time learning the airspace compared with participants who fixated less. The LAX airspace was unknown to the participants; they had to familiarize themselves with the airports and fixes to understand the flight paths of the aircraft. Participants who quickly learned the placement of the airports and fixes would need to fixate on them less than those who did not.

Previous research has shown a significant relationship between SA and task performance measures (e.g., Strybel, Vu, Kraft & Minakata, 2008; Prince, Ellis, Brannick, & Salas,2007). Consistent with previous findings, the current results showed that as current and overall SA scores increased, number of errors decreased. SA was not significantly related to number of actions remaining.

Separation Conflict Case Studies

It is important to consider other potential eye movement measures of SA within a scenario beyond the pre-specified aircraft queried. Several additional questions were included during the freeze break, including one asking participants to “list the pairs of aircraft that have currently lost separation or will lose separation if they stay on their current courses.” It was expected that this question would be used to determine if participants recognized a potential conflict before one occurred. If a separation conflict occurred in the scenario, participant responses would indicate if they had simply not recognized that the aircraft were in conflict or had recognized the potential conflict but had not taken the proper actions to avoid it. Pilot testing indicated that the scenarios were fairly difficult for trained controllers and participant errors, including separation conflicts, were expected. While it is fortunate that the actual test participants made very few errors, there were not enough conflicts with which to analyze and draw conclusions. Overall, there were three separation conflicts that occurred due to inattention to the situation. Two of these conflicts happened in one scenario and likely occurred because the participant was distracted by noise in the testing room. In addition to these three conflicts, two other separation conflicts occurred because of differences between the simulator and actual air traffic control conflict rules. Two controllers thought that once they handed off an aircraft to Tower or the next sector, these aircraft could no longer conflict with traffic in their sector. The simulator still generated errors when two aircraft violated space requirements, regardless of whether or not the aircraft was under the controller’s control.

Due to the low number of conflicts, eye movements for individual participants were examined to determine if there were differences in eye movement patterns between participants who had conflicts and those who did not. The experimenter determined that there were three types of response to a potential conflict. First, participants typically recognized a potential conflict and issued commands to avoid it as soon as the potential conflicting aircraft were both present in the scenario (preventive planning). Second, sometimes participants would not issue commands to avoid a conflict until much later, when the aircraft were in close proximity of one another (late conflict recognition). Third, very infrequently, the participants would not notice a conflict until it was too late for it to be avoided and a separation conflict would occur (separation conflict). The case studies discussed below include two examples from each of the three types of conflict response.

The first two examples illustrate preventive planning. Both examples illustrate an aircraft taking off from an airport in the sector that will conflict with another aircraft unless the controller gives instructions to prevent it. Both participants issued commands to avoid the conflict to the departure aircraft before it appeared on the radarscope. The third and fourth examples illustrate conflict recognition. The participants recognized a potential conflict and issued commands to avoid it, but not until both aircraft were on the radarscope and within relatively close proximity to one another. The fifth and sixth examples illustrate separation conflicts and the moments leading up to them.

For all of the charts (Figures 3.9, 3.10, 3.11, 3.12, 3.13 and 3.14), purple boxes indicate simulator generated actions, red boxes indicate controller generated actions, and green boxes highlight important fixation patterns. Screenshots are included in the charts to describe the situation at particular times. Only AOIs relevant to the conflict situation were included. In order to illustrate the proper timeline, overlapping AOIs fixations were removed from the data files. In most cases, this did not affect the relevant fixes. In cases where two relevant AOIs overlapped, the experimenter examined the participant’s gaze replay to determine which AOI was probably being fixated upon. In the charts, horizontal lines indicate fixations; diagonal lines indicate saccades or no data between two fixations. The time is in seconds.

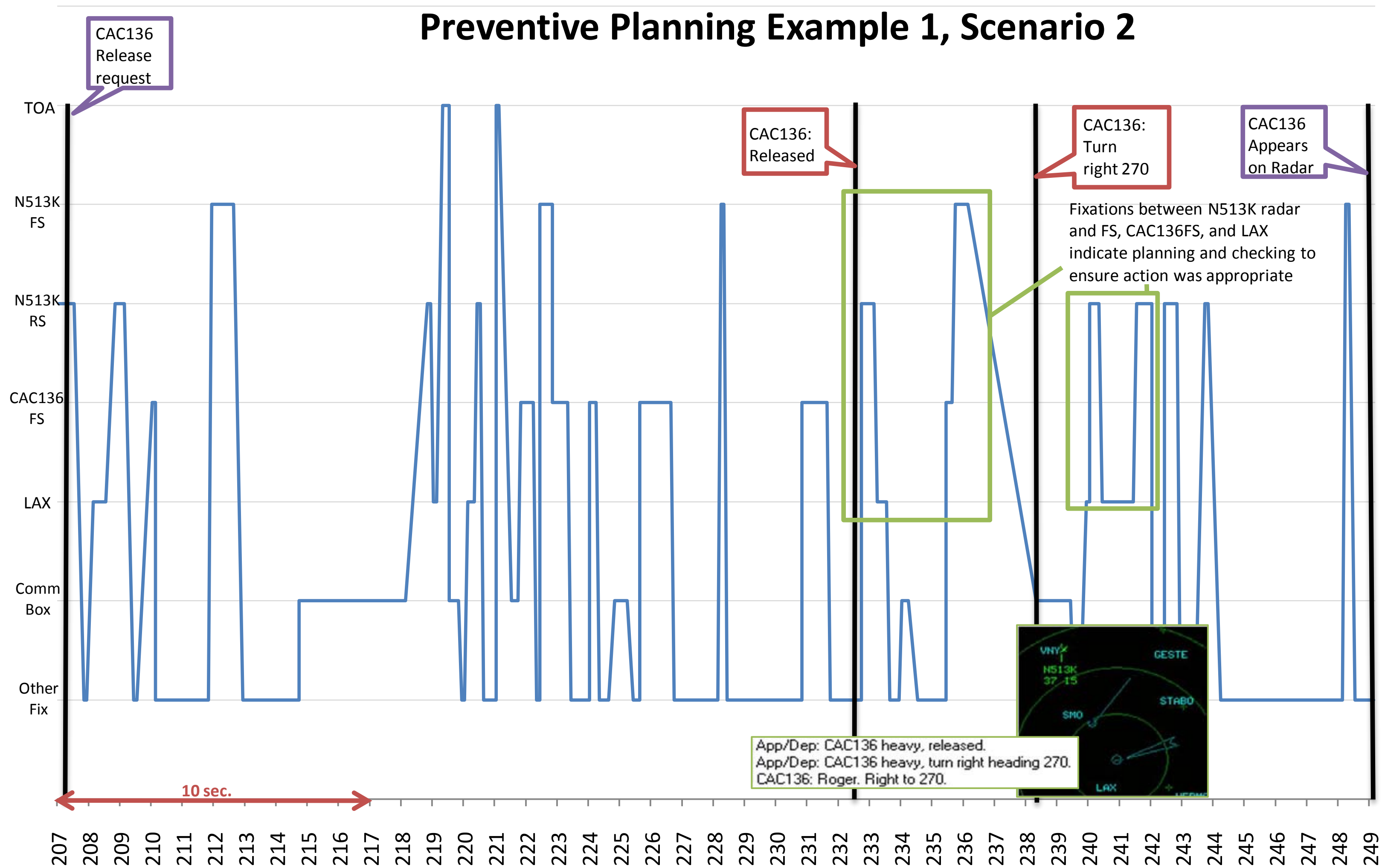


Figure 3.9. Preventive planning Example 1. The participant recognized a potential conflict prior as N513K moved across the airspace and CAC136 took off from LAX. The potential conflict was mitigated before CAC136 appeared on the radarscope.

