

**DOT/FAA/CT- 05/32**

Federal Aviation Administration  
William J. Hughes Technical Center  
Atlantic City International Airport, NJ 08405

# **Subjective Workload Ratings and Eye Movement Activity Measures**

Ulf Ahlstrom, Ph.D., AJP-7132  
Ferne Friedman-Berg, Ph.D., L-3 Communications Titan Corporation

December 2005

Technical Report

This document is available to the public through the National Technical Information Service (NTIS), Springfield, Virginia 22161. A copy is retained for reference by the William J. Hughes Technical Center IRC.



**U.S. Department of Transportation  
Federal Aviation Administration**

## NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof. The United States Government does not endorse products or manufacturers. Trade or manufacturer's names appear herein solely because they are considered essential to the objective of this report. This document does not constitute FAA certification policy. Consult your local FAA aircraft certification office as to its use.

This report is available at the Federal Aviation Administration William J. Hughes Technical Center's full-text technical reports web site:  
<http://actlibrary.tc.faa.gov> in Adobe Acrobat portable document format (PDF).

**Technical Report Documentation Page**

<b>1. Report No.</b> DOT/FAA/CT-05/32		<b>2. Government Accession No.</b>		<b>3. Recipient's Catalog No.</b>	
<b>4. Title and Subtitle</b> Subjective Workload Ratings and Eye Movement Activity Measures				<b>5. Report Date</b> December 2005	
				<b>6. Performing Organization Code</b> AJP-7132	
<b>7. Author(s)</b> Ulf Ahlstrom, Ph.D., AJP-7132 Ferne Friedman-Berg, Ph.D., L-3 Communications, Titan Corporation				<b>8. Performing Organization Report No.</b> DOT/FAA/CT-05/32	
<b>9. Performing Organization Name and Address</b> Federal Aviation Administration NAS Human Factors Group, AJP-7132 William J. Hughes Technical Center Atlantic City International Airport, NJ 08405				<b>10. Work Unit No. (TRAIS)</b>	
				<b>11. Contract or Grant No.</b>	
<b>12. Sponsoring Agency Name and Address</b> Federal Aviation Administration Chief Scientist for Human Factors, ATO-P 800 Independence Ave., S.W. Washington, DC 20591				<b>13. Type of Report and Period Covered</b> Technical Report	
				<b>14. Sponsoring Agency Code</b> ATO-P	
<b>15. Supplementary Notes</b>					
<b>16. Abstract</b> <p>In the present study, we evaluated the possibility of using eye movement activity measures as a correlate of cognitive workload. Using data from a high-fidelity human-in-the-loop weather simulation, we explored eye activity measures like pupil diameter, blink duration, and saccade distance, and assessed their relationship to subjective workload ratings. In our initial analysis, we established that although there was no significant effect of weather tool use on subjective workload ratings, there was a significant relationship between subjective workload ratings and our task load variable aircraft density. We found a linear increase in workload ratings with an increasing number of aircraft in the sector. In a subsequent analysis, we assessed the relationship between eye movement activity measures and aircraft density. We found that the mean blink duration and the mean saccade distance decreased as aircraft density increased, while the mean pupil diameter increased with an increasing number of aircraft in the sector. After establishing the relationship between these eye activity metrics and subjective workload, we evaluated whether we could use changes in eye movement activity along with other system state variables, like distance to weather from the outer marker, to measure ongoing controller workload. We developed both individual controller models and a general model (across controllers) to assess whether it was possible to predict the minute-by-minute number of aircraft in the sector. Using both multiple regression modeling and neural network models, we were able to produce individual controller models and general models with good prediction performances. We discuss possible applications for these findings in future air traffic control (ATC) research, in adaptive automation, and in ATC interface design.</p>					
<b>17. Key Words</b> Air Traffic Control Eye movements Workload				<b>18. Distribution Statement</b> This document is available to the public through the National Technical Information Service, Springfield, Virginia, 22161	
<b>19. Security Classif. (of this report)</b> Unclassified		<b>20. Security Classif. (of this page)</b> Unclassified		<b>21. No. of Pages</b> 37	<b>22. Price</b>



## Table of Contents

	Page
Executive Summary .....	v
1. Introduction.....	1
2. Method .....	2
2.1 Participants.....	2
2.2 Apparatus .....	2
2.3 GENERA TRACON Airspace.....	3
2.4 Traffic Scenarios .....	4
2.5 Air Traffic Standard Operating Procedures (SOP) .....	4
2.6 Advanced Weather and Precipitation Information .....	4
2.7 Simulation Design.....	5
2.8 Independent Variables .....	5
2.9 Procedures and Dependent Variables .....	7
3. Results.....	8
3.1 Data Sets .....	8
3.2 ATWIT Ratings .....	8
3.3 Controller Strategies for Regulation of Workload.....	9
3.4 ATWIT and Target Density .....	11
3.5 Eye Movement Measures.....	12
3.5.1 Blink Frequency and Blink Duration.....	12
3.5.2 Saccade Frequency and Saccade Distance.....	13
3.5.3 Pupil Diameter .....	15
4. Modeling.....	16
4.1 Multiple Linear Regression Modeling.....	16
4.2 Artificial Neural Network (ANN) Modeling .....	19
5. Discussion.....	24
References.....	27
Acronyms.....	31

## List of Illustrations

Figures	Page
Figure 1. The GENERA TRACON airspace .....	3
Figure 2. An illustration of precipitation levels and ITWS weather data used in the simulation.....	5
Figure 3. The WIDS and TCW configuration used during the simulation.....	6
Figure 4. Illustrations of the storm motion for WS 1 and WS 2.....	7
Figure 5a. Mean ATWIT ratings for WS 1 as a function of simulation condition and probe interval. ....	9
Figure 5b. Mean ATWIT ratings for WS 2 as a function of simulation condition and probe interval. ....	9
Figure 6. The mean distance to the closest Level 4 weather cell from the outer marker .....	10
Figure 7. Mean $z$ score values for distance to weather (36L), mean number of aircraft, and mean .....	11
Figure 8. Mean ATWIT ratings as a function of the number of aircraft in the sector.....	12

Figure 9. Mean blink frequency as a function of the number of aircraft in the sector .....	13
Figure 10. Mean blink duration as a function of the number of aircraft in the sector. ....	13
Figure 11. Mean saccade frequency as a function of the number of aircraft in the sector .....	14
Figure 12. Mean saccade distance and mean distance between aircraft as a function of the number of aircraft in the sector.....	15
Figure 13. Mean relative pupil diameter as a function of the number of aircraft in the sector ....	16
Figure 14. Individual low-performance model for observed and predicted number of aircraft as a function of scenario time .....	17
Figure 15. Individual high-performance model for observed and predicted number of aircraft as a function of scenario time .....	18
Figure 16. A general model for the prediction of the number of aircraft as a function of scenario time .....	19
Figure 17. An example of a standard Artificial Neural Network (ANN) model with backpropogation.....	20
Figure 18. Individual low-performance model for observed and predicted number of aircraft ...	21
Figure 19. Individual high-performance model for observed and predicted number of aircraft ..	21
Figure 20. A general model for observed and predicted number of aircraft for 25 exemplars ....	22
 Table	 Page
Table 1. Results of the Performance Comparison of ANN and Regression Models (RM) .....	23

## Executive Summary

Air traffic controllers respond to system demands generated by sector traffic, communications, and the overall air traffic control (ATC) environment. As the system's task load increases, there can be an increased demand on controller performance resulting in increased controller workload. Because workload levels can affect controller performance, researchers have spent a great deal of effort developing reliable measures of controller workload. One common measurement technique is to use self-reported workload ratings generated by controllers as they perform ATC tasks. Although these methods have proven to generate reliable workload estimates, there is a limit to their applicability in capturing real-time changes in workload levels. Therefore, as a supplement to subjective workload ratings, researchers have explored the use of eye movement parameters that correlate with cognitive demands. The most commonly used metrics are blink rate and duration, pupil diameter, saccadic extent, fixation frequency, and dwell time. By using real-time eye movement measures to evaluate workload, we could potentially detect changes in workload and identify what the controller was looking at when these changes occur.

In the present study, we explored the relationship between subjective workload ratings and eye activity measures from a high-fidelity human-in-the-loop weather simulation (Ahlstrom & Friedman-Berg, 2005). First, we assessed the relationship between subjective workload ratings and the task load variable, aircraft density. We found that workload ratings increased linearly with an increasing number of aircraft. Second, we assessed how eye activity measures of saccades, blinks, and pupil diameter correlated with aircraft density. We found that the mean blink duration and the mean saccade distance decreased as a function of an increased number of aircraft, while the mean pupil diameter increased as a function of an increased number of aircraft. Third, we explored the possibility of using this eye data to predict the minute-by-minute number of aircraft in the sector by means of multiple regression and neural network models. Using regression modeling, we were able to produce individual controller models ( $R$  ranging from .19 to .89) and a general model ( $R = .58$ ) with good mean prediction performances. Using neural network models, we produced individual controller models ( $r = .26$  to  $r = .92$ ) and a general model ( $r = .84$ ) with even greater mean prediction performances.

These results indicate the possibility of using real-time workload estimates derived from eye movement activity measures for the development and evaluation of ATC systems. For instance, we might use eye movement metrics to implement ATC systems that incorporate adaptive automation. The purpose of adaptive automation is to automate certain tasks when workload increases past a certain level, returning these tasks to the operator as workload decreases. Other potential operational uses are the development of workload watch systems to determine when to adjust staffing, when to divide sectors, or how to distribute tasks among controller teams. Real-time metrics would also be useful when developing new ATC procedures and ATC tool interfaces. The use of real-time workload measures would be particularly useful for the evaluation of displays or procedures in instances when the evaluation of controller performance does not detect any differences.

## 1. Introduction

Workers in many technical environments face increasing workload as system complexity increases. Many times, system operators perform tasks driven solely by system demands, leaving little opportunity for self-paced performance (Vicente, 1999). One example is air traffic control (ATC) operators performing under time and system event pressure in their complex sociotechnical domain (Ahlstrom, 2004; Stein, 1985). In the ATC domain, air traffic controllers must be constantly prepared to handle fluctuating system demands.

Even when system activity does not overburden operators, frequent fluctuations in system task load may create stress or impact vigilance, which in turn can have a negative impact on operator performance. All of these things may create potential operating hazards. Many different factors impact system demands and thus workload in the air traffic domain, including weather, sector traffic, communications, control actions, and the overall ATC environment. As the ATC system task load increases, there can be increasing demand on controller performance resulting in increased controller workload (Stein, 1985).

Because workload can have an immediate impact on operator performance, there are many potential uses for real-time measures of workload. For example, if we could accurately measure workload in real-time, we could create systems that produced an optimal level of workload or systems that could detect when workload exceeded a certain threshold (Schvaneveldt, Reid, Gomez, & Rice, 1998). We could also create systems that determined when workload was too low, using times of inactivity to schedule less critical tasks or deliver noncritical messages (Iqbal, Adamczyk, Zheng, & Bailey, 2005). By optimizing workload, we could potentially reduce controller stress, increase controller performance, and increase system safety. For instance, one technique for optimizing workload is adaptive automation (Parasuraman & Hancock, 2001). In adaptive automation, automation of select system activities occurs as workload increases, with control given back to the system operator as workload decreases. However, when system developers incorporate adaptive automation, they will need a method for measuring workload in real time.

Researchers have put a great deal of effort into developing measures and probes of controller workload to mitigate the impact of both high and low workload levels on controller performance and to reduce operating hazards. One commonly used measure in ATC research is the Air Traffic Workload Input Technique (ATWIT) (Stein, 1985, 1991). The ATWIT is a method that allows researchers to obtain an unobtrusive, unidimensional measure of workload as controllers perform ATC tasks. Using the ATWIT scale, controllers in high-fidelity human-in-the-loop simulations indicate their instantaneous subjective workload level by pressing buttons labeled from 1 (low workload) to 10 (high workload). Researches have used this method to measure workload in studies evaluating multi-sector control positions (Willems, Heiney, & Sollenberger, 2005), decision support tools (Sollenberger, Willems, Della Rocco, Koros, & Truitt, 2004), voice communications latency (Sollenberger, McAnulty, & Kerns, 2003), and ATC complexity (Yuditsky, Sollenberger, Della Rocco, Friedman-Berg, & Manning, 2002). However, this method does not provide researchers with a continuous, real-time measure of ongoing controller workload.

