DOT/FAA/TC-12/55

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

Effects of Weather Presentation Symbology on General Aviation Pilot Behavior, Workload, and Visual Scanning

Ulf Ahlstrom, FAA Human Factors Branch
Matthew Dworsky, Hi-Tec Systems, Inc.

November 2012

Technical Report

U.S. Department of Transportation
Federal Aviation Administration

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/TC-12/55		

4. Title and Subtitle	5. Report Date
Effects of Weather Presentation Symbology on General Aviation Pilot Behavior, Workload, and Visual Scanning	November 2012
	6. Performing Organization Code ANG-E25
7. Author(s) Ulf Ahlstrom, FAA Human Factors Branch Matthew Dworsky, Hi-Tec Systems, Inc.	8. Performing Organization Report No. DOT/FAA/TC-12/55
9. Performing Organization Name and Address Federal Aviation Administration Human Factors Branch William J. Hughes Technical Center Atlantic City International Airport, NJ 08405	10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Agency Name and Address Federal Aviation Administration Weather Technology in the Cockpit 800 Independence Avenue, S.W. Washington, DC 20591	13. Type of Report and Period Covered Technical Report 14. Sponsoring Agency Code ANG-C61

15. Supplementary Notes

16. Abstract

Objective: The purpose of this study is to explore the effects of cockpit weather presentation symbology on General Aviation (GA) pilot weather avoidance, weather presentation usage, and cognitive workload. **Background**: To support the Next Generation Air Transportation System (NextGen) program, on-going efforts focus on the implementation and use of weather technologies and weather presentations. Currently, there are no Federal Aviation Administration (FAA) or industry standards for the presentation of weather information in the cockpit. **Method**: Twenty-five instrument-rated GA pilots were randomly allocated to one of three simulation groups. During two 25-minute simulation flights, participants flew a Cessna 172 single-engine GA aircraft (using autopilot) under Visual Meteorological Conditions (VMC) and Instrument Meteorological Conditions (IMC). The pilots had to avoid the weather by using the cockpit weather presentation. We manipulated the cockpit weather presentation so that each pilot group used a different weather presentation symbology. **Results**: We found group differences in weather deviations, visual scanning behavior, and cognitive workload. **Conclusions**: Variations in weather presentations (colors and symbology) seem to affect pilot behavior and decision-making. **Applications**: This simulation is part of an on-going assessment of the effects of weather presentation symbology related to the standardization and optimization of weather presentations in cockpits.

17. Key Words	18. Distribution Statement
Eye Movement Analysis Functional Near Infrared General Aviation Human-in-the-Loop Instrument Meteorological Conditions Next Generation Air Transportation System Pilot Decision Making Weather Technology in the Cockpit Weather Avoidance	This document is available to the public through the National Technical Information Service, Alexandria, Virginia, 22312. A copy is retained for reference at the William J. Hughes Technical Center Library.

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	73	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

THIS PAGE IS BLANK INTENTIONALLY.

Table of Contents

List of Illustrations

Figures **Page**

Tables **Page**

Acknowledgments

The authors wish to thank Gary Pokodner, Program Manager of the Federal Aviation Administration (FAA) Weather Technology in the Cockpit (WTIC) Program Office, for sponsoring this research. We thank Ian Johnson (WTIC Human Factors Lead) and Roger Sultan (FAA Flight Technologies and Procedures Division, Aviation Safety Inspector) for their technical guidance of this research. In addition, we would like to thank Mark Mutchler (FAA Small Airplane Directorate) and Star McGettigan (FAA Weather Engineering & Evaluation Branch) for their assistance in the project, including test plan development.

We thank Albert Rehmann and the support of the Cockpit Simulation Facility. In particular, we would like to acknowledge April Stafford, Victoria Dzhussoeva, Kimberly Karaska, Robert Kusza, Thomas Granich, and George Chachis for their support with the Micro-Jet Simulator integration and flight scenario programming.

We would also like to thank Paul Robinson, David Forbes, Jason B. Prince (AeroTech Research, USA, Inc.), and Matthew Taylor (WSI Corporation) for their support with weather data and weather presentation software.

We thank Hilda Dimeo (Concepts & Systems Integration, Team Manager), Nick Marzelli, Josue Espinosa, Gordon Bond, and Dan Fumosa of the NextGen Integration and Evaluation Capability (NIEC) lab for analysis software, training, setup, and gracious use of their eyetracker system.