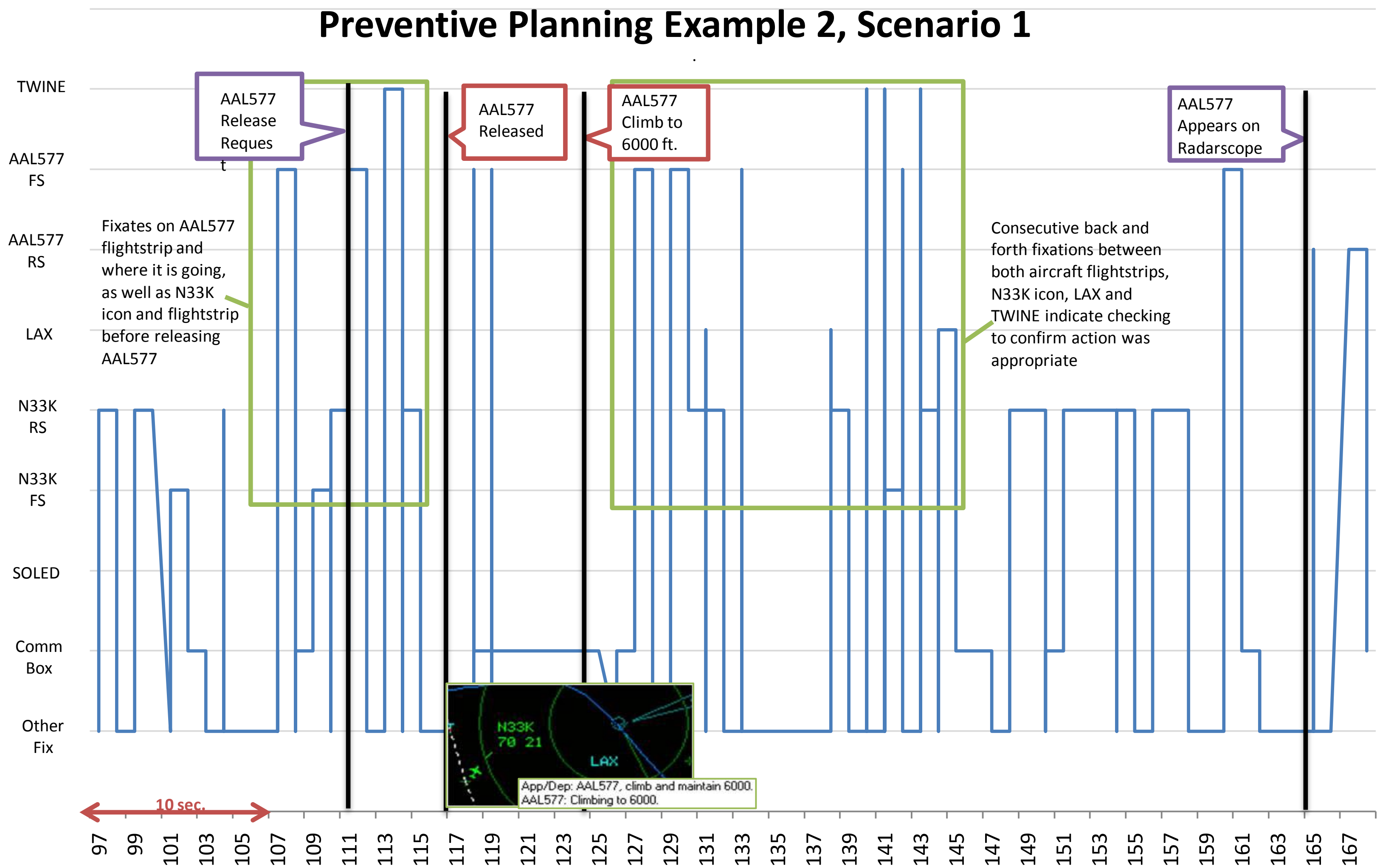


Figure 3.10. Preventive planning Example 2. N33K was crossing the airspace to SOLED. AAL577 was waiting for release from LAX. The two would have conflicted, but the participant issued an avoidance command to AAL577.

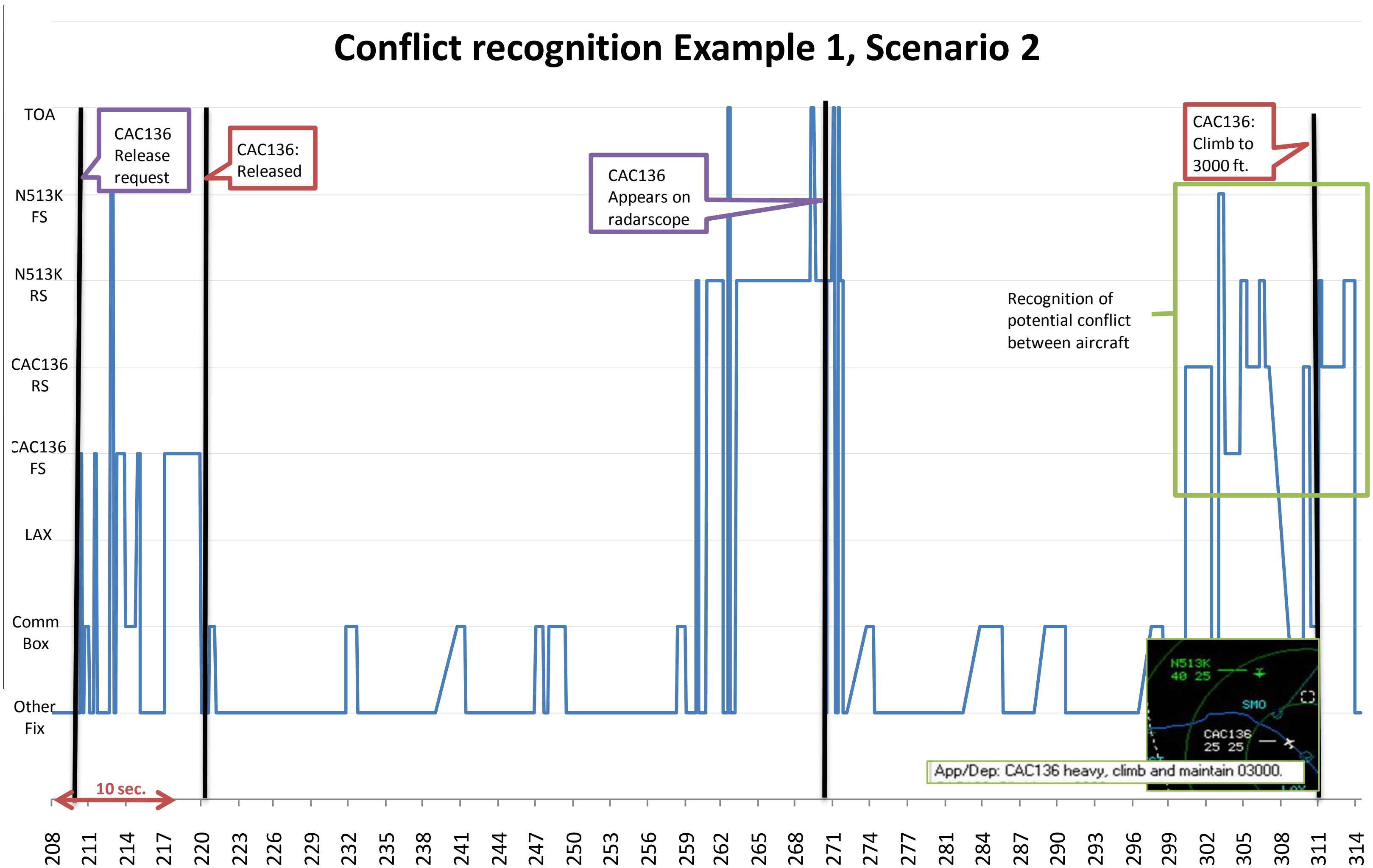


Figure 3.11. Conflict recognition Example 1. N513K moved towards TOA. CAC136 took off from LAX. The two aircraft may have conflicted, but the conflict was mitigated when CAC136 and N513K were in proximity of one another.