Some researchers have used eye activity measures that correlate with cognitive demands to measure real-time workload. For example, researchers have used metrics like blink rate and duration, dwell time, fixation frequency, pupil diameter, and saccadic extent to estimate



workload (Brookings, Wilson, & Swain, 1996; Van Orden, Jung, & Makeig, 2000; Van Orden, Limbert, Makeig, & Jung, 2001; Wilson, 2001; Wilson & Caldwell, 2002). Findings indicate that blink rate, blink duration, and saccade duration all decreased while pupil diameter, the number of saccades, and the frequency of long fixations all increased with increased workload (Iqbal, Adamczyk, Zheng, & Bailey, 2004, 2005; Iqbal, Zheng, & Bailey, 2004; Lin, Zhang, & Watson, 2003; Rognin, Grimaud, Hoffman, & Zeghal, 2004; Stein, 1992; Van Orden, 2000; Van Orden et al., 2000; Veltman & Gaillard, 1998; Zeghal, Grimaud, Hoffman, & Rognin, 2002). If we could validate these eye movement metrics in an ATC human-in-the-loop simulation, they would have the potential to provide the type of real-time measure we need.

In this study, we attempted to validate the relationship between subjective ATWIT ratings and eye activity measures from an ATC weather simulation (Ahlstrom & Friedman-Berg, 2005). First, we evaluated the relationship between ATWIT ratings and task load as measured by increases in aircraft density. We then examined how eye activity metrics like saccades, blinks, and pupil diameter, correlated with the task load variable. Because this data came from a previous study evaluating the impact of weather information on controller performance (Ahlstrom & Friedman-Berg), we attempted to use both eye activity metrics and a measure of weather proximity (to the outer marker) to develop multiple regression and neural network models to predict ongoing workload, as measured by the minute-by-minute sector aircraft count. We conclude this paper by discussing possible applications for these findings in ATC research, and possible issues with developing real-time models.

## 2. Method

### 2.1 Participants

Eleven full-performance level Terminal Radar Approach Control (TRACON) controllers participated in the study (mean job experience = 12 years). All controllers held a current medical certificate.

### 2.2 Apparatus

We conducted this simulation at the Federal Aviation Administration (FAA), William J. Hughes Technical Center's Research Development and Human Factors Laboratory (RDHFL). The simulation configuration consisted of the Distributed Environment for Simulation, Rapid Engineering, and Experimentation (DESIREE) and the Target Generation Facility (TGF). DESIREE emulates the Standard Terminal Automation Replacement System (STARS) functionality and receives input from the TGF for displaying radar targets. An Integrated Terminal Weather System (ITWS) simulator provided pre-recorded ITWS data to DESIREE for presentation on the TRACON controller workstation (TCW) or an auxiliary weather information display system (WIDS) (Ahlstrom, Keen, & Mieskolainen, 2004).

Two controllers operated sector traffic during each simulation run. One controller was responsible for West operations, while the other was responsible for East operations. The controllers issued commands to simulation pilots who maneuvered aircraft using keyboard commands and communicated with the controllers using appropriate ATC phraseology and procedures.

The West side controllers wore an Applied Science Laboratories (ASL) Model 5000 oculometer consisting of an eye and head tracking system. Prior to each simulation run, researchers calibrated the oculometer using a nine dot calibration grid (Willems et al., 2005). The system

measured both eye and head movement at 60 Hz to record points of gaze (POG) in x, y, and z coordinates relative to the scene plane. From the POG data, we were able to calculate fixation metrics, saccades metrics, blink metrics, and pupil diameter (Ahlstrom & Friedman-Berg, 2005, or Willems, Allen, & Stein, 1999).

A SUN workstation and a 10-button keypad collected and recorded controller responses. The controllers indicated their instantaneous workload by pressing one of the keypad buttons labeled from 1 (low workload - all tasks completed) to 10 (high workload - some tasks left uncompleted). The system prompted controllers for input every five minutes by emitting several beeps and lighting the buttons on the keypad. Controllers had 20 seconds to respond by pressing the button corresponding to their current workload level. If the controller gave no response within 20 seconds, ATWIT defaulted to a digit indicating that there was no response.

### 2.3 GENERA TRACON Airspace

Figure 1 shows an illustration of the GENERA TRACON airspace. The GENERA airspace extended for approximately 70 nm from north to south, and approximately 60 nm from west to east. Four en route sectors (GENERA Center), each with a separate arrival fix, surrounded the GENERA airspace. The primary arrival flows to Runway 36L came from the south and northwest arrival fixes. The primary arrival flows to Runway 36R came from the southeast and the northeast arrival fixes. Runway 5 was available for instrument approaches as needed. Runways 36L and 36R were also used for departures during the simulation, but the departure sector worked independently of the GENERA arrival sectors.

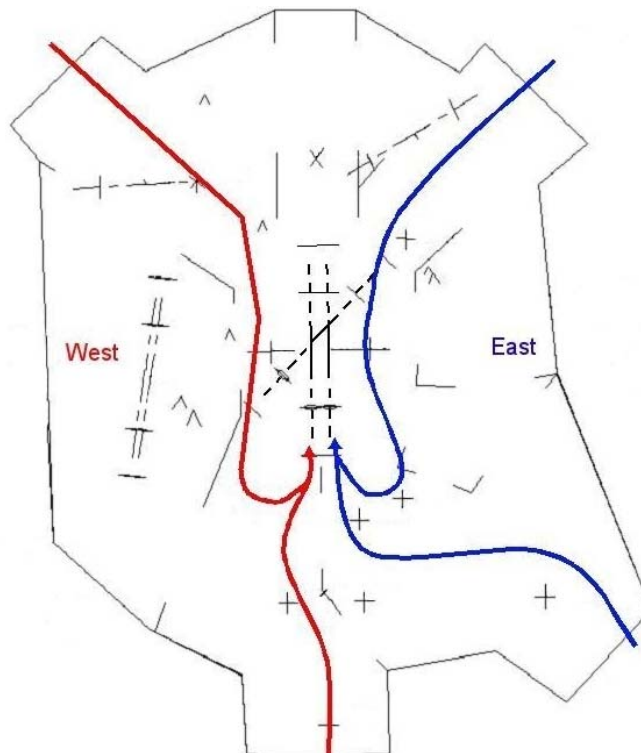


Figure 1. The GENERA TRACON airspace with arrival routes and Runways 36L, 36R, and 5.

## 2.4 Traffic Scenarios

Subject matter experts created six 50-minute traffic scenarios with moderate traffic levels. Each scenario had an equal number of arrivals and departures. The practice scenarios included 51 aircraft and test scenarios included 89 aircraft.

## 2.5 Air Traffic Standard Operating Procedures (SOP)

We used SOP developed for the simulation airspace. These procedures dictate operational responsibilities like separation, data entries, emergencies, handoffs, coordination, separation on final approach, and approaches to satellite airports to mention a few. Most importantly, we included an experimental procedure that assigned responsibility for keeping aircraft away from weather Levels 4, 5, and 6 to the controller.

## 2.6 Advanced Weather and Precipitation Information

During all simulation runs, the controllers had access to six levels of precipitation information. An illustration of the precipitation and ITWS information is shown in Figure 2. Precipitation Levels 1-3 were coded 'blue' and Levels 4-6 'brown'. Sparse stipples represented Level 2 ('blue') and 5 ('brown'). Dense stipples represented Level 3 ('blue') and Level 6 ('brown'). Level 1-2 represent light to moderate precipitation intensities with a possible light to moderate turbulence. Levels 3-6 represent moderate to extreme precipitation intensities with a possibility of severe turbulence. Pilots will generally avoid penetrations of precipitation Level 3 (i.e., 41 dBZ) or higher, and controllers begin to anticipate pilot requests for deviations when the weather approaches Level 3 (Rhoda & Pawlak, 1999).

The ITWS weather data provided graphical information about the location and movement of storm cells and gust fronts. It also provided information about the location and area of wind shear and microburst. Unfilled red circles indicated wind shear and semi-filled red circles indicated microburst. The white arrows denoted storm cell motion. Current storm cell position was depicted in magenta, and extrapolated positions (10 and 20 minutes) were shown as dotted magenta lines. A solid pink line defined current gust front position, and dotted pink lines defined extrapolated positions (10 and 20 minutes).

In addition to the ITWS information, we also used prototype displays of animated storm forecasts. These animations showed the current storm position and forecasted positions 15 and 30 minutes into the future.



Figure 2. An illustration of precipitation levels and ITWS weather data used in the simulation.

## 2.7 Simulation Design

We used a 3 (display location of weather information) x 2 (weather scenario) x 2 (sector) repeated measures design. Display location of weather information and weather scenario were within-subjects variables, while sector was a between-subjects variable. Therefore, all participants controlled traffic in all three display conditions for both weather scenarios for a total of six simulation runs. We counterbalanced the presentation order of these six simulation conditions by means of a randomized block design.

## 2.8 Independent Variables

The first independent variable was display location of advanced weather information that resulted in the following three conditions:

1. In the WIDS condition, we displayed weather information on the WIDS display located on top of the TCW (see Figure 3).
2. In the TCW condition, we displayed weather information directly on the TCW. The WIDS display was not used during this condition.
3. In the Control condition, we did not present any weather information to the controller. The WIDS display was not used during this condition. This condition represents current TRACON operations in the field.



Figure 3. The WIDS and TCW configuration used during the simulation (WIDS top display, TCW bottom display).

The second independent variable was weather scenario. We used two weather scenarios (Weather Scenario [WS] 1 and 2) of prerecorded ITWS data during the simulation. Both WS 1 and 2 contained ITWS weather information (i.e., storm motion, gust front, wind shear, and microburst information) but differed in the overall spatial distribution and temporal characteristics. Figure 4 shows an illustration of the storm motion for WS 1 and 2 superimposed on the GENERA airspace. WS 1 contained 15 wind shear alerts and WS 2 contained six alerts during the 50-minute scenarios. There were four microburst alerts in WS 1, whereas WS 2 only contained one microburst alert at the end of the scenario.

Our third independent variable was sector position, which was a between-subjects variable. We used two sector positions during the simulation:

1. GENERA West
2. GENERA East

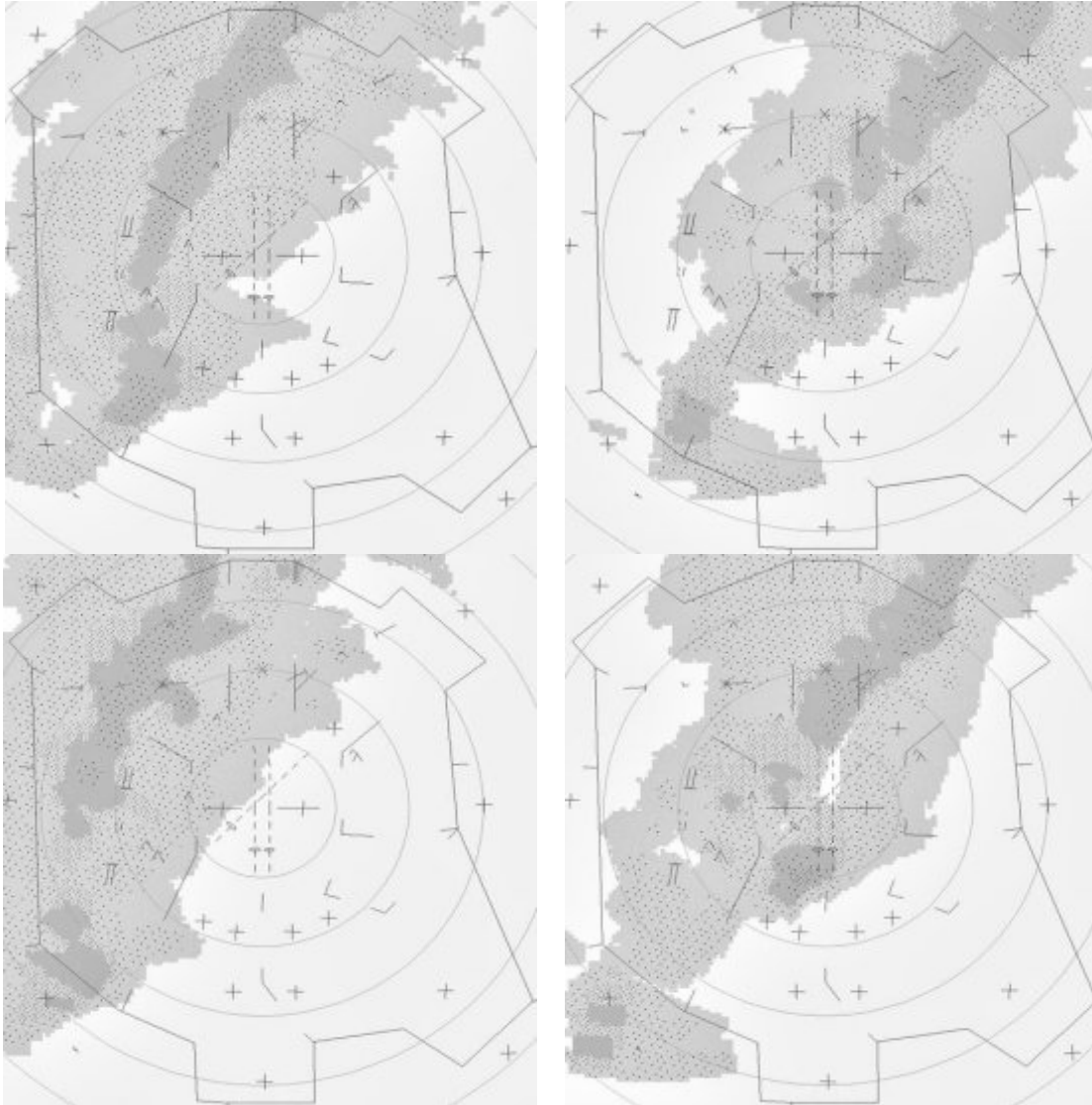


Figure 4. Illustrations of the storm motion for WS 1 and WS 2 (WS 1 top row, WS 2 bottom row). The left column shows the storm location at one minute into the scenario (WS 1 top left, WS 2 bottom left). The right column shows the storm location at the end (50 minutes) of the scenario (WS 1 top right, WS 2 bottom right).

## 2.9 Procedures and Dependent Variables

The controllers received one day of training on the GENERA airspace and the use of advanced weather information. During this training, controllers worked traffic in two 30-minute scenarios while using advanced weather information. We did not use any of these scenarios in the simulation.

During simulation scenarios, controllers worked traffic during adverse weather conditions. An automated data collection system recorded system operations and generated a set of standard ATC simulation dependent measures that included measures of safety, capacity, efficiency, severe weather avoidance, communications, and the use of weather tools. For instance, we

recorded the number of aircraft that penetrated precipitation Levels 4-6, the number of separation errors and wake turbulence violations, the number of completed flights, the number and duration of holds, the number of control commands (i.e., altitude, heading, and speed), and the number of handoffs. We also recorded the number and duration of push-to-talk communications and all instances of weather tool use. These results are reported elsewhere (Ahlstrom & Friedman-Berg, 2005).

For the current analysis, the relevant dependent measures included the ATWIT workload ratings made throughout the scenario by controllers and the eye movement metrics collected with the oculometer. When recording ATWIT ratings, researchers have sometimes replaced instances of nonresponses with a workload rating of 10 (Sollenberger, La Due, Carver, & Heinze, 1997), assuming that the controller was too busy to respond (i.e., high workload). In other instances, researchers have treated this data as missing data (Sollenberger et al., 2003). For the present study, we omitted all instances of these nonresponses prior to the analyses. Our recorded eye movement metrics included blink frequency and blink duration, saccade frequency and saccade duration, and pupil diameter.

### 3. Results

#### 3.1 Data Sets

We collected 10 ATWIT ratings for each controller during each simulation run. Due to technical problems and track losses, we only obtained 29 oculometer data sets from the 36 simulation runs. Three participants had data from all six conditions, and the other three participants had data from two, three, and four conditions, respectively. Of the 29 sets of data, 27 were from the entire 50 minute run, one was from the first 12 minutes of the scenario, and one was from the first 28 minutes of the scenario.

For each simulation condition, we determined the number of aircraft (target density) under the control of a particular controller for each minute of the 50 minute scenario. We used target density as our task load variable, with increasing target density corresponding to increasing workload (Van Orden et al., 2001). Across conditions and controllers, target densities per minute ranged from 1 to 11. We computed averages of the following eye activity metrics for each minute of the scenario: blink frequency and duration, saccade frequency and distance, and pupil diameter. We then computed an average for the eye activity metrics and the ATWIT ratings for the different levels of target density. Because we did not directly manipulate target density, we did not obtain complete data sets from each subject for each condition. However, by computing an average over target densities for all six conditions, we were able to obtain complete data sets from each subject for target densities 2 through 9.

#### 3.2 ATWIT Ratings

We computed the mean ATWIT ratings over subjects for each five minute interval in the three simulation conditions for both WS 1 and WS 2 (see Figures 5a and 5b). It is clear that in all conditions, workload ratings were low for both WS1 and WS2. We used a repeated measures analysis of variance (ANOVA) to test for an effect of scenario and tools on workload ratings over time. We found no significant ( $p < .05$ ) two-way (scenario x tool) or three-way (time x scenario x tool) interactions, indicating that when evaluating ATWIT response patterns over time, the ratings did not differ significantly for the different combinations of scenarios and tools.

Therefore, we collapsed the data over controllers, weather scenarios (WS 1 and 2), and simulation conditions (WIDS, TCW, and Control) for the present analysis.

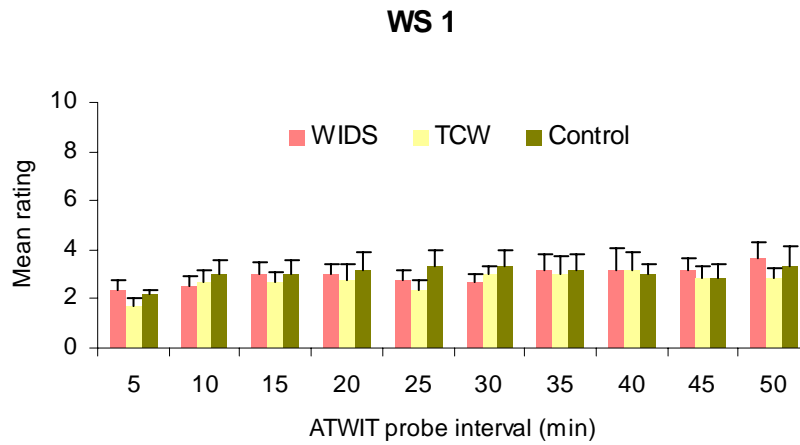


Figure 5a. Mean ATWIT ratings for WS 1 as a function of simulation condition and probe interval. Error bars are standard errors (*SE*).

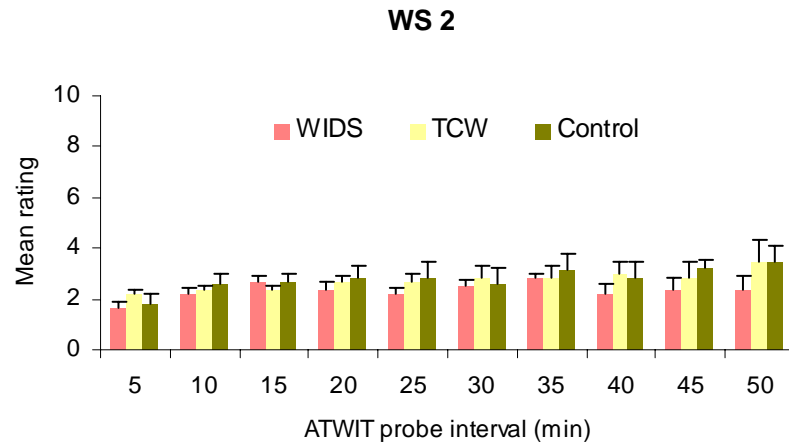


Figure 5b. Mean ATWIT ratings for WS 2 as a function of simulation condition and probe interval. Error bars are *SE*.

### 3.3 Controller Strategies for Regulation of Workload

Controllers employ many different strategies for regulating their workload as task load increases (Ahlstrom, 2004; Vicente, 1999). They may choose to decrease the time spent processing each aircraft, decrease communications with pilots, stop or offload less critical tasks, increase spacing, or prevent aircraft from entering their sector (Parasuraman & Hancock, 2001).

In this simulation, weather appeared to play a role in how controllers regulated workload. Controllers seemed to adjust the aircraft flow rate to conform to the weather situation. First, we looked at how the weather situation changed over the course of the scenario. We measured the distance to the outer marker because of its critical location in relation to the runways and its importance for pilots during approach. As the scenario progressed, the weather cells approached the outer marker for Runway 36L. Figure 6 graphically depicts the mean distance in relative



airspace units (pixels) to the closest Level 4 weather cell from Runway 36L’s outer marker as a function of simulation time. We measured the distance from the Level 4 weather cells because higher Levels (5 and 6) always appeared within Level 4 cell areas in our weather scenarios. As can be seen in Figure 6, the distance to weather from Runway 36L decreased as a function of simulation time,  $R^2 = .94$ ,  $SE = 7.78$ ,  $F(1,49) = 753.27$ ,  $p < .001$ .

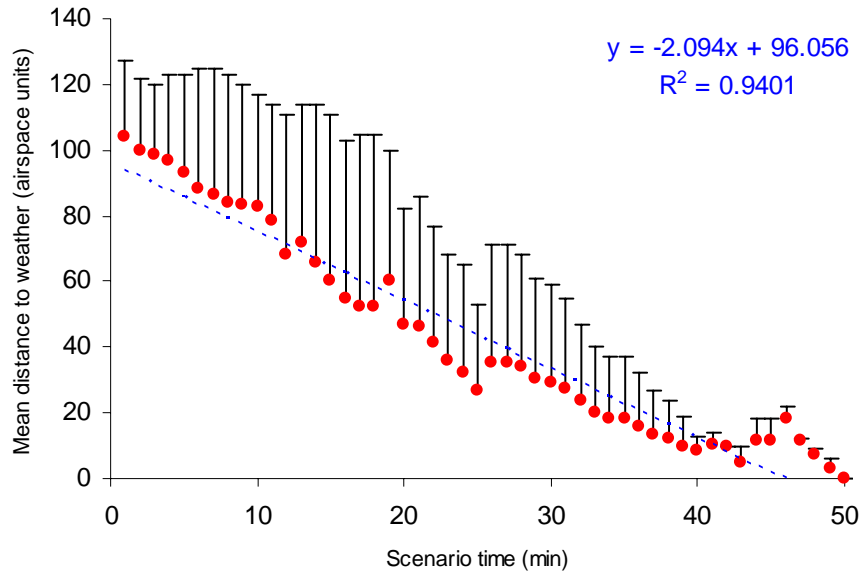


Figure 6. The mean distance to the closest Level 4 weather cell from the outer marker (Runway 36L). The distance measure is the mean distance from WS 1 and WS 2. The bars are SE.

Next, we examined how changing weather patterns influenced aircraft arrivals over the course of the scenarios. During the simulation, standard operating procedures allowed controllers to have aircraft penetrate Level 1-3 weather cells, but controllers had specific instructions to avoid aircraft penetrations of Level 4-6 cells (severe weather avoidance). The presence of Level 4 weather cells in close proximity to or over the outer marker of a runway, prevented aircraft from approaching that runway. Therefore, when Level 4-6 weather cells were blocking the approach to both sector runways, controllers stopped accepting aircraft into the sector. Controllers also regulated traffic by reducing arrivals to the runway blocked by Level 4-6 weather cells, funneling traffic to the nonblocked runway. As weather cells moved away and left a clear path to a runway, controllers began to accept arrivals, with the rate of acceptance influenced by the current weather pattern.

By looking at the number of aircraft in the sector at a given time, along with ATWIT ratings and distance to weather, we can evaluate weather’s influence on how controllers regulated workload over the course of the scenario. Here, we graphically illustrate the relationship between these three variables over time. First, we normalized these variables by converting them to  $z$  scores, where the  $z$  score indicates how much and in what direction a score deviates from the mean score as expressed in standard units. The standardized  $z$  scores have a mean of zero and a standard deviation of one. Figure 7 depicts the  $z$  scores for the mean distance to weather, the mean number of aircraft, and the mean ATWIT ratings as a function of simulation time.