We thank Albert Macias, Otto Smith, John Dilks, and Wallace Daczkowski (FAA Human Factors Laboratory Concepts & Systems Integration Team staff) for providing audio and video recording as well as for eyetracker analysis support.

Thanks also to Professor Kurtulus Izzetoglu and Josh Harrison, from Drexel University, for lending us the functional near-infrared (fNIR) equipment and helping out with the fNIR analysis.

We thank Hal Olson (Hi-Tec Systems, Inc., Pilot SME) for providing simulation-pilot support. We thank Bill Thomas (Hi-Tec Systems, Inc., ATC SME), Bruce Slack (FAA Pilot, SME), and Matthew Kukorlo (FAA Pilot, ATC SME) for developing the flight scenarios and serving as SMEs to assist with this research.

THIS PAGE IS BLANK INTENTIONALLY.

Executive Summary

The Next Generation Air Transportation System (NextGen) encompasses many different aspects of aviation, with two key objectives of enhancing safety and efficiency. To support the NextGen program, this Weather Technology in the Cockpit (WTIC) project is focusing on enabling the use of weather technologies and weather presentations. New weather presentations need comparisons with current weather presentations to assure enhanced usability and safe support for pilots during operations.

Currently, there are no Federal Aviation Administration (FAA) or industry standards for the presentation of weather information in the cockpit. This results in large symbology variations between vendors. The lack of standardization can mean that some weather presentations are suboptimal. Suboptimal presentations can decrease usability and degrade safety margins.

The purpose of this study is to explore the effects of cockpit weather-presentation symbology on General Aviation (GA) pilot weather avoidance, weather presentation usage, and cognitive workload. Twenty-five instrument-rated GA pilots participated as volunteers during the simulation. From this pool of volunteers, we randomly allocated each pilot to one of three simulation groups. During two 25-minute simulation flights, participants flew a Cessna 172 single-engine GA aircraft (using autopilot) under Visual Meteorological Conditions (VMC) and Instrument Meteorological Conditions (IMC). The scenarios consisted of two flight segments on a route from Milton (MIP) in PA to Bradley International Airport in CT. Both scenarios started in midflight and there were no take-offs or landings. For both scenarios, pilots started in VMC but encountered IMC within 5 minutes into the flight. At 20 minutes into the flight, the weather conditions changed and the pilots were under VMC for the remainder of the flight.

During the simulation, we manipulated the cockpit weather presentation so that each pilot group used a different weather presentation symbology. The weather data were the same in all three weather presentations, but each presentation used different symbols and colors. The weather presentations were based on three commercially available symbology sets. In addition to a moving geographical terrain map and aircraft position symbol, each weather presentation contained precipitation information, meteorological report (METAR), significant meteorological information (SIGMET), and lightning information. Pilots could view the weather presentation at three different zoom levels (5 nmi, 20 nmi, and 50 nmi; also indicated on the weather display by range rings).

During a pre-simulation briefing, pilots were informed that a "significant squall-line thunderstorm associated with a fast moving cold front extending from Pittsburgh to Maine is moving southeast at 25 knots. This front has scattered, moderate to heavy thunderstorms and Instrument Flight Rules (IFR) areas below 200 Above Ground Level (AGL)." Pilots were told that they were aware of the weather and had received a preflight weather briefing but had chosen to proceed with their flight. Pilots were told to fly the scenario as if they were flying Visual Flight Rules (VFR) with two passengers aboard while receiving flight following from New York Center. For purposes of the study, the New York Center had a radar malfunction, making controllers unable to provide vectors around the weather. The pilots had to avoid the weather on their own using the cockpit weather presentation.

To assess if variations in weather presentation symbology affect pilot decision-making and flight behavior, we recorded simulation data in the following categories: decision-making (weather deviations), weather avoidance (distance to weather), weather presentation usage (zoom levels), radio communication (pilot/controller), system performance (altitude and heading changes), eye movements (visual scan patterns), and mental workload (oxygenation changes).