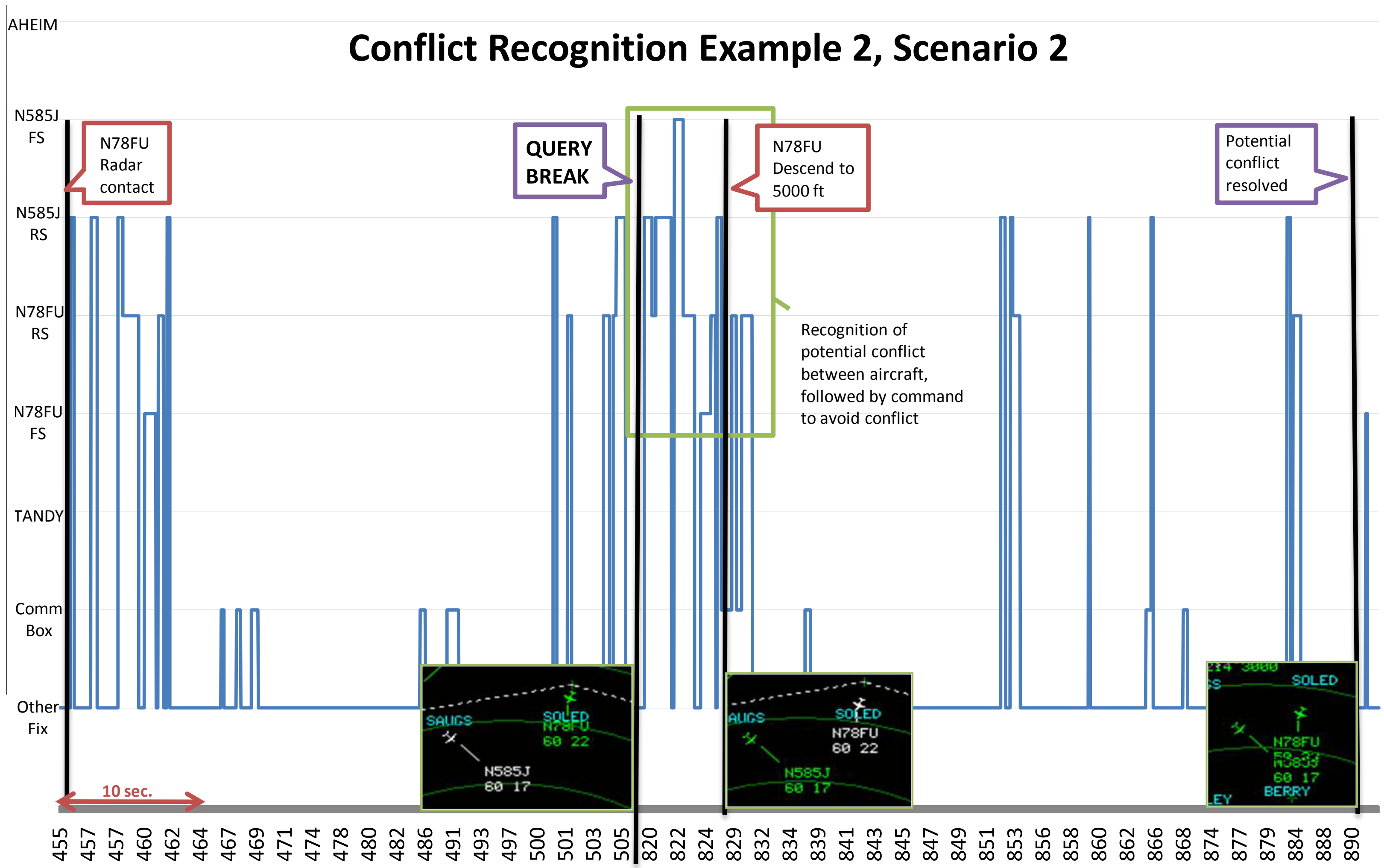


Figure 3.12. Conflict recognition Example 2. The participant recognized a potential conflict and issued a command to avoid it.

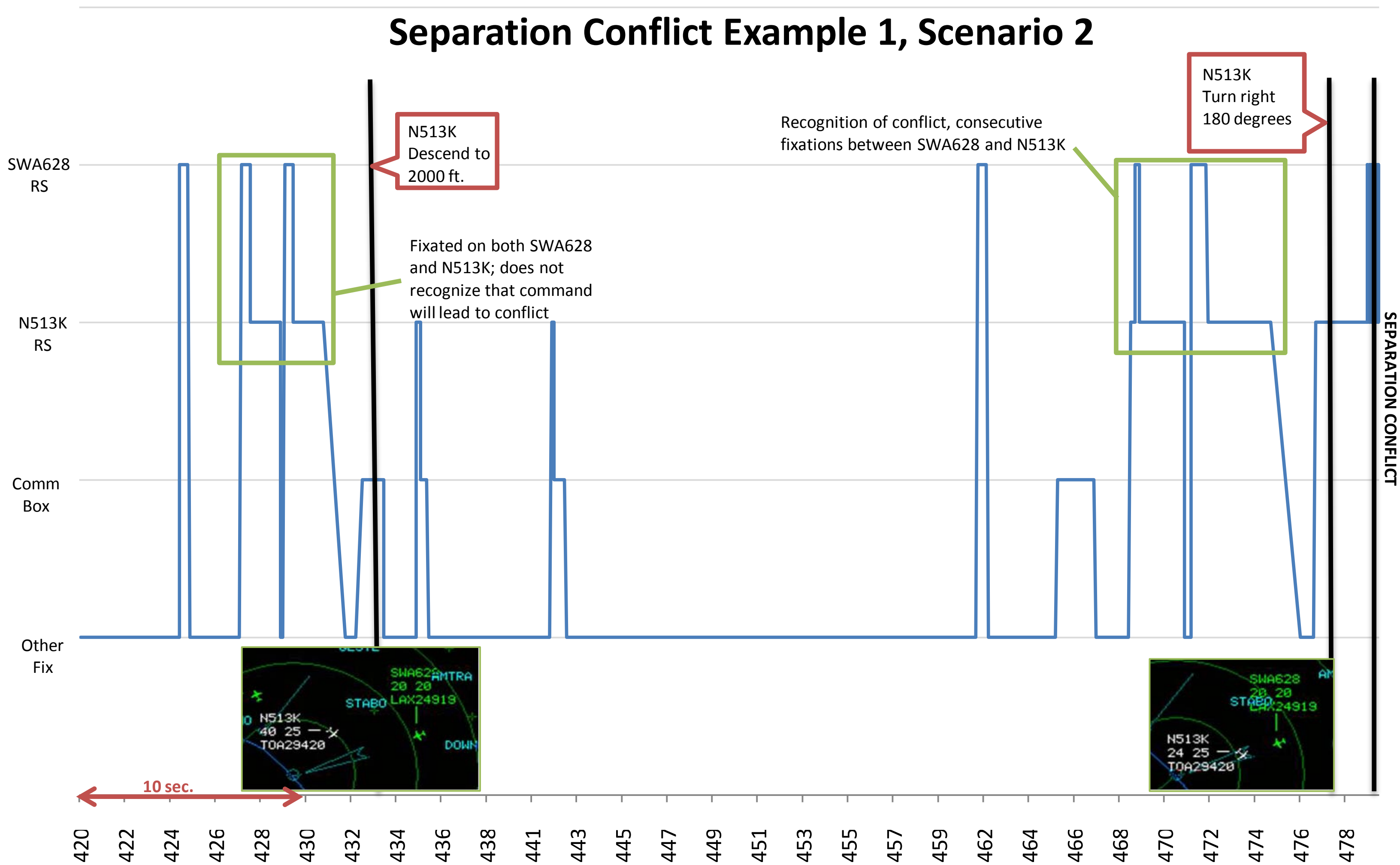


Figure 3.13. Separation conflict Example 1. N513K was headed to TOA to land. It was directed to traverse the airspace over LAX, where SWA628 was landing. The participant didn't issue a command to avoid the conflict until it was too late to avoid.

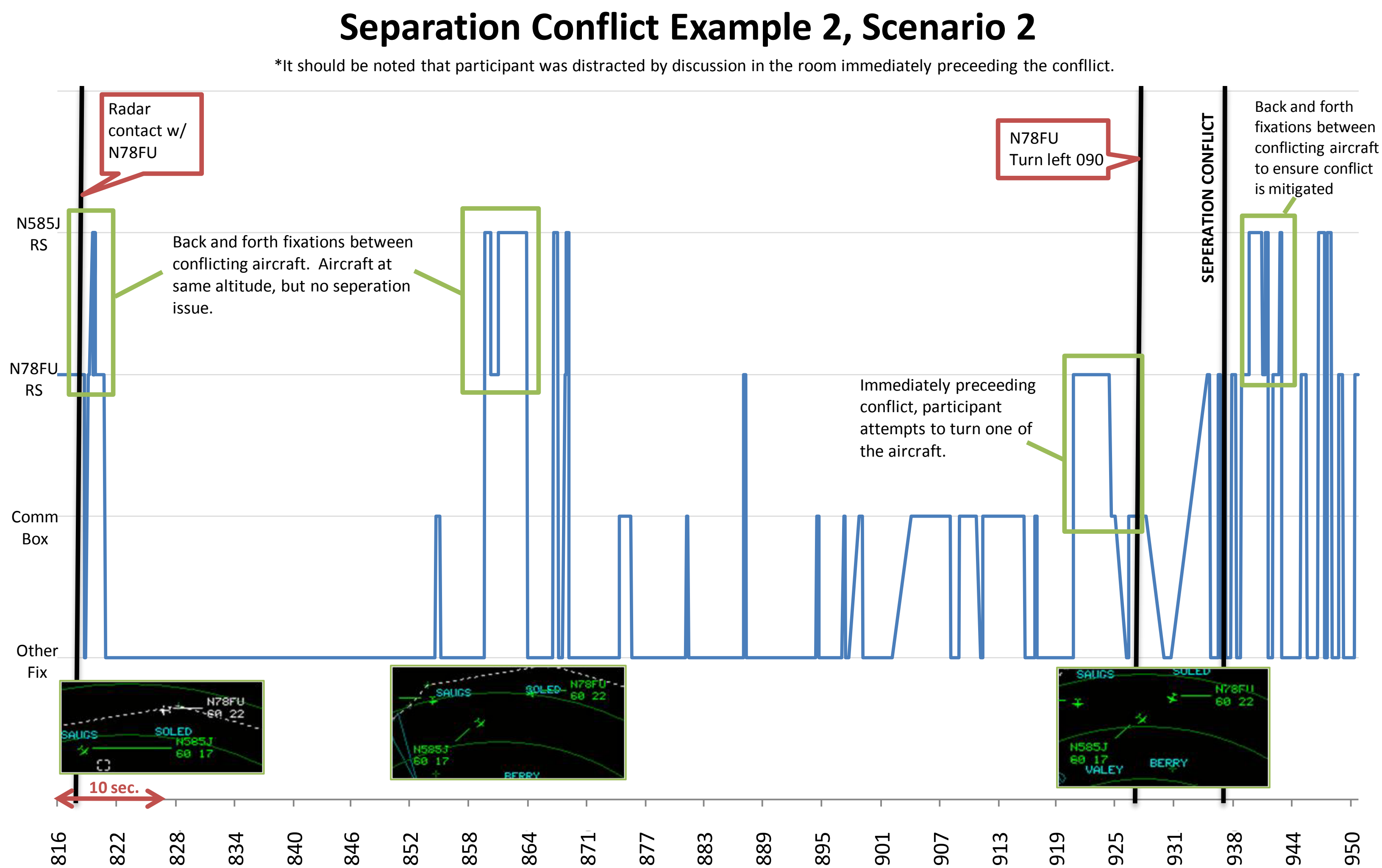


Figure 3.14. Separation conflict Example 2. N585J and N78FU were at the same altitude, headed towards one another. The participant did not issue a command to avoid the conflict until immediately preceding it.