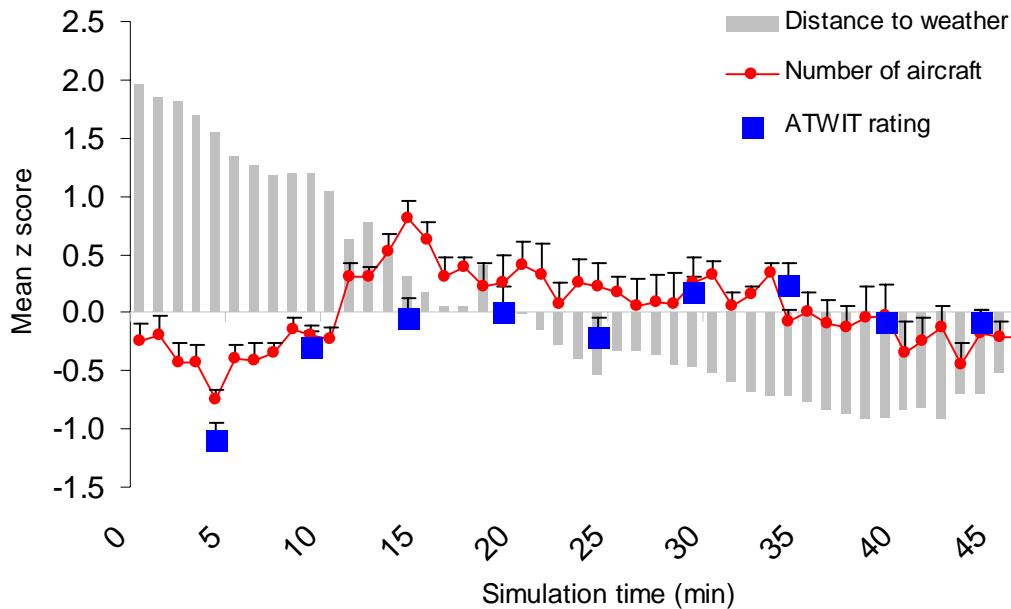


Figure 7. Mean z score values for distance to weather (36L), mean number of aircraft, and mean ATWIT ratings as a function of simulation time. Error bars are SE.

We can see that at the beginning of the simulation, the Level 4 cells were farther from the outer marker ( $|z \text{ score}| > 1.5$ ) than they were later in the simulation ( $|z \text{ score}| < 1.5$ ). Also at this time, both the number of aircraft in the sector and the ATWIT ratings were low. At 12 to 15 minutes into the scenario, the number of aircraft increased and the weather moved closer to the outer marker for Runway 36L. After 15 minutes, controllers seemed to begin reducing the number of arrivals, probably due to approaching weather cells. Because controllers seemed to reduce the number of aircraft in the sector to a manageable level after the 15 minute mark, the ATWIT ratings also were low and remained so up to the 30 minute mark. At the 30 and 35 minute marks, there was a slight increase in workload ratings at the same time that there were two peaks in the aircraft count. At the 35 minute mark, controllers again began to decrease the number of aircraft in the sector. At this point, weather cells were still close to the outer marker for Runway 36L and Runway 5 (distance to Runway 5 not shown). We hypothesize that because weather cells at the end of the scenario were moving eastward and were impacting both Runways 36L and 5, controllers found it increasingly difficult to engage in severe weather avoidance. This may have led to an increase in subjective workload at the end of the scenario, as reflected in the increase in ATWIT ratings.

### 3.4 ATWIT and Target Density

In the present analysis, we established the relationship between subjective ATWIT ratings and an objective measure of task load. Van Orden et al. (2001) found that target density functioned well as a task load variable. Therefore, we examined the ATWIT ratings and their relationship to the number of aircraft in the sector at the time of the rating.

Figure 8 plots the mean ATWIT ratings against the number of aircraft in the sector, and shows a clear increase in the ATWIT ratings as the number of aircraft increased. We performed a linear regression analysis, regressing the ATWIT ratings on the number of aircraft, and found that the

ATWIT ratings increased linearly with an increasing number of aircraft,  $R^2 = .75$ ,  $SE = .20$ ,  $F(1, 6) = 17.97$ ,  $p = .005$ .

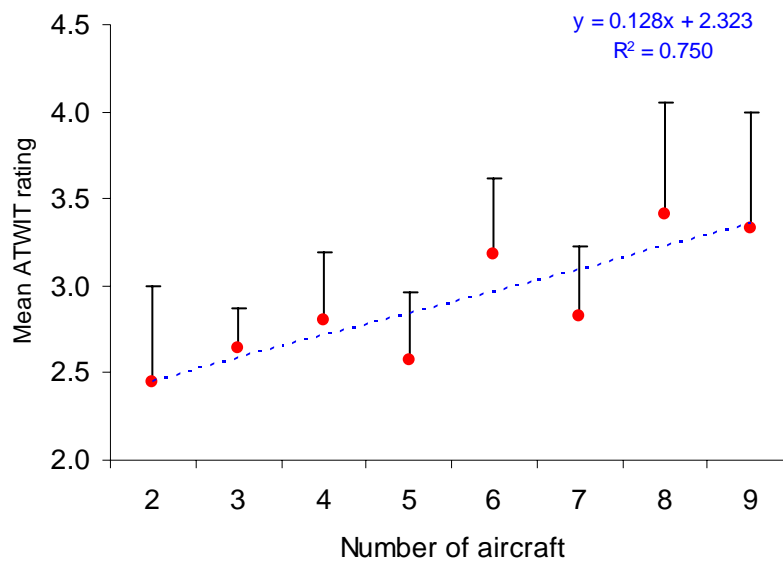


Figure 8. Mean ATWIT ratings as a function of the number of aircraft in the sector. Error bars are  $SE$ .

### 3.5 Eye Movement Measures

In addition to the ATWIT ratings, we also evaluated other, more objective measures to determine their suitability for use as measures of workload. Although ATWIT is a fairly reliable measure of workload, it has a number of issues related to its use. It can be an imprecise measure when trying to detect small changes in workload. ATWIT ratings primarily reflect workload at a particular instant and are, therefore, static and are not an effective way to measure moment-to-moment changes in workload over time. Because it is a subjective rating, it may not always accurately reflect true workload. For instance, there may be social pressures on controllers to report low workload levels. In addition, ATWIT ratings are unidimensional, but workload is seen as multifaceted (Parasuraman & Hancock, 2001); affected by environmental factors (e.g., aircraft count), controller activity level (e.g., communications and computer interactions), and operator state. Researchers would be well-served by having multiple metrics for measuring workload and for capturing its many different facets (Metzger, 2001). We explored whether eye movement activity metrics could serve in this capacity by evaluating whether they were sensitive to our measure of task load, target density, in our human-in-the-loop simulation.

#### 3.5.1 Blink Frequency and Blink Duration

We first evaluated blink frequency and blink duration. Prior findings indicate that both blink frequency and blink duration decrease as workload increases (Van Orden et al., 2001). In this study, we did not find that blink frequency decreased as the number of aircraft in the sector increased; the slope of the regression line was not statistically different from zero (see Figure 9). However, we did find that blink duration increased as the number of aircraft increased,  $R^2 = .595$ ,  $SE = .008$ ,  $F(1, 6) = 8.81$ ,  $p = .025$  (see Figure 10).

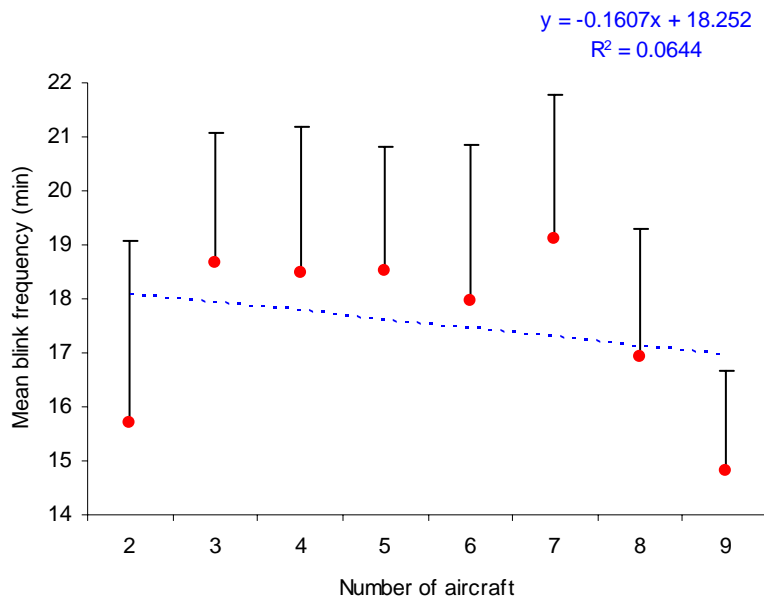


Figure 9. Mean blink frequency as a function of the number of aircraft in the sector. Error bars are *SE*.

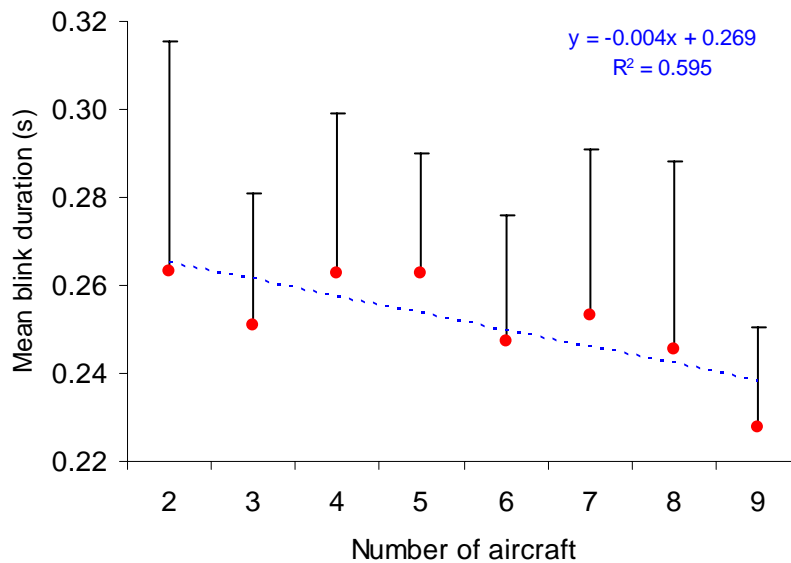


Figure 10. Mean blink duration as a function of the number of aircraft in the sector. Error bars are *SE*.

### 3.5.2 Saccade Frequency and Saccade Distance

Next, we analyzed the relationship between both saccade frequency and saccade distance and the number of aircraft in the sector. Although research has demonstrated that saccade frequency sometimes increases with increasing workload (Zeghal et al., 2002), we did not find the

relationship between saccade frequency and workload to be significant (see Figure 11).

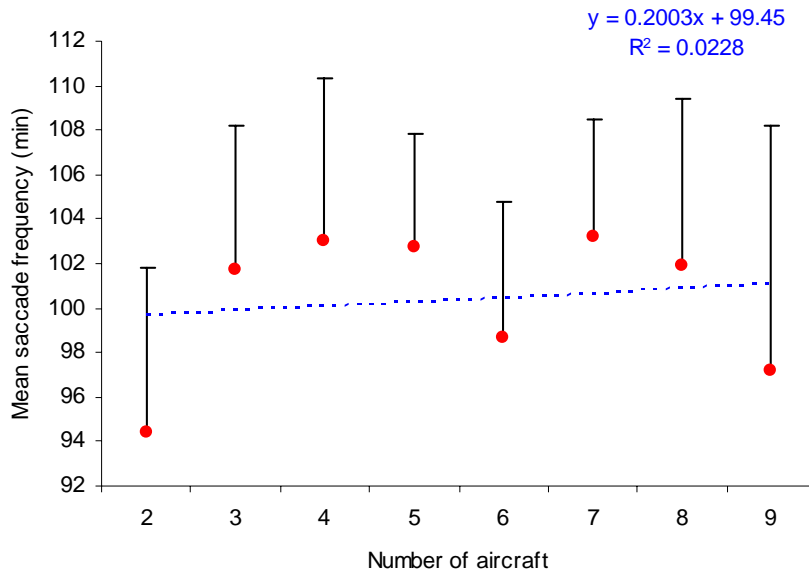


Figure 11. Mean saccade frequency as a function of the number of aircraft in the sector. Error bars are *SE*.