Under the tacit assumption that matched pilot groups will perform similarly (on average) when confronted with the same operational tasks, we used Bayesian model comparison to analyze our simulation data. Two model hypotheses were used in the analysis. The first hypothesis, the null, states that there is no difference in performance or behavior between pilot groups. The second hypothesis, the alternative, states that there are differences in performance or behavior between pilot groups. In addition to allowing inferences about the presence of effects (by means of Bayes factors), this framework also provides categories that classify the evidential strength from "anecdotal" to "extreme."

For null results of no difference between groups, we have the distance-to-weather measure where groups essentially performed the same and kept a safe distance (on average 14 nmi to 40 nmi) from the 30 dBZ precipitation intensities, despite variations in precipitation symbology. We found a similar null result for weather zoom usage and for pilot/controller communications. For radio communications, the number and durations for communications were similar across pilot groups. For the zoom usage, all groups displayed a similar use of the weather presentation zoom, and they displayed each zoom level for similar durations. Additional null results are pilot heading and altitude changes. Overall, the three pilot groups kept the same average scenario heading and altitude.

In contrast to these null results, we find effects that support the alternative and reflect a differential behavior between groups. Although most of these effects are anecdotal in support of the alternative, they are nevertheless operationally important due to the potential impact on GA pilots. These alternative effects include weather deviations, visual scan behavior, and cognitive workload. In contrast to the anecdotal effects, we find credible effects that range from "substantial" to "strong." Here we have group differences in weather deviations and visual scanning patterns. Based on the weather presentation, the pilot groups chose different deviation routes from start to destination. For the visual scanning behavior, we found group differences in scan patterns that included an increased number of fixations and saccades on the weather presentation and cockpit glass display.

To conclude, in this study, we found credible effects on pilot behavior, workload, and decision-making from the use of weather presentations. Variations of colors and symbols create very different images that affect interpretation and information retrieval. Most important, variations in colors and weather symbology seem to affect pilot behavior and decision-making. This answers our first question "Are there effects on pilot decision-making and behavior from manipulations of weather presentation symbology?" With these results in mind, we need to answer the next question "Are these symbology effects operationally important?" This entails the tailoring of other scenario and weather parameters, not investigated in this study, to assess their operational impact on pilot behavior.

1. INTRODUCTION

The Federal Aviation Administration (FAA) is involved in creating a Next Generation Air Transportation System (NextGen). This system will be an improvement upon today's National Airspace System (NAS) infrastructure. NextGen encompasses many different aspects of aviation, with two key objectives of increasing safety and efficiency. To support the NextGen program, on-going efforts focus on the implementation and use of weather technologies and weather presentations. New weather presentations require comparisons with current meteorological (MET) presentations to assure an enhanced usability and safe support for pilots during operations. The Weather Technology in the Cockpit (WTIC) program office (within the FAA Weather Research Branch) helps to accomplish this task.

The WTIC portfolio contains multiple efforts related to the standardization of MET information presentation in cockpits. These efforts are based on the premise that the lack of standardization is degrading pilot decision-making capability on weather-related events. Pilots should be able to interpret weather information clearly during dynamic and multitasking situations (FAA, 2010). This intuitive presentation allows the pilots to take actions considering future weather status even when they are under high task load (McAdaragh, 2002). In order for pilots to use this information, it is crucial that these weather presentations integrate multiple weather sources to provide accurate depictions of future weather hazards. The pilot should be able to recognize the weather condition, identify individual weather symbols from other symbols in the same area, and decide a course of action (Grasse, Schilke, & Schifele, 2008). Weather decision-making is enhanced only if the weather information presentation has a temporal resolution within the time horizon of the weather decision and is spatially accurate. The interfaces for these weather presentations should support rapid interpretation and understanding, thereby attenuating the mental resources needed to analyze and interpret weather data.

Currently, there are no FAA or industry standards for the presentation of MET information in the cockpit, which results in large symbology variations between vendors. Although not an official standard, the Radio Technical Commission for Aeronautics (RTCA) provides some guidance for the display of weather data in the cockpit (RTCA, 2004). The lack of standardization can mean that some presentations of MET information are suboptimal. Suboptimal information presentation can decrease usability and degrade safety margins. Therefore, this study seeks to examine the impact of weather presentation in the cockpit with the goal of increasing data usability and enhancing safety. Previous research has mainly focused on qualitative analyses, without much quantitative data on the effects of weather presentations on pilot behavior and decision-making. This study will provide quantitative data that can help the WTIC program to prioritize efforts and focus on important human factors issues related to the use of MET displays.