The preventive planning examples highlight the most efficient way controllers avoided conflicts. In these examples, the majority of actions controllers took occurred as soon as they recognized two aircraft could potentially conflict. In Preventive Planning Example 1 (Figure 3.9), an aircraft with the call sign N513K is moving across the airspace towards Torrance airport (TOA) at 5000 ft. At time 207, a Los Angeles airport (LAX) tower controller asked the air traffic controller for permission to release an aircraft with the call sign CAC136. The flightpaths of the two aircraft intersect slightly northwest of LAX. In Figure 3.9, the y-axis is labeled with both aircraft radarscope (RS) icons and flightstrips (FS), the destination airport of N513K (TOA), the airport CAC136 is departing from (LAX) and the Communications Box (Comm Box). Therefore, CAC136 AOIs includes LAX and CAC136 FS and N513K AOIs include TOA, N513K FS, and N513K RS. The vertical black lines highlight important communications between the controller and the aircraft, as well as important actions of the aircraft on the radarscope or communications box. The timeline begins when CAC136 requested to be released from LAX and ends when it appeared on the radarscope. During the time between the CAC136 release request and when the controller granted the request (especially between 218 and 227 sec), the controller focused on the AOIs for both aircraft consecutively several times. This pattern of eye movements appears to indicate planning on the part of the controller. The fixations first go from N513K RS to where CAC136 will be departing from (LAX) to where N513K is headed (TOA). They then go from LAX back to N513K RS, to TOA, then CAC136 FS, then N513K FS and again back to CAC136 FS. There are several additional fixations on CAC136 FS before it is released.

Once CAC136 is released but before it has taken off from LAX (between 233 and 237 sec), the controller appears to check the problem area; he fixated on N513K RS, LAX (where CAC136 is awaiting takeoff), and then N513K FS before instructing CAC136 to turn after takeoff in order to avoid a conflict with N513K. After giving the turn command (between 239 to 243 sec), the controller appears to double check to ensure that it was the appropriate action; he fixated on N513K RS, LAX, and again on N513K RS. After confirmation, the controller turned his attention away from the AOIs before CAC136 takes off and appears on the radarscope.

In Preventive Planning Example 2 (Figure 3.10), an aircraft (call sign AAL577) is again released for takeoff from LAX and headed to the TWINE fix and another aircraft (call sign N33K) is crossing over the same airspace to the SOLED fix. AAL577 AOIs include AAL577 RS and FS, LAX (departure airport), and TWINE (destination fix). N33K AOIs include N33K RS and FS and SOLED (destination fix). The timeline begins 15 seconds before AAL577 requested to be released from LAX. Before the release request, the controller attended to N33K RS and FS, but also fixated on AAL577 FS. Once the release request was made, the controller fixated on AAL577 FS, TWINE, and then N33K RS before releasing AAL577. Almost immediately after AAL577 was released (but before it had taken off), the controller instructed AAL577 to climb to 6000 ft, 1000 ft lower than N33K. Once the command was issued (from 127 to 131 sec), the controller fixated on AAL577 FS, N33K RS and LAX; apparently to double check that the action was appropriate. He then fixated back and forth between both aircraft AOIs for eight seconds (between 138 and 145 sec). His attention then went to N33K RS and briefly to AAL577 FS before AAL577 appeared on the radarscope at 164 seconds.

Each of these Preventive Planning examples shows multiple instances of eye movements in close temporal proximity that connect the flightpaths of the two potentially conflicting aircraft. Notably, these connecting eye movements occur before the departing aircraft has taken off and entered the airspace. Possibly as a result of this early attention to the potential conflict, the controllers issued commands to the departing aircraft that result in the conflict being avoided while it was still on the ground.

Conflict Recognition Example 1 (Figure 3.11) is the same aircraft configuration as Preventive Planning Example 1. CAC136 requested to be released for takeoff from LAX at the beginning of the time line. The controller fixated several times on CAC136 FS and briefly on N513K FS before releasing CAC136. Once released, the controller did not fixate on CAC136 AOIs (CAC136 RS and FS, LAX) until 30 seconds after it appeared on the radarscope. Once the controller fixated on CAC136 RS (~301 sec) after this long delay, he seemed to recognize the conflict, as he then fixated on N513K FS, CAC136 FS, and then between N513K RS and CAC136 RS before a late command was issued to CAC136 to ascend to 3000 ft, 1000 ft lower than N513K. After the command, the controller fixated again on N513K RS, CAC136 RS and back to N513K RS, presumably to ensure that the command was executed.

Both the Preventive Planning Example 1 controller and the Conflict Recognition Example 1 controller fixated on CAC136 FS multiple times between the release request and release, but the Preventive Planning controller took around twice as long to release the same departure aircraft as the Conflict Recognition controller (22 sec compared to 11 sec). Unlike Conflict Recognition 1, Preventive Planning 1 appeared to look ahead to determine whether or not CAC136 would be in conflict with N513K before releasing it (between 218 and 224 sec) and issued a command to ensure there would be no conflict before CAC136 appeared on the screen. In contrast, the Conflict Recognition 1 controller did not fixate on CAC136 FS after it had been released until it appeared on the radarscope. Possibly because of this, the conflict avoidance command was given much later in Conflict Recognition 1.

Conflict Recognition Example 2 (Figure 3.12) took place over a longer timeline because the query break occurred in the middle. The query break lasted five minutes while the participant answered SA questions about the aircraft on the radarscope. Both aircraft (N585J and N78FU) were overflights entering the sector at the same altitude. If no action was taken, they would have conflicted. The time line begins when the controller made radar contact with N78FU. When contact was made, the controller fixated briefly on both aircraft (i.e., on N585J RS three times, then on N78FU RS and FS, then again on N585J RS). Around 40 seconds later, the controller again focused on both aircraft RS AOIs immediately before the query break. Once back from the break (five minutes later), the controller fixated between both aircraft RS and FS AOIs before instructing N78FU to descend to 5000 ft., 1000 ft. lower than N585J. Immediately after the command, the controller fixated several times on N78FU RS, likely to ensure that it was descending as instructed. N78FU did not complete its descent until 890 seconds; around one minute after the command was issued. The last screenshot illustrates the spatial proximity of the two aircraft when they reached appropriate separation. If the controller had waited 10 to 15 seconds longer to issue the command, a separation conflict would have likely occurred. In this example of late conflict recognition, the controller looked at the two aircraft only briefly when the second one entered the airspace, and then ignored them (visually) for 40 seconds until near the last possible moment when he could avoid a conflict.

Separation Conflict Example 1 (Figure 3.13) is interesting because the controller issued a command which directly led to the conflict. One aircraft's flightpath (call sign N513K) was directly over LAX airport, where another aircraft (call sign SWA628) was landing. N513K was at an appropriate flight level to avoid conflict (4,000 ft.), but the controller issued a command to descend to the same level as the SWA628 (2,000 ft.). Immediately preceding the command, the controller looked at both aircraft two times (between 427 and 431 seconds). After issuing the command, the controller does not look at the two aircraft together for 37 seconds, when (between 468 and 475 seconds) he recognized and attempted to correct the conflict. The first set of fixations was to SWA628 RS then N513K RS, a short fixation elsewhere, then again to SWA628 RS then N513K RS. The second set of fixations was to N513K RS, then SWA628 RS and back to N513K RS, a short fixation elsewhere, then again to N513K RS and then SWA628 RS. Interestingly, the controller's eye movements connecting the two aircraft early in the episode are very similar to the movements later on when he belatedly attempted to correct the conflict. The only differences appear to be that when the second set occurred, the aircraft were in closer spatial proximity and the fixations were considerably longer on

N513K RS. It is interesting that the controller fixated on both aircraft shortly before issuing the command that led to the conflict, indicating that he was at least aware of both aircraft and their positions at this early juncture. The controller's actions suggest that although he recognized the aircrafts' spatial proximity at this time, he did not realize that the command would lead to conflict. In other words, SA for the future movements of the two aircraft was low. It is also interesting that following the early command to N513K to descend, the controller fixated on it only very briefly, without also fixating the other aircraft. In the other examples, when a command was issued, the controllers appeared to 'follow-up' to ensure the command was executed and there were no additional conflicts by fixating on both aircraft.