Although the relationship between saccade distance and workload can vary greatly, depending upon the task, Van Orden et al. (2001) found a decrease in saccade distance related to increased workload. Our findings also showed the same pattern (see Figure 12), with a decrease in saccade distance related to an increasing number of aircraft,  $R^2 = .686$ ,  $SE = .08$ ,  $F(1, 6) = 13.12$ ,  $p = .011$ . However, this result could potentially be due to a covariation between aircraft density and saccade distance. With an increasing number of aircraft in the sector, objects could be more densely spaced, which could lead to a decrease in the mean saccade distance. However, the complex relations that existed in the simulation between the number of aircraft and the distance between them due to weather movements, holding patterns, variation in traffic flow adjustments, runway selection, and controller style, did not contribute to a linear decrease in distances between aircraft with an increasing number of aircraft in the sector. Figure 12 shows the mean distance between aircraft in the sector for the eight levels of aircraft density. As we can see in the figure, the slope of the mean distance between aircraft is positive ( $R^2 = .576$ ,  $SE = .93$ ,  $F(1, 6) = 8.18$ ,  $p = .028$ ) and opposite the slope of the mean saccade distance. This indicates that the reduction in mean saccade distance was workload related and not caused by a reduced distance between aircraft in the sector.

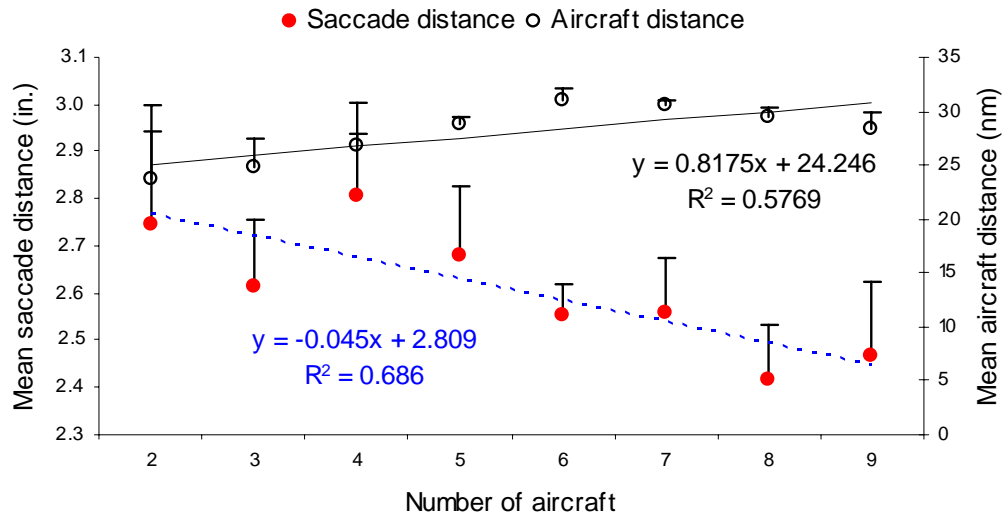


Figure 12. Mean saccade distance and mean distance between aircraft as a function of the number of aircraft in the sector. Error bars are *SE*.

### 3.5.3 Pupil Diameter

Many researchers (Iqbal et al., 2004, 2005; Van Orden et al., 2001; Zehgal et al., 2002) have found that relative pupil diameter increases as workload increases. In our study, we evaluated the relationship between relative pupil diameter and the number of aircraft in the sector. As can be seen in Figure 13, relative pupil diameter did increase as the number of aircraft increased,  $R^2 = .697$ ,  $SE = .02$ ,  $F(1, 6) = 13.83$ ,  $p = .01$ . We did use relative units in our measurement of pupil diameter, but these values may be converted to metric units (mm) by multiplying the relative units by a scaling factor of .044 (e.g., a relative unit value of 81 is approximately equal to a pupil diameter of 3.56 mm). The large *SE* bars in Figure 13 are due to individual differences in baseline pupil diameter size among controllers. The smallest measures of pupil diameter from individual controllers ranged from 54.5 to 87.9 ( $M = 77.1$ ,  $SE = 5.2$ ), while the largest individual measures ranged from 59.5 to 99.68 ( $M = 84.5$ ,  $SE = 5.9$ ).

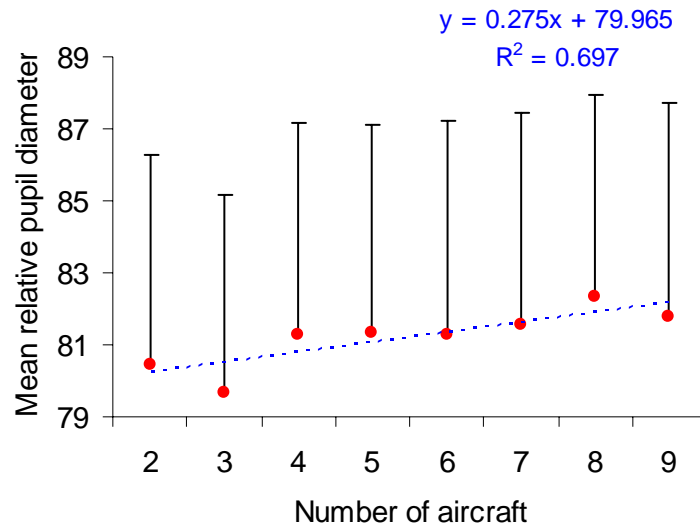


Figure 13. Mean relative pupil diameter as a function of the number of aircraft in the sector. Error bars are *SE*.

#### 4. Modeling

##### 4.1 Multiple Linear Regression Modeling

In our workload analysis, we showed that ATWIT ratings increase linearly with an increasing number of aircraft in the sector. We also demonstrated that changes over time in workload and traffic density measures correlate with the distance to weather from the outer marker. In our eye movement activity analysis, we showed that an increasing number of aircraft linearly decreased the mean blink duration and the mean saccade distance, while increasing the mean pupil diameter. Due to the systematic effects of target density on workload and eye activity measures, and their mutual dependence on the distance to Level 4 weather cells from the outer marker, we wanted to explore the possibility of using these parameters to predict the minute-by-minute target density in our simulation data (Van Orden et al., 2001).

Because of the apparent linear relationships in our data, we first explored the possibility of developing linear multiple regression models using stepwise regression. We developed prediction equations for individual controllers by using the normalized (*z*) minute averages for distance to weather (DW), blink duration (BD), saccade distance (SD), and pupil diameter (PD). Using the 27 complete data sets and the one data set with 28 minutes of oculometer data, we produced 28 individual regression models (one for each controller and simulation condition). These individual models showed considerable variability in their ability to predict the observed number of aircraft with a mean performance of  $R = .53$ , and a range of .19 to .89. Figures 14 and

15 show, respectively, examples of two individual models with low ( $R = .19$ ,  $F(4, 44) = .42$ ,  $p = .790$ ) and high ( $R = .89$ ,  $F(4, 45) = 41.70$ ,  $p < .001$ ) prediction performance. The prediction equation for the low-performance model is:

$$\text{Number of aircraft} = 0.020 \text{ DW} + 0.055 \text{ BD} - 0.194 \text{ SD} + 0.022 \text{ PD},$$

and the prediction equation for the high-performance model is:

$$\text{Number of aircraft} = -0.805 \text{ DW} + 0.084 \text{ BD} - 0.049 \text{ SD} + 0.167 \text{ PD}.$$

As can be seen in Figure 14, the low-performance model was unable to predict rapidly fluctuating aircraft numbers and missed all the peaks and valleys for the observed number of aircraft over time. The high-performance model in Figure 15 does not perform perfectly either, but the performance is much better than that of the low-performance model in Figure 14.

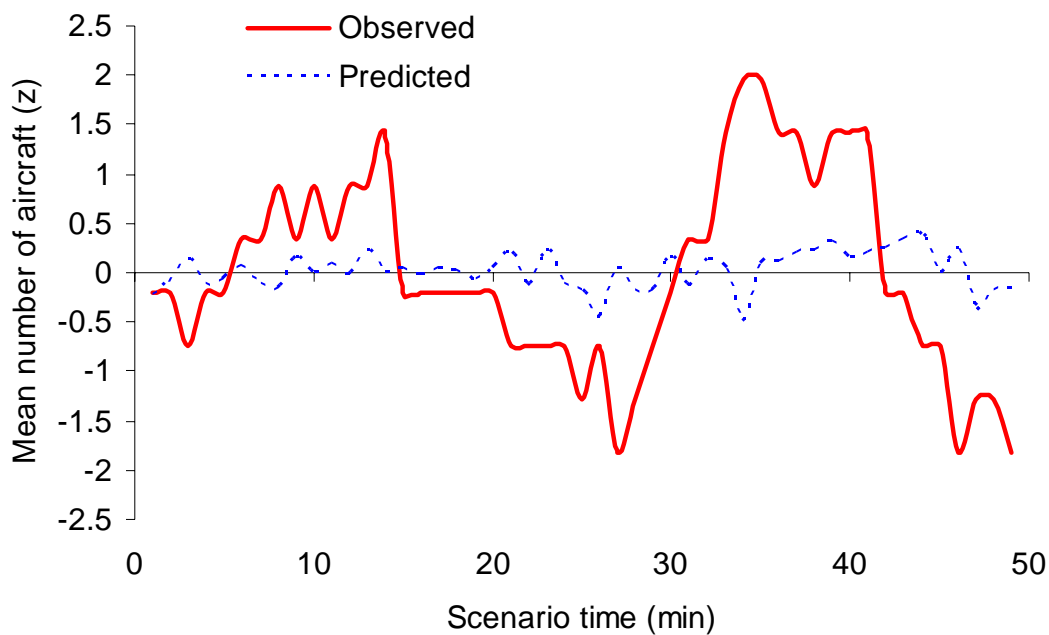


Figure 14. Individual low-performance model for observed and predicted number of aircraft as a function of scenario time ( $R = .19$ ).



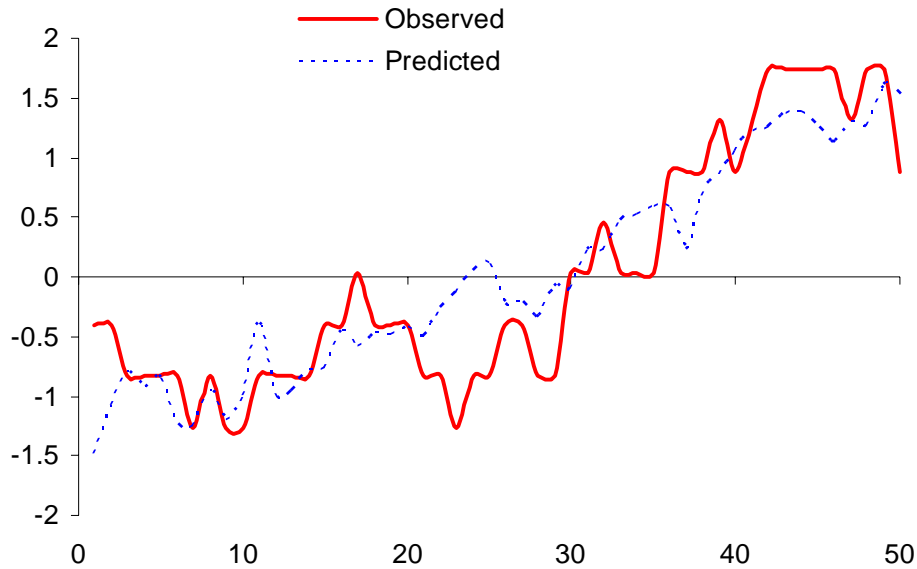


Figure 15. Individual high-performance model for observed and predicted number of aircraft as a function of scenario time ( $R = .89$ ).