1.1. Background

Little is known about the effects on pilot behavior from variations in weather presentation symbology. Previous research has mostly focused on pilot use of precipitation information. For example, research by Beringer and Ball (2004) compared the situation awareness of pilots using the Next Generation Radar (NEXRAD) information at varying levels of resolution. They found that pilots who relied more on high-resolution NEXRAD images attempted to navigate between weather more than pilots with low-resolution presentations and windscreen verification did. This suggests that pilots will take higher risks going through weather systems if they are using the NEXRAD high-resolution system. There are other issues with NEXRAD presentations as well. Elgin and Thomas

(2004) stress the fact that NEXRAD information may be 5 minutes old when it reaches the weather service provider. It takes another minute or two for the service provider to broadcast the data. On top of that, the cockpit presentation only updates once every 5-7 minutes. This process results in weather data presentations that can be more than 14 minutes old, so that by the time the pilot sees the information, it may no longer be accurate. This is a serious problem when pilots use this weather information as a guide to fly between weather areas when in reality the window of opportunity to perform the appropriate and safe maneuver has already disappeared. This temporal uncertainty of weather data induces greater risks and can potentially cause damage to the plane and passengers if the plane is trapped in a storm. Furthermore, not only are there latencies in generating and presenting the weather data, but any time stamps on the data could be erroneous or misleading depending upon how the time stamp is generated.

Yuchnovicz, Novacek, Burgess, Heck, and Stokes (2001) found that the compelling nature of NEXRAD images caused some pilots to depend too heavily on the presentation and ignore other sources. This research also showed that pilots' use of a graphical radar display does not ensure optimal decision-making. Another NEXRAD study on General Aviation (GA) pilots (Chamberlain & Latorella, 2002) found that most pilots felt less than 50% confident in their awareness of the age of incoming data. Nevertheless, some still felt comfortable flying between convective weather in Instrument Meteorological Conditions (IMC), a potentially dangerous combination. In addition, NEXRAD does not detect nonconvective hazards, such as clear air turbulence (CAT), which is a serious issue in GA (Comerford, 2004). These issues make it important for the pilot to obtain other corroborating information, such as Meteorological Report (METAR) text data and Air Traffic Control (ATC) information, so that the accuracy of data can be verified.

Other researchers have analyzed Pilot Decision Making (PDM) or the overall decision-making process of pilots. The integrated model of naturalistic decision-making developed by Elgin and Thomas (2004) looks at three possible PDM modes depending upon the characteristics of the phase of flight (e.g., task load and time stress). This information suggests that pilots under high stress use their intuitive processing, personal skill, and pattern matching to reduce cognitive load. This PDM mode must be taken into account when designing optimal weather presentations, because the data presentation must be compatible with the user's level of urgency, stress, and workload.

Currently, we have little empirical knowledge on how weather symbology presentations support pilots during flight, although symbology research has suggested that presentation symbology can affect pilot decision-making and behavior. For example, presentations that deviate from the safety color series (green/yellow/red) or other familiar schema has the potential to cause misinterpretation of weather data (Arend, 2003). Hegarty, Canham, and Fabrikant (2010) examined weather map salience and display design, and found that display symbology and cue salience can have large effects on task performance. Today's commercial MET products contain a wealth of weather symbols. However, due to the lack of standardization and little empirical data, it is unclear what parts of symbology presentations enhance pilot behavior and decision-making and what parts have the opposite effect. Therefore, there is a need to assess the effects of weather presentation symbology on pilot behavior to enhance development of presentations for the safe, efficient, and tactical needs required to handle adverse weather conditions.

1.2. Purpose

This study explores the effects of cockpit MET presentations on GA pilot weather avoidance, weather presentation usage, and workload.