Separation Conflict Example 2 occurred in the same scenario and involved the same aircraft as Conflict Recognition Example 2. Separation Conflict 2 (Figure 3.14) likely occurred due to inattention to the two conflicting aircraft (N585J and N78FU). The controller executed back and forth eye movements between N78FU RS and N585J RS when contact was made with N78FU, and again 40 seconds later. The controller fixated on N78FU RS only one additional time and did not fixate on N585J again in the time frame between the second set of back and forth movements and the conflict recognition, a time span of almost 50 seconds. After the conflict occurred, the controller fixated on both N585J RS and N78FU RS multiple times to ensure proper separation.

Separation Conflict 2 likely occurred due to inattention to the aircraft. If the controller recognized that both aircraft were at the same altitude when the back and forth eye movements occurred (between 816 and 820 seconds and again between 859 and 864 seconds), a command should have been issued to avoid the conflict. Unfortunately, one drawback of the eye movement data is that it alone is not enough to determine if the controller recognized the potential separation issue at the time the two aircraft were fixated upon or not. The conflict likely occurred due to external distraction to the controller. During this period of time, two individuals entered the data collection room and were speaking to one another. The controller was clearly distracted by the conversation, even though his eyes remained focused on the monitor. Both Separation Conflict 1 and 2 controllers recognized and attempted to correct the situations immediately before the conflicts, but were not able to avoid them.

When comparing Conflict Recognition 2 and Separation Conflict 2 (which involved the same aircraft), neither controller issued a command to avoid a conflict when radar contact was made, even though they looked at both aircraft consecutively back and forth. The Conflict Recognition controller issued a command the second time the aircraft were consecutively fixated upon, whereas the Separation Conflict controller did not.

When all of these case studies are compared, one thing that stands out is that the controllers in the Preventive Planning examples focused on the relevant AOIs more than the others and also had superior performance. Not only did the Preventive Planning controller fixate on the aircraft icons on the radarscope (RS's), they also fixated on the flightstrips (FS's) and the destination fixes of the aircraft. The controllers in the other examples rarely or never fixated on the flightstrips or destination fixes.

The case studies lend support to the ability to better understand how planning occurs using eye movements. The Preventive Planners' eye movements appear to show planning between the potentially conflicting aircraft flightstrips and their current and future locations on the radarscope. The Conflict Recognition 1 controller who avoided a conflict but waited until the aircraft appeared on the radarscope did not exhibit planning eye movements. Instead, the back and forth eye movements between the two aircraft on the radarscope immediately preceding a command indicate recognition of a potential conflict. The Conflict Recognition 2 controller did show a planning pattern of fixating on both aircrafts' RS and FS immediately before issuing the command to avoid the conflict, but conflict recognition took more time than in the preventive planning examples.

The case studies highlight that eye movements do not always clearly indicate what controllers were intending. Separation Conflict 1 shows that even when eye movements are very similar, they can lead to different outcomes. In addition, Separation Conflict 2 underscores how even short distractions can greatly affect concentration and performance, similar to the differences in performance seen in Study 1 between participants in the mobile phone and control conditions.

The next step in this case study analysis is to see if further examples from this study support this distinction between patterns of eye movements that do and do not suggest planning. If further support is found, then further steps would be to: 1. operationally define a pattern of eye movements that demonstrates planning; 2. systematically (i.e., not via case studies) identify every instance of this planning pattern in the data; 3. test whether planning eye movements are positively correlated with good control actions such as early conflict avoidance.

CHAPTER FOUR

GENERAL DISCUSSION

In recent years, situation awareness has become a catch-phrase in the media, used when human error occurs across numerous occupations and industries. In October, 2009, two pilots were using their laptops while on a domestic route and overflew their destination airport by 150 miles. A number of news outlets who reported the story stated in their reports that the pilots had "lost situational awareness." Researchers continue to study the construct of SA because of the potentially catastrophic circumstances that can arise when a loss of SA occurs. In this paper, SA is defined as knowledge of the current and expected future states of a situation.

While the theoretical underpinnings of SA continue to be the topic of debate, almost everyone can agree that understanding how to develop and maintain SA, as well as how losses of SA occur, are all relevant and necessary research areas. Measuring SA in real-world, operational environments is currently only able to be accomplished through observer reports or post-event questionnaires or interviews, because other more intrusive measures, such as online queries, would disrupt performance. A viable online SA measurement for operational environments needs to be developed in order to allow researchers the opportunity to continue to improve their knowledge and understanding of a still relatively uncharted construct. The current research results suggest that eye tracking may be employed in dynamic situations to measure SA.

The use of physiological methods such as eye tracking for measurement have only recently become feasible for a larger population of researchers due to greater accessibility and technological improvements. Physiological measures are now a credible option for measuring SA in real time situations. The two studies presented give support to the use of eye tracking as a measure of SA.

The current research explored the relationship between eye movements and direct measures of SA as a first step to determine how eye tracking can be used to measure SA in previously unexplored task domains. In both Study 1 and Study 2, SA was measured by interrupting operators (drivers or controllers) as they performed a real-time task in a simulator and querying them about task-relevant aspects of the preceding scenario. Thus our operational definition of SA in these studies was accuracy in answering the queries. Both studies showed that the more individuals fixated on an important, task-relevant event, the higher their SA for that event (as measured by accuracy of query responses). The studies also provide evidence that the way operators allocate attention (i.e., distributed widely or narrowly) affects their SA as well as their task performance. Finally, the studies showed positive correlations between SA and task performance.

In Study 1, participants who were distracted (in the mobile phone condition) had lower SA for hazardous events in driving scenarios compared with participants who were not distracted (in the no-phone condition). In the non-distraction condition, participants who spent a higher percentage of time fixating on the event were more accurate on SA questions about the event. Time spent fixating the event did not predict SA in the distraction condition. In terms of attention allocation strategy, one possibility is that participants who were not distracted had a wider distribution of fixations, which increased the likelihood of noticing and fixating on hazardous events compared with the distracted participants. Though not tested, this idea would be supported if non-distracted participants exhibited more fixations with shorter durations leading up to the event compared with the distracted participants.

The traffic level was also manipulated in Study 1. The number of fixations on the entire scene during an event decreased as SA increased for the low traffic level, and increased as SA increased for the high traffic level. Though not significant, there was a trend for mean fixation duration (for fixations anywhere in the scene) to increase as SA increased in the low traffic condition. Thus, in low traffic, as number of scene fixations decreased and scene fixation durations increased, SA improved. In other words, a narrow attention allocation strategy during a hazardous event improved SA for that event. It is unclear why the result was in the opposite direction for the high traffic condition, though it should be noted that the number of scene fixations in high traffic scenarios for both high and low SA were lower than for low traffic scenarios. The time period leading up to the event should be analyzed for both traffic levels to determine how participants' attention strategies affected performance on the SA queries.

Compared to Study 1, Study 2 scenarios were longer and were designed to allow participants time to develop SA before answering questions about information in the scene. Unlike Study 1, the eye movements for the entire time leading up to the SA queries, and sometimes for the entire scenario, were analyzed. This allowed for analysis of how participants' eye movements over extended periods affected SA, rather than simply how eye movements during an event affected SA for that specific event. In Study 2, a variety of eye movement measures predicted SA. A higher percentage of fixations on the queried aircraft AOIs and on flightstrips were associated with higher SA. In other words, fixating more on high-priority events, including the movements of important aircraft and the flightstrips, led to higher SA. These findings replicated the Study 1 finding that more time fixating on hazardous driving events led to higher SA for those events.

Study 2 also showed that a lower standard deviation of percentage fixations on aircraft was associated with higher SA. In other words, allocating attention widely across the aircraft led to higher SA. This finding supports the finding from Hauland's (2008) ATC study that radar controllers who distributed focal attention widely performed better than those who allocated attention narrowly.

In Study 2, several eye movement measures also predicted ATC performance. As number of scene fixations increased, ATC errors decreased. Since more fixations may be an indicator of a wider distribution of attention, this finding also shows the value of distributing attention widely during air traffic control. Larger NNI values were correlated with an increase in the number of errors. The NNI was initially used as a measure of workload. Camilli et al. (2008) found that as NNI neared 1 (i.e., fixations were more random), participant workload increased. In the context of the current study, as perceived workload increased the number of errors committed increased.