After generating the predictions for the set of individual models, we attempted to develop a general model (across controllers) for predicting the number of aircraft in the sector. By taking the average of all conditions from each controller, we computed a grand average from all six controllers. For this general model, we developed a final prediction equation using stepwise regression, including the three significant regression variables: saccade distance ( $t(46) = -3.70$ ,  $p < .001$ ), blink duration ( $t(46) = 3.32$ ,  $p < 0.001$ ), and pupil diameter ( $t(46) = 2.33$ ,  $p < 0.023$ ). Distance from weather was not significant and was not included in the final equation. The final prediction equation was:

$$\text{Number of aircraft} = -0.507 \text{ SD} + 0.535 \text{ BD} + 0.466 \text{ PD}.$$

Figure 16 shows the performance ( $R = .58$ ) of the general model for predicting the number of aircraft in the sector ( $F(3, 46) = 7.99$ ,  $p < .001$ ).

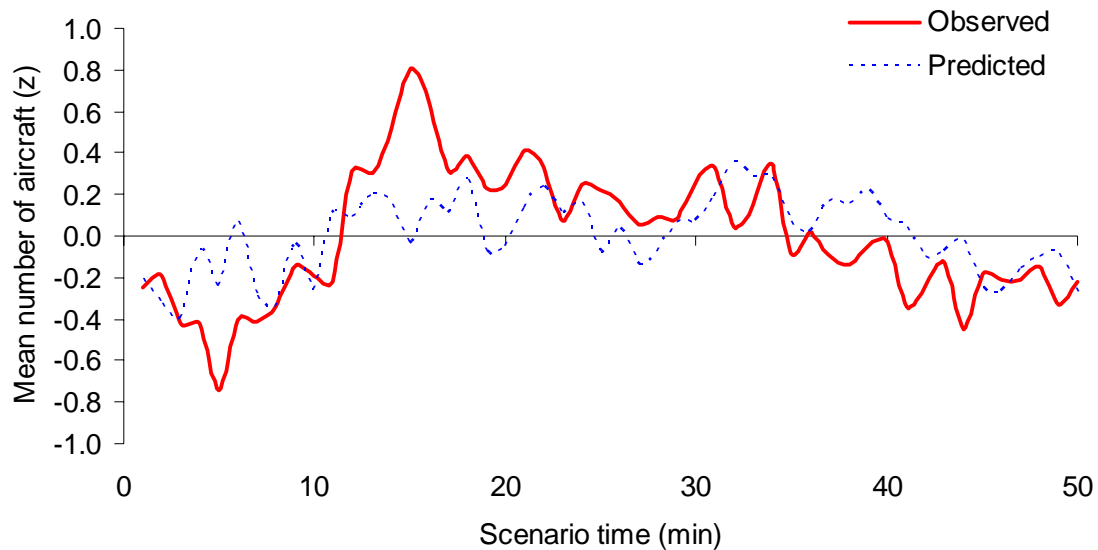


Figure 16. A general model for the prediction of the number of aircraft as a function of scenario time ( $R = .58$ ).

As can be seen in Figure 16, the performance of the general model is moderately good and similar to the mean performance of the individual models (see Figure 16). However, similar to the performance of the individual models, the general model also has a limited capability to capture rapid fluctuations in aircraft density.

#### 4.2 Artificial Neural Network (ANN) Modeling

Besides multiple regression techniques, researchers have also used ANN models in workload (Van Orden et al., 2001) and adaptive aiding research (Wilson, Lambert, & Russell, 1999). An advantage of ANN models is their ability to evaluate complex, non-linear, mathematical relationships between many variables. Also, ANN models usually yield higher prediction performances compared to regression models (Van Orden et al., 2001). Therefore, as an alternative to our multiple regression modeling approach, we also evaluated multilayer ANN models as possible candidates for predicting aircraft counts from our eye movement data. When training an ANN model, researchers present variables to the model. The model then tries to perform a classification based on the values of the input variables or else uses these values to predict the value of some other variable (e.g. aircraft count). As the ANN model receives more training, it gets better at classifying or predicting the outcome variable. After enough training, researchers can use ANN model to classify or predict outcomes for untrained items (see Shanks, 1995, for details). However, sometimes a network can overlearn training examples and then cannot classify or predict outcomes for new data. To counter this problem, researchers typically use part of their data to train the model and part of their data to test the model.

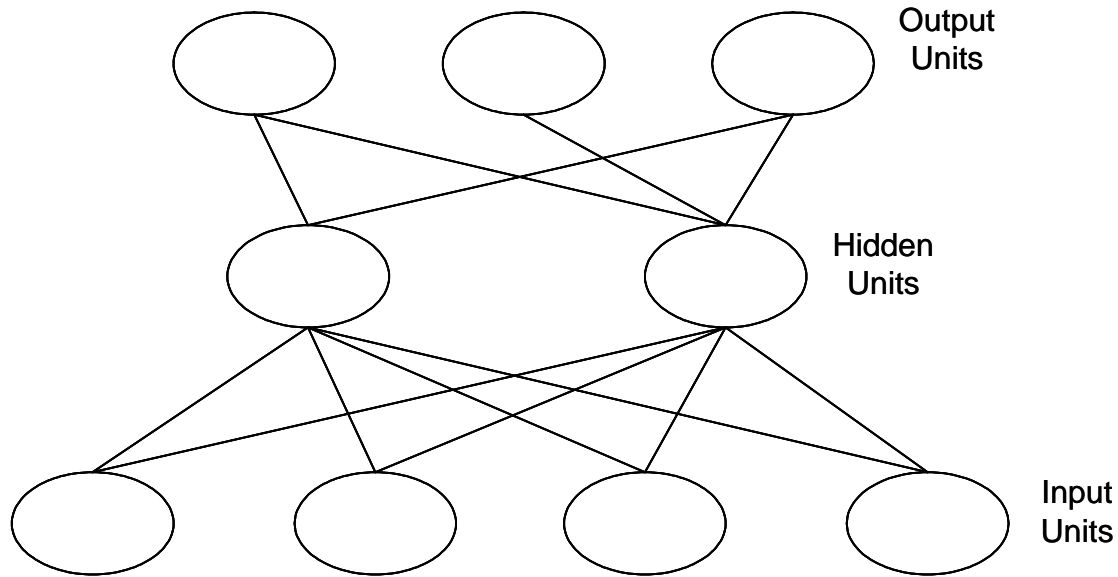


Figure 17. An example of a standard Artificial Neural Network (ANN) model with backpropagation.

We developed individual models for each controller in a given condition using DW, BD, SD, and PD to predict aircraft density. We developed 27 individual models, one for each controller and simulation condition. For each individual model, we randomized the data and selected half for training, using the remaining half to test the trained network. By increasing the number of test items, we increased the generalizability of the model. We stopped each training session after 50,000 learning trials. For each individual model, we performed five different training sessions. Out of these five sessions, we selected the one that produced the highest correlation between predicted aircraft count and actual aircraft count. Our analysis was only a preliminary attempt to demonstrate the value of these models, and we could potentially develop much more accurate models by averaging over seconds, not minutes, or by using data smoothing techniques (see Van Orden et al., 2001, for a description). Because we were not attempting to develop operational models, and because creating and running these models is very time consuming and costly, we did not attempt any further refinement of the models. However, for operational use, a prediction performance of 90% or greater is desirable (Schvaneveldt et al., 1998).

The 27 individual ANN models showed considerable variability in their ability to predict the observed number of aircraft, with a mean correlation in the cross-validation data between the actual number of aircraft and the predicted number of aircraft of  $r = .56$ , and a range of .26 to .92. Figures 18 and 19 show examples of two individual models with low ( $r = .26$ ) and high ( $r = .92$ ) prediction performance, respectively. Figure 18 demonstrates, the low performance model is unable to predict many peaks and valleys in the observed number of aircraft. On the other hand, the high-performance model in Figure 19 performs very well, outperforming the low-performance model by a wide margin.

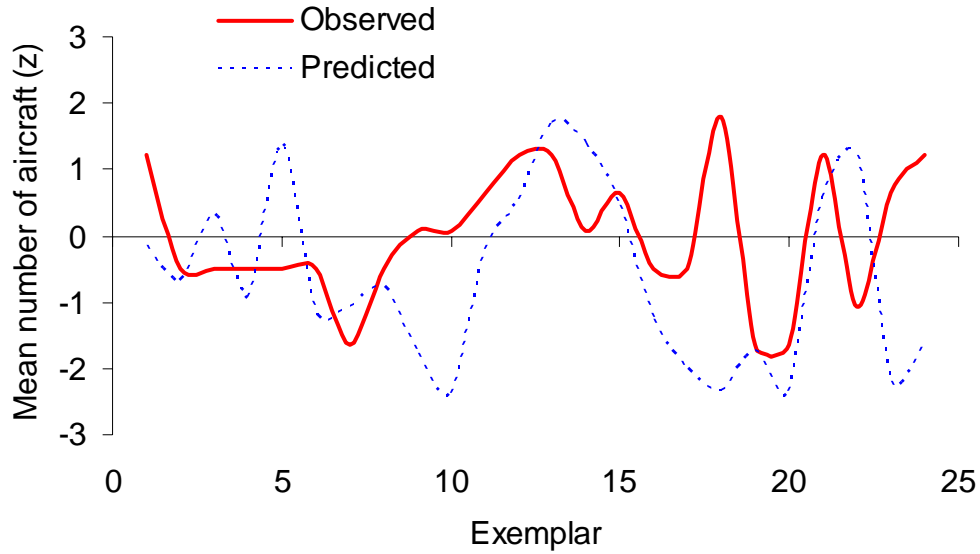


Figure 18. Individual low-performance model for observed and predicted number of aircraft for 24 exemplars ( $r = .26$ ).

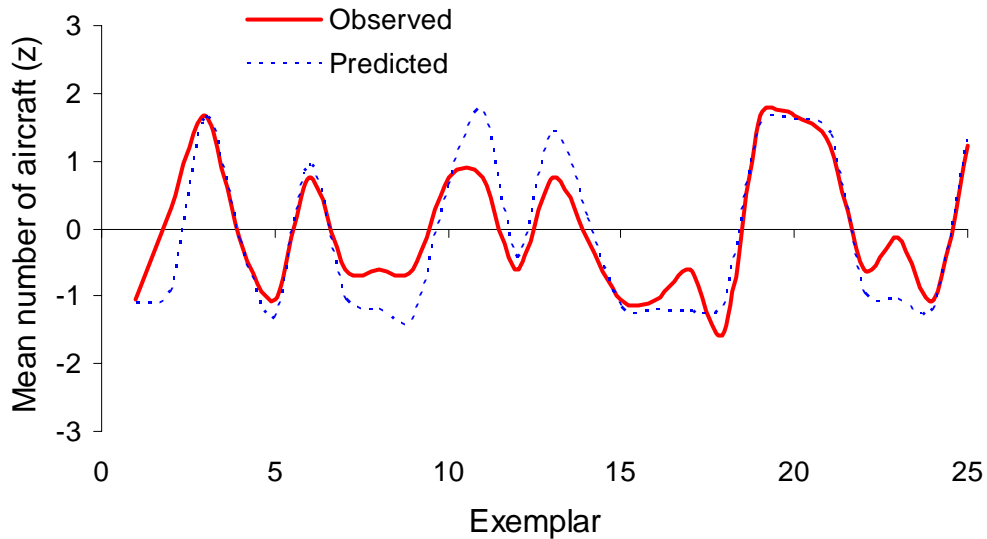


Figure 19. Individual high-performance model for observed and predicted number of aircraft for 25 exemplars ( $r = .92$ ).

After generating the set of individual models, we attempted to develop a general model (across controllers) for predicting the number of aircraft in the sector. We first took the minute-by-minute average across conditions from each controller. We then computed a grand mean of the minute-by-minute data over all six controllers. We then randomized the data and selected half the data (25 exemplars) for the training data set and half the data (25 exemplars) for the test data set. For the general model, we again predicted the number of aircraft using DW, BD, SD, and

PD. Figure 20 shows the performance of this general model for predicting the number of aircraft in the sector,  $r = .84$ .