2. METHOD

2.1. Participants

Twenty-five instrument-rated GA pilots participated as volunteers during the simulation. The participants came from a pool of qualified pilots working at the FAA William J. Hughes Technical Center (WJHTC) and other FAA facilities. From this pool of volunteers, we randomly allocated each pilot to one of three simulation groups (Groups 1-3). Before the simulation, each participant provided biographical data to determine age, level of pilot experience, and experience with weather presentations. We use the median (middle value of a data set), First Quartile (Q1, median of the lower half of the data set), Third Quartile (Q3, the median for the upper half of the data set), and Interquartile Range (IQR, the spread of the middle 50% of the values) when reporting age and flight data. For the initial group of 25 pilots, the median age = 58 years (Q1 = 48, Q3 = 64, IQR = 16), the median total flight hours = 4000 (Q1 = 1350, Q3 = 5150, IQR = 3800), the median instrument flight hours = 300 (Q1 = 100, Q3 = 800, IQR = 700), and the median instrument flight hours within the previous six months = 3 (Q1 = 0, Q3 = 21.5, IQR = 21.5).

The flight hours were estimates given by each pilot before the simulation. Table 1 shows the same data for the three simulation groups after the random allocation procedure.

Table 1. Simulation Group Data for Age and Flight Hours

	Group 1	Group 2	Group 3
Age	Median = 64 Q1 = 59 Q3 = 69.5 IQR = 10.5	Median = 56 Q1 = 49.5 Q3 = 61.5 IQR = 12	Median = 53 Q1 = 42 Q3 = 61.5 IQR = 19.5
Total flight hours	Median = 3500 Q1 = 1750 Q3 = 6330 IQR = 4580	Median = 3100 Q1 = 675 Q3 = 5150 IQR = 4475	Median = 4000 Q1 = 1600 Q3 = 5600 IQR = 4000
Instrument flight hours	Median = 350 Q1 = 225 Q3 = 850 IQR = 625	Median = 150 Q1 = 29 Q3 = 1250 IQR = 1221	Median = 300 Q1 = 175 Q3 = 575 IQR = 400
Instrument flight hours within the previous 6 months	Median = 2 Q1 = 0 Q3 = 6.5 IQR = 6.5	Median = 2 Q1 = .5 Q3 = 21.5 IQR = 21	Median = 7.5 Q1 = 2 Q3 = 30 IQR = 28

Note. Q1 = First Quartile; Q3 = Third Quartile; IQR = Interquartile Range.

2.1.1. Participant Familiarity with Weather Data Sources

As Table 2 shows, the three simulation groups were adequately matched according to age and flight hours. In this section, we report the group data for the use of external sources (weather data), experience with weather systems, and weather training.

Table 2. External Sources Used by Group

	ATC	Flight watch	DUAT	Fore flight	FSS	Websites	Garmin	Radio	NOAA	AOPA	Flight follow	None
Group 1	1	1	1	0	4	0	0	1	1	1	1	1
Group 2	1	2	2	1	3	4	1	1	2	0	0	0
Group 3	3	2	0	0	3	1	0	1	1	0	0	0
Total	5	5	3	1	10	5	1	3	4	1	1	1

Note. ATC = Air Traffic Control; DUAT = Direct User Access Terminal; FSS = Flight Service Station; NOAA = National Oceanic and Atmospheric Administration; AOPA = Aircraft Owners and Pilot Association.

2.1.1.1. External sources used

The pilots commented that they used many different types of external sources for gathering weather information while in flight. According to our query, 24 out of 25 pilots interviewed used more than one external source while in flight. Only one pilot stated no use of an external source. Of the remaining 24 pilots, eight pilots used only one external source while flying, seven pilots used two sources, and nine pilots used three or more sources during flight. The interviewed pilots across groups relied on Flight Service Station (FSS) and ATC information the most. As shown in Table 2, Foreflight (Group 2), Garmin (Group 2), Aircraft Owners and Pilot Association (AOPA, Group 1), and Flight Following (Group 1) were all used by only one pilot each. Direct User Access Terminal (DUAT) service (two pilots in Group 2, and one pilot in Group 1) and Radio systems (one pilot in each group) were all used by three pilots each. Pre-flight Planning tools, such as Flight Watch (two pilots in Group 2, and two pilots in Group 3) and National Oceanic and Atmospheric Administration (NOAA) weather (one pilot in Group 1, and one pilot in Group 3, and two pilots in Group 2) were all used by four pilots each. One pilot in Group 1, one pilot in Group 2, and three pilots in Group 3 used ATC contact. We saw the use of in-air websites by four pilots in Group 2 and by one pilot in Group 3. Three pilots in Group 3, three pilots in Group 2, and four pilots in Group 1 used FSS services.