A higher percentage of fixations and longer fixation durations on queried aircraft and more time fixating on airports and fixes all led to more actions remaining after the scenario, a measure of controller efficiency. Overfocusing on high priority aircraft may increase number of actions remaining due to less attention to the other objects in the scenario. In addition, overfocusing on lower priority objects, such as airports and AOIs, may also reduce the efficiency of controllers. Regardless of object priority, an overfocusing (i.e., narrowed attention) strategy reduced controller efficiency.

The case studies were a first step in examining participant fixation patterns for different situations which arose during the scenarios. Though exploratory, the patterns seen in the case studies appear to lend support to the significant findings of Study 2. The case studies showed that eye movement patterns differed between participants in preventive planning situations, where conflicts were resolved early, compared with eye movements during late conflict resolutions and actual separation conflicts.

Study 2 results showed that fewer fixations on important aircraft AOIs, higher number of overall fixations, and smaller NNI values all led to higher performance. Participants in the preventive planning examples recognized and resolved conflicts quickly by scanning the flightpaths of the aircraft (proper attention allocation) and giving commands to avoid a potential conflict. By focusing attention narrowly on potential conflicts early and quickly resolving them, the amount of time spent fixating on the two aircraft over the course of the scenario was likely reduced because the participant no longer needed to monitor the potential conflict and ensure proper separation later in the scenario. This might explain why the high-performing preventive planners would have relatively few fixations on important aircraft. Once there was no possibility of conflict, the participant was able to attend to the rest of the aircraft in their airspace, using a wide distribution of attention that might lead to more overall fixations.

Participants in the late conflict recognition examples did not exhibit flightpath scanning and did not resolve the potential conflict until both aircraft were on the radarscope, requiring them to use a narrow attention strategy later in the scenario to ensure proper separation of the aircraft. Taken together, preventive planning participants would be more likely than late conflict resolution participants to have a higher number of total fixations (wide distribution of attention). Preventive planners would also be more likely to have smaller NNI values (lower perceived workload) than controllers who do not handle potential conflicts until they are close to one another. It is important to recognize that the case study analyses only considered a small portion of eye movements that occurred for that particular time frame; all eye movements for that time frame would need to be analyzed to draw more firm conclusions about the relationships between how participants handled conflicts, their attention strategies, and the results of Study 2.

Durso and Sethumadhavan (2008) explain that SA research is split into two lines.

> One line focuses on the product of SA, uses recall techniques, is domain specific, and uncovers that of which the operator is consciously aware. Another line focuses on the processes of SA; uses a variety of techniques… and uncovers the underlying mechanisms and processes… that allow an operator to understand the situation (Durso & Sethumadhavan, 2008, pg. 444).

The results of the two studies and the Study 2 case studies begin to piece together the perceptual and attentional processes that underlie the product of SA, which likely involves explicit knowledge maintaied in working memory. The underlying attention allocation strategies of participants appear to affect both SA and task performance. The participants in Study 2 were all trained air traffic controllers and both SA and task performance were high. However, Study 2 demonstrated that even with high levels of

SA and performance, differences in eye movements can predict both SA and performance.

The results of Study 1 and Study 2 also indicate that distraction affects SA in complex ways. In Study 1, participants in the mobile phone (distraction) condition spent a similar amount of time fixating on the hazardous events, regardless of whether they answered the SA queries correctly. This lends support to the "look-but-not-see" phenomenon. Though participants were not intentionally distracted in Study 2, when one participant was distracted unexpectedly, a separation conflict occurred. This type of distraction is an example of a momentary loss of SA. With only a short distraction, the participant did not recognize a conflict even though he/she continued to attend to the radarscope. The fine-grained analysis of this participant's eye movement data in one of the case studies allowed for a more intricate understanding of why this conflict occurred. The relationship between eye tracking, types of distraction, and SA should be examined in further experiments, as distraction appears to be a major contributing factor to a loss of SA.

There were several limitations in the current studies. The biggest limitation of Study 1 was that the analysis was performed on a data set from a previous experiment, completely removing experimental control. With the knowledge of the design and results from Study 2, Study 1's experimental design should be improved to include a more robust SA measure with additional queries related to what was occurring during each scenario. In addition, the scenarios should be redone to last longer than 30 seconds in order to allow SA to be built over time and the analysis should include the time frame leading up to the events.

Study 2's main limitation was the low variance in several of the measures. The future SA scores were out of only 3 possible points for each aircraft, and performance overall on that measure was very high. The task performance measure number of errors only ranged between 0 – 2. Though it is a positive indicator that certified air traffic controllers have high SA and low errors, a more normal distribution of scores would improve the hierarchical linear regression model fit.

Another limitation in the current studies was the amount of time it took to prepare the data for analysis. One key determination for researchers is the cost/benefit trade-off of eye tracking data when considering it as a viable measure of a psychological construct. The costs of research and development for the current studies included manually defining scenes and AOIs within scenes, determining and defining eye movement predictor variables, and breaking down eye movements within an individual participant's data set to detect patterns applicable to all participants, among others. Although this resulted in much time and research dollars dedicated to the data analysis, the benefits include a better understanding of the relationship between eye movements and SA and performance. In addition, while future researchers will still be required to consider the cost/benefit trade-off, the time it takes to go from raw eye movement data to analyzable variables may continue to decrease.

The current research results highlight that the relationship between eye movement measures and SA is complex and in need of further exploration. The eye movement measures chosen for analysis were only a selection from many potential analysis options. They were chosen based on their use in previous research studies, but additional variables likely play a role in an individual's SA development and performance. Further analyses on the current data set could include analyzing fixation durations in categories based on amount of time fixating (e.g., fixations less than 150 ms, fixations from 151 to 300 ms, etc.) as opposed to mean fixation duration, which ignores the distribution of fixations (e.g., Harris & Wiggins, 2008; Velichkovsky, Joos, Helmert, & Pannasch, 2005). Moreover, several eye movement variables could be included that have been shown to estimate cognitive requirements and workload in previous research, including pupil diameter, number of saccades, and duration of saccades (Ahlstroma & Friedman-Berg, 2006).

The results also underscore the drawbacks of eye tracking. The case studies illustrate how similar eye movements resulted in different outcomes. Even though participants' eye movements leading up to events could not always explain their actions, it is important to recognize that participants' probable reasoning for their choices could be identified in a majority of the case studies. One argument against eye tracking in the past has been that it can only illustrate where someone is looking and cannot determine comprehension or understanding. While it is true that comprehension is internal to the individual, the case studies conducted here suggest that, with further development, eye movements and actions of participants can give valuable insight into what was an individual was thinking and/or planning in a large number of situations. In addition, further analysis of eye movement patterns will lead to a deeper understanding of how planning develops and affects performance as well as identify common patterns of attention allocation in experts and others with high levels of SA. Understanding the components of successful planning would be beneficial to aid in training novices and improving scanning patterns of skilled operators.

Though additional research is needed to further validate the results, the current study's findings support the use of eye tracking as a measure of SA in situations where direct measures are not currently feasible. The analysis of eye movements in the current study was time intensive, but it is expected that the time it took to define AOIs and aggregate data for analysis could be greatly reduced in future studies through practice and by using ATC simulators that are better programmed to facilitate eye movement analyses.

Direct measures of SA are often criticized for only measuring SA at specific points in time and, in turn, ignoring the processes that occur leading up to and following SA measurement. Study 2 results were able to begin to examine the processes that affected performance on the SA queries and the importance of attention distribution on successful performance. Eye movements not only showed where a participant was looking, but also when participants were planning future actions and when they were not. Further refinement of the current analyses and results will influence the development of eye tracking measures of SA during actual on-the-job situations.

APPENDICES

Appendix A

Situation Awareness Queries

For Aircraft _____ (Call Sign:_________), indicate all of the following that you can recall:

1. Altitude (in feet): _______________
2. Groundspeed (in knots): _______________
3. Heading (between 0 - 360°): _______________
4. Circle One: Climbing? Descending? or Level?
5. Right turn? Left turn? or Straight?
6. Arrival? Departure? or Overflight?

7. If this aircraft is an arrival, at which airport will it be landing? _____________
   a. Is this aircraft currently at the correct landing altitude for its arriving airport? Y N

8. If this aircraft is a departure or an overflight, at which fix will it be leaving your sector? _______
   a. Is this aircraft currently at the correct altitude level for hand off? Y N

9. This aircraft is: *Higher* than *Lower* than the *Same Altitude* as Aircraft _____.
10. This aircraft is: *Faster* than *Slower* than the *Same Speed* as Aircraft _____.