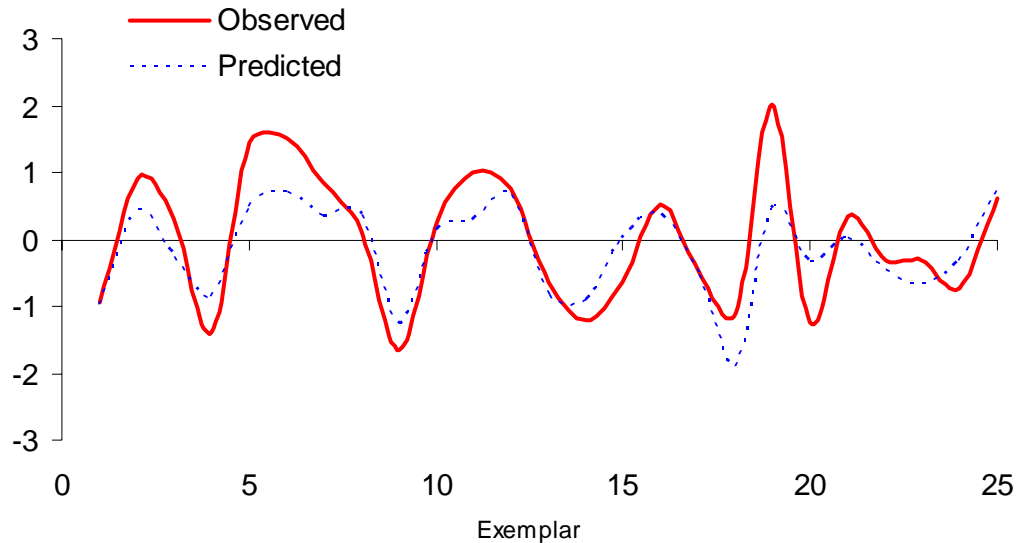


Figure 20. A general model for observed and predicted number of aircraft for 25 exemplars ( $r = .84$ ).

As seen in Figure 20, the performance of the general model is good and is somewhat higher than the mean performance of the 27 individual models. As with some of the individual models, the general model underpredicts some of the peaks and valleys in the actual data.

We compared the performance of the ANN models and the linear regression models for each subject in each condition (see Table 1 for complete results). We evaluated whether the correlation coefficient between the predicted number of aircraft and actual number of aircraft for individual conditions were significantly ( $p < .05$ ) different from zero. Six correlations for the regression models were not significantly ( $p > .05$ ) different from zero and one correlation coefficient for the ANN model was not significantly ( $p > .05$ ) different from 0,  $t_{crit} = 2.01$  for  $df = 48$ ,  $t_{crit} = 2.01$  for  $df = 47$ . We directly compared the correlation coefficients for the two model types using the Fisher Z transform, and found that in five instances the correlation coefficients for the two types of models were significantly ( $p < .05$ ) different from each other,  $z_{crit} = 1.96$ . Four of these instances favored the ANN model, while one favored the regression model. We also computed an average correlation for the 27 individual models for both model types. For both model types, these average correlations were significantly different ( $p < .05$ ) from 0,  $t_{crit} = 2.06$  for  $df = 25$ , but were not significantly ( $p > .05$ ) different from each other,  $z_{crit} = 1.96$ . We also computed correlation coefficients for the general models. We found that the correlation coefficients for both general models were significantly ( $p < .05$ ) different from 0,  $t_{crit} = 2.01$  for  $df = 48$ . In addition, the correlation coefficient for the general ANN model was significantly ( $p < .05$ ) higher than the correlation coefficient for the linear regression model,  $z_{crit} = 1.96$ .

The large number of tests involved in this analysis could be cause for some concern. Each comparison in Table 1 is equivalent to a  $t$  test with a 5% chance of a type one error (i.e., wrongly rejecting the null hypothesis). However, our main purpose in performing these tests was not to

make operational decisions based on a statistical analysis, but to explore the performance of ANN and regression models on our simulation data. Additionally, these statistical results are unlikely to be the sole outcome of type one errors. The binomial probability of obtaining 26 significant ( $p < .05$ )  $t$  tests out of 27 cases in our ANN data is  $p < .001$ . Similarly, the binomial probability of obtaining 21 significant ( $p < .05$ ) tests by chance alone from 27 cases in our regression data is  $p = .002$ .

Table 1. Results of the Performance Comparison of ANN and Regression Models (RM)

Case	Controller	Condition	Correlation between predicted and actual aircraft count – ANN	Correlation between predicted and actual aircraft count - RM
1	1	B	0.45*	0.60*
2	1	C	0.46*	0.59*
3	1	D	0.92*	0.89*
4	1	E	0.87*	0.79*
5	2	A	0.62*	0.53*
6†	2	B	0.40*	0.70*
7†	2	C	0.67*	0.28
8	2	D	0.70*	0.81*
9	2	E	0.50*	0.56*
10	2	F	0.67*	0.74*
11	3	B	0.50*	0.67*
12	3	D	0.61*	0.64*
13	4	A	0.41*	0.27
14	4	B	0.52*	0.50*
15†	4	E	0.77*	0.38*
16†	4	F	0.42*	0.36*
17	5	A	0.40*	0.68*
18	5	B	0.55*	0.54*
19	5	C	0.30*	0.60*
20	5	D	0.35*	0.25
21	5	E	0.54*	0.21
22	5	F	0.40*	0.46*
23	6	B	0.26	0.19
24†	6	C	0.58*	0.28
25	6	D	0.75*	0.30*
26	6	E	0.86*	0.79*
27	6	F	0.59*	0.68*
Average $r$ over conditions			0.56	0.53
General model †			.84	.59

\* Indicates a correlation that is significantly different from 0 ( $t = r \sqrt{\frac{N-2}{1-r^2}}$ ).

† Indicates a case with two correlations that are significantly different from one another using the Fisher Z transform, where

$$Z = \operatorname{arctanh} r, \left( z = \frac{Z_{r_1} - Z_{r_2}}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}} \right).$$

## 5. Discussion

One of the most important and difficult tasks for human factors researchers is to conceptualize and measure human operator workload. As more complex and dynamic systems are introduced in the work domain (Vicente, 1999), it is important to identify human operator capabilities and limitations in their response to system demands (i.e., task load). High task loads can lead to an increase in visual and cognitive demands and thereby increase operator workload. An increase in operator workload can negatively affect performance and lead to potentially higher risks for operating hazards.

As Stein (1998) aptly phrased it:

Workload is a construct. Constructs are used to label processes or experiences that cannot be seen directly but must be inferred from what can be seen and measured. As new person-machine systems are designed, the nature of workload and human error may also experience changes. No one wants to design systems for failure by overloading the operator. (p. 155)

Researchers have commonly used both self-reported workload ratings and physiological measures to gauge ATC controller workload during different task load conditions (Averty, Athenes, Collet, & Dittmar, 2000; Hilburn, 1997). In the present study, we explored how subjective workload ratings (ATWIT) correlated with the known ATC task load variable aircraft density (Lee, 2005; Lee, Prevot, Mercer, Smith, & Palmer, 2005; Rantanen, Maynard, & Özhan, 2005; Stein, 1985; Van Orden et al., 2001). First, we analyzed controller ATWIT ratings as a function of aircraft density. We found that controllers' subjective workload ratings increase linearly with an increasing number of aircraft in the sector. This result supports findings from previous research that aircraft density functions well as a task load variable. Second, we explored how aircraft density correlated with systematic changes in blinks, saccades, and the pupil diameter. If the task load variable aircraft density produces systematic changes in eye movement activity, we can use these changes as an indirect measure of controller workload. Our eye movement analysis confirmed the presence of these systematic relationships in our data. The mean blink duration and the mean saccade distance decreased as a function of an increased aircraft density, while the mean pupil diameter increased as a function of an increased aircraft density. This indicates that we can use changes in eye movement activity as a correlate for controller workload during ATC tasks. As a final test of these relationships, we explored the possibility of using blink duration, saccade distance, pupil diameter, and distance to weather data, to predict the minute-by-minute aircraft count from our recorded simulation data. Using both multiple linear regression and ANN modeling, we were able to produce individual controller models (regression  $R$  ranging from .19 to .89 and ANN  $r$  ranging from .26 to .92) and general models (regression  $R = .58$  and ANN  $r = .84$ ) with good performances in predicting the minute-by-minute number of aircraft in our recorded scenarios.

This raises the question as to how we can apply these measures in research and in operational settings in the ATC domain. Van Orden et al. (2001) proposed the use of eye activity correlates of workload and ANN modeling as a means to monitor operator workload and scanning efficiency. Others have proposed the use of blink activity and physiological variables in conjunction with ANN models for adaptive aiding (Wilson et al., 1999). We see a potential use for eye activity measures in conjunction with ATWIT ratings for ATC research. Although self-reported workload ratings are the most widely used technique for measuring workload in ATC

research (Stein, 1998), there are limits to how frequently researchers can probe a controller without interfering with the task at hand. Commonly, researchers have used five-minute probe intervals for ratings (Ahlstrom & Friedman-Berg, 2005), but shorter temporal delays with two- (Kuk, Arnold, & Ritter, 1999) and one-minute (Stein, 1985) intervals have also been used. From a practical standpoint, the one-minute interval is probably the shortest interval researchers could use without affecting and interfering with the controller's operations. Therefore, what is missing is a supplemental method for collecting workload estimates in-between self-reported workload ratings. We argue that eye movement activity measures, because of their role as workload correlates, could fill this gap. Researchers could use these measures to monitor real-time changes in workload during ATC simulations, and this would be especially useful for the development and testing of interface designs and procedures where we might expect rapid fluctuations in operator workload.

As far as modeling workload fluctuations, we used both multiple linear regression models and ANN models in the present study. There are advantages and disadvantages to using both methods. In general, neural network algorithms provide increased predictive performance compared to linear regression algorithms, especially if the data contains significant non-linear components (DeTienne, Lewis, & DeTienne, 2003; Ethridge & Zhu, 1996). This trend is noticeable in the present study, our ANN models appeared to have on average a higher prediction performance compared to our regression models. However, there were few statistically significant differences in the performance comparison between the multiple linear regression outcomes and the ANN model outcomes. Therefore, the relatively good performances of the linear models might favor their real-time applications in ATC research. Also, multiple linear regression models are less computationally intense to implement and require no model training, and no initial selection of training parameters. Therefore, they have advantages over ANN models in research settings where several models might be needed for different operator environments. Neural networks also do not present a model of the data that is easy to interpret. This is a disadvantage for exploratory research where the goal is to understand the underlying relationship among variables. In this respect, an ANN model is more of a 'black box' that delivers an output without an explanation. Furthermore, it is more difficult to incorporate a neural network model into an existing system or a simulation engine, and there is a question about the daily reliability of ANN models in operational settings (Wilson et al., 1999). Additionally, it has yet to be proven if one general ANN model can be used for all individual operators, or whether there is a need to establish a separate model for each operator.

However, there are also problems related to the use of regression models. First, regression models make certain parametric assumptions that researchers must not violate. Researchers do not need to make the same assumptions for neural network models. For instance when using regression models, researchers assume that the independent variables are not correlated (multicollinearity) and that errors are normally distributed and independent (Myers & Well, 2003). In many cases these assumptions are not met. Regression models are also unable to deal effectively with outliers while neural networks perform best when there is noise in the data (Marquez, Hill, Worthley, & Remus, 1991), including the noise introduced by outliers (DeTienne et al., 2003). Stepwise regression also has issues related to its use. Although it provides the best prediction equation, there is no guarantee that the resulting equation has any explanatory or theoretical value (Myers & Well). Furthermore, despite the fact that our general ANN and regression models (across controllers) provide good prediction performances, they are



both the result of post-hoc analyses. Therefore, we cannot take these results as a guarantee that real-time analyses with general models will perform equally well in operational settings.

Finally, although the present results corroborate previous research (Van Orden et al., 2001), there is an important difference in our approach and analysis. Most importantly, we did not manipulate traffic load in our weather simulation in order to maximize responses in eye movement activity and workload. Despite this, we demonstrated in our analysis that systematic relationships between traffic density, workload ratings, and eye activity measures exist, and that it is possible to assess these relationships even in high-fidelity human-in-the-loop simulation data. Although the systematic directions of the eye movement effects in our study were similar to those found by Van Orden et al. (2001), we found much smaller effect magnitudes. We attribute this fact to a lack of a systematic manipulation of task load variables in our simulation. It is interesting to note, however, that the eye movement activity effects are nevertheless present in simulation conditions where controllers operate under low to moderate workload conditions. Furthermore, while we found traffic density to be the primary task load variable, we also found that distance to weather from the outer marker also worked as a predictor of workload. However, this variable was only a significant predictor in half of the individual regression models, and we therefore excluded it from the final regression equation. Other researchers have explored variables like pilot-controller communications and their usefulness in predicting subjective estimates of controller workload (Manning, Mills, Fox, Pfleiderer, & Mogilka, 2001).