Please answer the following questions regarding pending and active aircraft (Please use the number corresponding to the aircraft radar icon or pending flightstrip):

1. If applicable, list the pairs of aircraft that have currently lost separation or will lose separation if they stay on their current courses:

   _______ & _______; _______ & _______; _______ & _______; _______ & _______;

   _______ & _______; _______ & _______; _______ & _______; _______ & _______

2. If applicable, list the pairs of aircraft that would have lost separation had you not issued commands to adjust their courses:

   _______ & _______; _______ & _______; _______ & _______; _______ & _______;

   _______ & _______; _______ & _______; _______ & _______; _______ & _______

3. Of the aircraft currently in your ***pending*** flightstrips list:

   a. How many are arrivals or overflights? _______

      i. List all fixes the aircraft will be entering your sector at:

         ____

   b. How many are departures? _______

      i. List all airports the aircraft will be taking off from:

         ____

Please list your command priority level for each aircraft using the following priority levels:

1. High priority (*H*): ______ ______ ______ ______ ______ ______ ______ ______
2. Medium priority (*M*): ______ ______ ______ ______ ______ ______ ______ ______
3. Low priority (*L*): ______ ______ ______ ______ ______ ______ ______ ______

Appendix B

Demographic Questionnaire

1. How many years air traffic control experience do you have? _______ years

   a. How many years TRACON experience do you have? _______ years

   b. How many years has it been since you have worked TRACON control? _______ years

   c. How many years radar experience do you have? _______ years

   d. How many years has it been since you have used radar? _______ years

2. Are you familiar with Los Angeles (LAX) airport and its surrounding airspace (including VNY, SMO, TOA, & LGB airports)? Y N

   a. If yes, please explain why: ____

3. Have you previously used a TRACON computer simulator? Y N

   If you answered yes, please answer the following, if not proceed to Question 4.

   a. What was the name of the simulator? ________________________________

   b. When was the last time you used a TRACON computer simulator (Month/Year)?

   ____

   c. How proficient would you rate yourself on the TRACON computer simulator?

   Beginner Moderately Proficient Expert

4. About how many **hours per month** do you **currently** play **computer games or simulations** (also called video games)? _________ hours/month

5. For how many **years** have you played **computer games or simulations** (also called video games)? _________ years

6. How proficient would you rate yourself on **computer games or simulations** (also called video games)?

   No experience Beginner Moderately Proficient Expert

REFERENCES

Adams, M., Tenney, Y., & Pew, R. (1995). Situation awareness and the cognitive management of complex systems. *Human Factors, 37*(1), 185–104.

Ahlstrom, U. & Friedman-Berg, F.J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics, 36*, 623-636.

Balk, S.A., Moore, K.S, Steele, J.E., & Spearman, W. (2006). Mobile phone use in a driving simulation task: Differences in eye movements. *Presented as a poster presentation at the annual meeting of the Vision Sciences Society.*

Camilli, M., Nacchia, R., Terenzi, M., & Di Nocera, F. (2008). ASTEF: A simple tool for examining fixations. *Behavior Research Methods, 40*(2), 373-382.

Chapman, P.R. & Underwood, G. (1998). Visual search of driving situations: Danger and experience. *Perception, 27*(8) 951-964.

Cooke, N.J., Stout, R., & Salas, E. (2001). A knowledge elicitation approach to the measurement of team situation awareness. In M.D. McNeese, E. Salas, & M. Endsley (Eds.), *New Trends in Cooperative Activities: System Dynamics in Complex Settings (*pp. 114-139). Santa Monica, CA: Human Factors.

Crane, P.M. (1992). Theories of expertise as models for understanding situation awareness. In *Proceedings of the 13th Annual Symposium on Psychology in the Department of Defense* (pp. 148-52).

Crundall, D.E., Underwood, G., & Chapman, P.R. (1998). How much do novice drivers see? The effect of demand on visual search strategies in novice and experienced drivers. In G. Underwood (Ed.), *Eye Guidance in Reading and Scene Perception* (pp. 396-417). Amsterdam: Elsevier.

Dekker, S., & Hollnagel, E. (2004). Human factors and folk models. *Cognition, Technology, and Work, 6*(2), 79–86.

Durso, F.T., Bleckley, M.K., & Dattel, A.R. (2006). Does situation awareness add to the validity of cognitive tests? *Human Factors, 48*(4), 721-733.

Durso, F.T. & Gronlund, S.D. (1999). Situation awareness. In F. T. Durso, R. Nickerson, R. Schvaneveldt, S. Dumais, M. Chi, & S. Lindsay (Eds.), *The Handbook of Applied Cognition* (pp. 283-313). London: Wiley.

Durso, F.T., Hackworth, C.A., Truitt, T.R., Crutchfield, J., Nikolic, D., & Manning, C.A. (1998a). Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly, 6*(1), 1-20.

Durso, F. T., Rawson, K., & Girotto, S. (2007). Comprehension and situation awareness. In F. T. Durso, R. Nickerson, S. Dumais, S. Lewandowsky, & T. Perfect (Eds.), *Handbook of Applied Cognition (2nd) (*pp. 163-193), Chicester: Wiley.

Durso, F.T. & Sethumadhavan, A. (2008). Situation awareness: Understanding dynamic environments. *Human Factors, 50*(3), 442-448.

Durso, F.T., Truitt, T.R., Hackworth, C., Crutchfield, J., & Manning, C.A. (1998b). En route operational errors and situation awareness. *International Journal of Aviation Psychology, 8*(2), 177-194.

Durso, F. T., Truitt, T. R., Hackworth, C., Crutchfield, J., Nikolic, D., Moertl, P M., Ohrt, D. & Manning, C.A. (1995). Expertise and chess: A pilot study comparing situation awareness methodologies. In D. J. Garland & M. R. Endsley (Eds.), *Experimental Analysis and Measurement of Situation Awareness* (pp. 295–304). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Endsley, M.R. (l987). *SAGAT: A methodology for the measurement of situation awareness* (NOR DOC 87-83). Hawthorne, CA: Northrop Corp.

Endsley, M.R. (1988). Situation Awareness Global Assessment Technique (SAGAT). In *Proceedings of the National Aerospace and Electronics Conference* (pp. 789-795). New York: IEEE.

Endsley, M. R. (1990). Predictive utility of an objective measure of situation awareness. *Proceedings of the Human Factors Society, 34*, 41-45.

Endsley, M.R. (1995a). A taxonomy of situation awareness errors. In R. Fuller, N. Johnson, & N. McDonald (Eds.), *Human Factors in Aviation Operations* (pg. 287-292). Aldershot, England: Avebury Aviation, Ashgate Publishing Ltd.

Endsley, M.R. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors, 37*(1), 65-84.

Endsley, M.R. (1995c). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32-64.

Endsley, M.R. (2000). Direct measurement of situation awareness: Validity and use of SAGAT. In M.R. Endsley & D.L. Garland (Eds.), *Situation Awareness Analysis and Measurement* (pg. 147-175). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Endsley, M. R. & Jones, D. G. (1995). Situation awareness requirements analysis for TRACON air traffic control (TTU-IE-95-01). Lubbock, TX: Texas Tech University.

Endsley, M.R. & Rodgers, M.D. (1994). *Situation awareness information requirements analysis for en route air traffic control* (DOT/FAA/AM-94/27). Washington, DC: Federal Aviation Administration, Office of Aviation Medicine.

Endsley, M.R. & Rodgers, M.D. (1996). Attention distribution and situation awareness in air traffic control. In *Proceedings of the 40th annual meeting of the HFES,* 82-85.

Endsley, M.R., Selcon, S.J., Hardiman, T.D., & Croft, T.D. (1998). A comparative analysis of SAGAT and SART for evaluations of situation awareness. In *Proceedings of the 42nd Annual Meeting of the Human Factors and Ergonomics Society*.

Endsley, M.R., & Smolensky, M.W. (1998). Situation awareness in air traffic control: The picture. In M. Smolensky & E. Stein (Eds.), *Human Factors in Air Traffic Control* (pp. 115-154). New York: Academic Press.

Flach, J. (1995). Situation awareness: Proceed with caution, *Human Factors*, 37(1), 149-157.

Gaba, D.M., Howard, S.K., & Small, S.D. (1995). Situation awareness in anesthesiology. *Human Factors, 37*, 20-31.

GAO, Runway safety: Progress on reducing runway incursions impeded by leadership, technology, and other challenges, GAO-08-481T (Washington, D.C.: Feb. 13, 2008). Retrieved on October 26, 2009, http://www.gao.gov/new.items/d08481t.pdf.

Gorman, J.C., Cooke, N.J., & Winner, J.L. (2006). Measuring team situation awareness in decentralized command and control environments. *Ergonomics, 49*(12–13), 1312–1325.

Guan, Z., Lee, S., Cuddihy, E., & Ramey, J. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Gugerty, L.J. (1997). Situation awareness during driving: Explicit and implicit knowledge in dynamic spatial memory. *Journal of Experimental Psychology: Applied, 3,* 42-66.

Harris, J.M. & Wiggins, M.W. (2008). Evaluating cognitive competence: Does eye movement behavior represent the missing piece of the puzzle? In *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting*, 2077-2081.