The most important outcome of this study is our demonstration of a strong correlation between subjective ATWIT ratings and eye movement activity measures. This result corroborates previous research that has validated ATWIT ratings as a reliable measure of operator workload, but it extends upon this research by highlighting the link between subjective workload ratings and physiological measures. Blink duration, saccade distance, and pupil diameter are reliable correlates of ATC operator workload. We have outlined here how these measures, in conjunction with ATWIT ratings, can improve upon our ability to measure rapid fluctuations in controller workload. Moreover, we also show how traffic density and weather cell proximity affect controller workload during severe weather avoidance. As traffic increases and heavy weather cells get closer to runways, there is a corresponding increase in controller workload. To regulate this increase in workload, controllers adopt a strategy where they decrease the number of arriving aircraft. This strategy shift underlines the importance of assessing ATC domain constraints (Ahlstrom, 2004) that affect controller operations and workload. We hypothesize that other domain constraints could affect workload as well. Future research should explore other constraints like aircraft state (e.g., configuration, time in sector, separation) and operator state (e.g., performance metrics), to assess their impact on regression and ANN model predictions. By exploring these constraints, we can increase the performance of these models, not only for research purposes, but also to the point where we can apply them successfully in real-world applications (Schvaneveldt et al., 1998; Van Orden, 2000).

## References

- Ahlstrom, U. (2004). *TRACON controller weather information needs: II. Cognitive work analysis* (DOT/FAA/CT-TN04/09). Atlantic City International Airport, NJ: Federal Aviation Administration, William J. Hughes Technical Center.
- Ahlstrom, U., & Friedman-Berg, F. (2005). *TRACON controller weather information needs: III. Human-in-the-loop simulation* (DOT/FAA/CT-05/12). Manuscript in preparation.
- Ahlstrom, U., Keen, J., & Mieskolainen, A. J. (2004). Weather Information Display System (WIDS). *Journal of Air Traffic Control*, 46(3), 7-14.
- Averty, P., Athenes, S., Collet, C., & Dittmar, A. (2000). Evaluating a new index of mental workload in real control situation using psychophysiological measures. In *Proceedings of the 21st Digital Avionics Systems Conference* (pp. 1-13). Irvine: IEEE.
- Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology*, 42, 361-377.
- DeTienne, K. B., Lewis, L. W., & DeTienne, D. H. (2003). Artificial neural networks for the management researcher: The state of the art. *2003 Research Methods Forum, Volume 8*. Retrieved August 17, 2005, from <http://aom.pace.edu/rmd/2003forum.html>
- Ethridge, D., & Zhu, R. (1996). Prediction of rotor spun cotton yarn quality: A comparison of neural network and regression algorithms. In *Proceedings of the Beltwide Cotton Conference* (pp. 1314-1317). Memphis, TN: National Cotton Council.
- Hilburn, G. G. (1997). Free flight and air traffic controller mental workload. In *Proceedings of the 9<sup>th</sup> International Symposium on Aviation Psychology*. Columbus: Ohio State University.
- Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., & Bailey, B. P. (2004). Changes in mental workload during task execution. In *Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology*. Santa Fe, NM. Retrieved July 12, 2005, from <http://orchid.cs.uiuc.edu/publications/UISTshort.pdf>
- Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., & Bailey, B. P. (2005). Towards an index of opportunity: Understanding changes in mental workload during task execution. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, Portland, Oregon. Retrieved July 12, 2005, from <http://orchid.cs.uiuc.edu/publications/p313-iqbal.pdf>
- Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). Task-evoked pupillary response to mental workload in human-computer interaction. In *Proceedings of the 2004 Conference on Human Factors in Computing Systems (CHI 2004)*, Vienna, Austria. Retrieved July 12, 2005, from <http://orchid.cs.uiuc.edu/publications/acm-chi-2004-iqbal.pdf>
- Kuk, G., Arnold, M., & Ritter, F. E. (1999). Effects of light and heavy workload on air traffic tactical operations: A hazard rate model. *Ergonomics*, 42(9), 1133-1148.
- Lee, P. U. (2005). A non-linear relationship between controller workload and traffic count. In *Proceedings of the Human Factors and Ergonomic Society 49th Annual Meeting* (pp. 1129-1133). Santa Monica, CA: Human Factors and Ergonomic Society.

- Lee, P. U., Prevot, T., Mercer, J., Smith, N., & Palmer, E. (2005). Ground-side perspective on mixed operations with self-separating and controller-managed aircraft. In *Proceedings of the 6th USA/Europe Air Traffic Management Research and Development Seminar (ATM2005)*. Baltimore, MD.
- Lin, Y., Zhang, W. J., & Watson, L. G. (2003). Using eye movement parameters for evaluating human-machine interface frameworks under normal control operation and fault detection situations. *International Journal of Human Computer Studies*, 59(6), 837-873.
- Manning, C., Mills, S., Fox, C., Pfleiderer, E., & Mogilka, H. (2001). The relationship between air traffic control communication events and measures of controller taskload and workload. In *Proceedings of the 4<sup>th</sup> USA/Europe Air Traffic Management R & D Seminar*. Santa Fe, NM.
- Marquez, L., Hill, T., Worthley, R., & Remus, W. (1991). Neural network models as an alternative to regression. In *Proceedings of the Twenty-Fourth Hawaii International Conference on System Sciences*, 4, 129-135.
- Metzger, U. (2001). *Automated decision aids in future air traffic management: Human performance and mental workload*. Unpublished doctoral dissertation, Technische Universität Darmstadt, Germany.
- Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis* (2nd ed.). New Jersey: Lawrence Erlbaum Associates.
- Parasuaraman, R., & Hancock, P. A. (2001). Adaptive control of workload. In P. A. Hancock & P. Desmond (Eds.), *Stress, workload, and fatigue*. Mahwah, NJ: Erlbaum.
- Rantanen, E. M., Maynard, P. W., & Özhan, D. (2005). The impact of sector characteristics and aircraft count on air traffic control communications and workload. In *Proceedings of the 13th International Symposium on Aviation Psychology* (pp. 491-496). Dayton, OH.
- Rhoda, D. A., & Pawlak, M. L. (1999). *An assessment of thunderstorm penetrations and deviations by commercial aircraft in the terminal area*, Report No. NASA/A-2. Lincoln Laboratory, MA: Institute of Technology: MA.
- Rognin, L., Grimaud, I., Hoffman, E., & Zeghal, K. (2004). Assessing the impact of a new instruction on air traffic controller monitoring tasks. In *Proceedings of the International Conference on Human-Computer Interaction in Aeronautics (HCI-Aero)*. Toulouse, France.
- Schvaneveldt, R. W., Reid, G. B., Gomez, R. L., & Rice, S. (1998). Modeling mental workload. *Cognitive Technology*, 3, 19-31. Retrieved July 12, 2005, from <http://www.interlinkinc.net/Roger/Papers/Workload.pdf>
- Shanks, D. R. (1995). *The psychology of associative learning*. New York: Cambridge University Press.
- Sollenberger, R. L., La Due, J. L., Carver, B., & Heinze, A. (1997). *Human factors evaluation of vocoders for air traffic control (ATC) environments phase II: ATC simulation (DOT/FAA/CT-TN97/25)*. Atlantic City International Airport, NJ: Federal Aviation Administration, William J. Hughes Technical Center.
- Sollenberger, R. L., McAnulty, D. M., & Kerns, K. (2003). *The effect of voice communications latency in high density, communications-intensive airspace (DOT/FAA/CT-TN03/04)*.

Atlantic City International Airport, NJ: Federal Aviation Administration, William J. Hughes Technical Center.

- Sollenberger, R. L., Willems, B., Della Rocco, P. S., Koros, A., & Truitt, T. (2004). *Human-in-the-loop simulation evaluating the collocation of the user request evaluation tool, traffic management advisor, and controller-pilot data link communications: Experiment I - Tool combinations* (DOT/FAA/CT-TN04/28). Atlantic City International Airport, NJ: Federal Aviation Administration, William J. Hughes Technical Center.
- Stein, E. S. (1985). *Air traffic controller workload: An examination of workload probe* (DOT/FAA/CT-TN84/24). Atlantic City International Airport, NJ: DOT/FAA, William J. Hughes Technical Center.
- Stein, E. S. (1991). Evaluating air traffic controller workload using real time person in the loop simulation. *Journal of Air Traffic Control*, 33(4), 55-58.
- Stein, E. S. (1992). *Air traffic control visual scanning* (DOT/FAA/CT-TN92/16). Atlantic City International Airport, NJ: Federal Aviation Administration, William J. Hughes Technical Center.
- Stein, E. S. (1998). Human operator workload in air traffic control. In M. W. Smolensky & E. S. Stein (Eds.), *Human factors in air traffic control* (pp. 155-183). New York: Academic Press.
- Van Orden, K. F. (2000). *Real-time workload assessment and management strategies for command and control watchstations: Preliminary findings*. Retrieved July 12, 2005, from <http://www.dtic.mil/matris/sbir/sbir011/Navy89b.doc>
- Van Orden, K. F., Jung, T. P., & Makeig, S. (2000). Combined eye activity measures accurately estimate changes in sustained visual task performance. *Biological Psychology*, 52, 221-240.
- Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T. P. (2001). Eye activity correlates of workload during a visuospatial memory task. *Human Factors*, 43(1), 111-121.
- Veltman, J. A., & Gaillard, A. W. K. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, 41(5), 656-669.
- Vicente, K. J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. Mahwah, NJ: Erlbaum.
- Willems, B., Allen, R. C., & Stein, E. S. (1999). *Air traffic control specialist visual scanning II: Task load, visual noise, and intrusions into controlled airspace* (DOT/FAA/CT-TN99/23). Atlantic City International Airport, NJ: Federal Aviation Administration, William J. Hughes Technical Center.
- Willems, B., Heiney, M., & Sollenberger, R. (2005). *Study of an ATC baseline for the evaluation of team configurations: Effects of allocating multisector control functions to a radar associate or airspace coordinator position* (DOT/FAA/CT-05/07). Atlantic City International Airport, NJ: Federal Aviation Administration, William J. Hughes Technical Center.
- Wilson, G. F. (2001). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology*, 12(1), 3-18.

- Wilson, G. F., & Caldwell, Jr., J. A. (2002). Cardiac and eye activity correlates of sleep loss in helicopter pilots. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*, (pp. 126-129). Santa Monica, CA.
- Wilson, G. F., Lambert, J. D., & Russell, C. A. (1999). Performance enhancement with real-time physiologically controlled adaptive aiding. In *Proceedings of the Human Factors and Ergonomics Society 44th Annual Meeting*, (pp. 61-64). Santa Monica, CA.
- Yuditsky, T., Sollenberger, R. L., Della Rocco, P. S., Friedman-Berg, F., & Manning, C. A. (2002). *Application of color to reduce complexity in air traffic control* (DOT/FAA/CT-TN03/01). Atlantic City International Airport, NJ: Federal Aviation Administration, William J. Hughes Technical Center.
- Zeghal, K., Grimaud, I., Hoffman, E., & Rognin, L. (2002). Delegation of spacing tasks from controllers to flight crew. Impact of controller monitoring tasks. In *Proceedings of the IEEE/AIAA Digital Avionics Systems Conference*, Irvine, CA.

## Acronyms

ANN	Artificial Neural Network
ANOVA	Analysis of Variance
ASL	Applied Science Laboratories
ATC	Air Traffic Control
ATWIT	Air Traffic Workload Input Technique
BD	Blink Duration
DESIREE	Distributed Environment for Simulation, Rapid Engineering, and Experimentation
DW	Distance to Weather
FAA	Federal Aviation Administration
ITWS	Integrated Terminal Weather System
PD	Pupil Diameter
POG	Points of Gaze
RDHFL	Research Development and Human Factors Laboratory
RM	Regression Model
SD	Saccade Distance
<i>SE</i>	Standard Errors
SOP	Standard Operating Procedures
STARS	Standard Terminal Automation Replacement System
TCW	TRACON controller workstation
TGF	Target Generation Facility
TRACON	Terminal Radar Approach Control
WIDS	Weather Information Display System
WS	Weather Scenario