Hauland, G. (2002). *Measuring team situation awareness in training of en route air traffic control: Process oriented measures for experimental studies* (Risø-R-1343[EN]). Risø National Laboratory. Retrieved September 15th, 2008, from http://www.risoe.dk/rispubl/SYS/ris-r-1343.htm

Hauland, G. (2008). Measuring individual and team situation awareness during planning tasks in training of en route air traffic control. *The International Journal of Aviation Psychology, 18*(3), 290-304.

Hogg, D. N., Folleso, K., Strand-Volden, F., & Torralba, B. (1995). Development of a situation awareness measure to evaluate advanced alarm systems in nuclear power plant control rooms, *Ergonomics, 11*, 2394-2413.

Horrey, W.J., Wickens, C.D., & Consalus, K.P. (2006). Modeling drivers' visual attention allocation while interacting with in-vehicle technologies. *Journal of Experimental Psychology: Applied, 12*(2), 67-78.

Horswill, M.S. and McKenna, F.P. (2004). Drivers' hazard perception ability: Situation awareness on the road. In S. Banbury and S. Tremblay (Eds.), *A Cognitive Approach to Situation Awareness*, pp. 155–175 (Aldershot, UK: Ashgate).

Hull, M. & Christie, R. (1992). Hazard Perception Test: The Geelong Trial and Future Development, paper presented to National Road Safety Seminar, Wellington, New Zealand, November.

Jacob, R.J., & Karn, S.K. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In J. H. Radach & H. Deubel (Eds.), *In the mind's eye: Cognitive and applied aspects of eye movement research* (pp. 573–605), Amsterdam: Elsevier Science.

Kaber, D.B., Endsley, M.R., Wright, M.C., Warren, H. (2002). The effects of levels of automation on performance, situation awareness, and workload in an advanced commercial aircraft flight simulation. Final Report: NASA Langley Research Center (Grant #NAG-1-01002). NASA Langley Research Center, Hampton, VA.

Kaber, D.B., Perry, C.M., Segall, N., McClernon, C.K., & Prinzel III, L.J. (2006). Situation awareness implications of adaptive automation for information processing in an air traffic control-related task. *International Journal of Industrial Ergonomic, 36*, 447-462.

Kintsch, W. (2005). An overview of top-down and bottom-up effects in comprehension: The CI perspective. *Discourse Processes, 39*(2 & 3), 125-128.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, U.K.: Cambridge University Press.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(2), 163-182.

Mecklinger, A., Kramer, A.F., & Strayer, D.L. (1992). Event related potentials and EEG components in a semantic memory search task. *Psychophysiology, 29*(1), 104-119.

Mertler, C.A. & Vannatta, R.A. (2005). Advanced and multivariate statistical methods: Practical application and interpretation (3rd Edition). Pyrczak Publishing, Glendale, CA.

Neisser, U. (1976). *Cognition and reality: Principles and implications* of *cognitive psychology.* San Francisco: W.H. Freeman.

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

Osborne, J.W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research, and Evaluation, 7*(1), Available online at http://pareonline.net/getvn.asp?v=7&n=1.

Press, M. (1986). *Situation awareness: Let's get serious about the clue-bird.* Unpublished manuscript.

Prince, C., Ellis, E., Brannick, M.T., & Salas, E. (1997). Measurement of team situation awareness in low experience level aviators. *The International Journal of Aviation Psychology, 17*(1), 41-57.

Rahimi, M., Briggs, R.P., & Thom, D.R. (1990). A field evaluation of driver eye and head movement strategies toward environmental targets and distracters. *Applied Ergonomics, 21*(4), 267-274.

Recartes, M.A. & Nunes, L.M. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Psychology: Applied, 6*(1), 31-43.

Recartes, M.A. & Nunes, L.M. (2003). Mental workload while driving: Effects on visual search, discrimination, and decision making. *Journal of Experimental Psychology: Applied, 9*(2), 119–137.

Salmon, P.M., Stanton, N.A., Walker, G.H., Green, D. (2006). Situation awareness measurement: A review of applicability for C4i environments. *Applied Ergonomics, 37*, 225-238.

Salmon, P.M., Walker, G.H., Ladva, D., Stanton, N.A., Jenkins, D.P., & Rafferty, L. (2007). Measuring situation awareness in command and control: Comparison of methods study. *Proceedings of the ECCE 2007 Conference*, London, UK, 27-34.

Salmon, P.M., Stanton, N.A., Walker, G.H., Jenkins, D., Ladva, D., Rafferty, L., & Young, M. (2008a). Measuring situation awareness in complex systems: Comparison of measures study. *International Journal of Industrial Ergonomics*. Corrected proof available online December 6, 2008.

Salmon, P.M., Stanton, N.A., Walker, G.H., Baber, C., Jenkins, D.P., McMaster, R., Young, M.S. (2008b). What really is going on? Review of situation awareness models for individuals and teams. *Theoretical Issues in Ergonomics Science, 9*(4), 297–323.

Sarter, N. B., & Woods, D. D. (1991) Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology, 1,* 45-57.

Sarter, N.B., & Woods, D.D. (1995). How in the world did I ever get into that mode? Mode error and awareness in supervisory control. *Human Factors*, *37*(1), 5-19.

Smith, K. & Hancock, P.A. (1995). Situation awareness is adaptive, externally directed consciousness. *Human Factors, 37*(1), 137–148.

Stanton, N.A., Stewart, R., Harris, D., Houghton, R.J., Baber, C., McMaster, R., Salmon, P., Hoyle, G., Walker, G., Young, M.S., Linsell, M., Dymott, R., & Green, D. (2006). Distributed situation awareness in dynamic systems: Theoretical development and application of an ergonomics methodology. *Ergonomics*, *49*(12-13), 1288-1311.

Strayer, D.L., Cooper, J.M., & Drews, F.A. (2004). What do drivers fail to see when conversing on a cell phone? In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting*, 2213-2217.

Strayer, D.L., Drews, F.A., & Johnston, W.A. (2003). Cell phone-induced failures of visual attention during simulated driving. *Journal of Experimental Psychology: Applied*, *9*(1), 23-32.

Strybel, T.Z., Vu, K.L., Kraft, J., & Minakata, K. (2008). Assessing the situation awareness of pilots engaged in self spacing. In *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting*, 11-15.

Taylor, R.M. (1990). Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design. In *Situational Awareness in Aerospace Operations (AGARD-CP-478*, pp. 3/1 –3/17). Copenhagen, Denmark: NATO-Advisory Group for Aerospace Research and Development.

Tenney, Y. J., & Pew, R. W. (2006). Situation awareness catches on: What? So What? Now What? *Reviews of the Human Factors and Ergonomics Society, 1,* 1-35.

Tobii Eye Tracker user manual (2006). Tobii Technology AB.

TRACON II multi-player air traffic control simulator (1990). Austin, TX: Wesson International.

Velichkovsky, B.M., Joos, M., Helmert, J.R., & Pannasch, S. (2005). Two visual systems and their eye movements: Evidence from static and dynamic scene perception. *Proceedings of the XXVII Conference of the Cognitive Science Society,* 2283-2288.

Vidulich, M.A. (2000). The relationship between mental workload and situation awareness. In *Proceedings of the IEA 2000/HFES 2000 Congress,* 3-460 - 3-463.

Vortac, O.U., Edwards, M.B., Fuller, D.K., & Manning, C.A. (1993). Automation and cognition in air traffic control: An empirical investigation. *Applied Cognitive Psychology, 7,* 631-651.

Walker, N. & Fisk, A.D. (1995). Human factors goes to the gridiron. *Ergonomics in Design, 3*(3), 8-13.

Wheater, C.P. & Cook, P.A. (2000). *Using statistics to understand the environment*. New York, NY: Routledge.

Wickens, C. D. (1996). Situation awareness: Impact of automation and display technology. In *Situation Awareness: Limitations and Enhancement in the Aviation* Environment, Brussels: AGARD Conference Proceedings, pp. k2.1-k2.13.

Wickens, C.D. (2008). Situation awareness: Review of Mica Endsley's 1995 articles on situation awareness theory and measurement. *Human Factors, 50*(3), 397-403.

Wickens, C.D., McCarley, J.S., Alexander, A.L., Thomas, L.C., Ambinder, M., & Zheng, S. (2005). Attention-situation awareness (A-SA) model of pilot error. Aviation Human Factors Division Institute of Aviation Technical Report (AHFD-04-15/NASA-04-5).

Williams, T.R., Mulligan, C., Koprowicz, K., Miller, J., Reimann, C., Wang, D. (2005). Does isolating a visual element call attention to it? Results of an eye-tracking investigation of the effects of isolation on emphasis. *Technical Communication, 52*(1), 21-26.

Woodhouse, R., & Woodhouse, R.A. (1995). Navigation errors in relation to controlled flight into terrain (CFIT) accidents. *8th International Symposium on Aviation Psychology,* Columbus, OH.