2.1.1.2. Simultaneous use of sources

Pilots from Group 1 and Group 3 used the most sources simultaneously; with the highest being four sources at once. In Group 1 and Group 3, two pilots used more than one source compared to three pilots who used more than one source in Group 2. One pilot in Group 1 and one pilot in Group 2 used four sources simultaneously.

2.1.2. Participant Experience and Training

2.1.2.1. Weather system experience

Six pilots had flight experience with Sirius XM weather presentations and one pilot in the Group 3 condition was familiar with the WSI radar presentation. Three pilots in Group 1 and Group 2 used Sirius XM on the Garmin systems. Two pilots, one in Group 2 and one in Group 3, had G1000 glass NEXRAD display experience. One pilot in Group 1 was familiar with the GT 330 MODE S display for weather presentation. One pilot in Group 1 acknowledged the use of a U.S. Air Force display in the cockpit. We found commercial product experience, such as Flight Watch, Foreflight, and Garmin Strike Finder, only in Groups 2 and 3. The Aircraft Communication Addressing and Reporting System (ACARS), Automatic Terminal Information Service (ATIS), Automatic Dependent Surveillance-Broadcast (ADS-B), and Flight Information Services-Broadcast (FIS-B) information were seen in Group 1 only.

2.1.2.2. Weather system training received

We asked pilots if they had any previous training in weather interpretation or meteorology. Roughly, half of the pilots (12 out of 25) had no previous training. Out of these individuals without training, there were three pilots in Group 2, four pilots in Group 3, and five pilots in Group 1. A single Group 3 pilot acknowledged being a tester and evaluator for weather presentations. One pilot in Group 2 and one pilot in Group 3 had aviation weather and radar training for their ratings and one pilot in Group 1 reported being self-taught. The remainder of the pilots had formal training experiences. Six pilots had military recurrent training (U.S. Air Force, U.S. Army, U.S. Coast Guard), with Groups 1 and 2 having three and two pilots, respectively. Only one pilot in Group 3 had any military training. The remaining four pilots had meteorology courses in college. These four pilots spread across Group 2 and Group 3 with two in each group. No pilots in Group 1 had formal training other than military.

2.2. Equipment

During this simulation, we used pilot simulation equipment and specialized data collection equipment for workload and eye-movement recordings. In this section, we will report the various systems and devices we used for data recordings.

2.2.1. Micro-Jet Simulator

The Micro-Jet simulator, which is housed in the Cockpit Research Facility at the WJHTC, uses one 13-inch flat screen to show cockpit controls, two 17-inch monitors for controlling simulator flight data (technician use), and one large curved monitor to portray the "out-the-window" view (see Figure 1).

Figure 1. The Micro-Jet cockpit simulator.

2.2.2. Eyetracker System

During simulation runs, participants wore a Mobile Eye Eyetracker, equipped with a Northern Digital Inc. (NDI) company Headtracker device. The Eyetracker from Applied Science Laboratories (ASL; refer to the Web site: www.asleyetracking.com/) was used to assess the number, duration and location of fixations on the weather presentations. The Eyetracker records Point of Gaze (POG) and pupil diameter by using near-infrared reflection outlines from the pupil and the cornea[1] (ASL Inc., 1991). Eye-movement activity correlates with cognitive workload, and it is used to assess operator scan patterns (Ahlstrom & Friedman-Berg, 2006). In addition to the eye-tracking capability, the system includes a head-motion tracker to capture head and neck movements to support the Eyetracker analysis. The sampling rate of the Eyetracker system is 60 Hz.

2.2.3. Functional Near-Infrared System

Pilot cognitive workload was measured during flights using an objective method to record functional cortical activity by means of a portable Functional Near-Infrared (fNIR; refer to www.biomed.drexel.edu/fNIR/CONQUER/Optical_Brain_Imaging.html) spectroscopy system (Izzetoglu, Bunce, Shewokis & Ayaz, 2010; Izzetoglu, Bunce, Izzetoglu, Onaral & Pourrezaei, 2007).

[1] During our IRB review, we submitted detailed information on the eye-tracking system to the Aerospace Medicine Division. The IRB review of the eye-tracking system and our intended use of the system did not reveal any unsafe conditions for participants.

The fNIR technology uses specific wavelengths of light to measure changes in cortical activity through the relative ratios of deoxygenated hemoglobin and oxygenated hemoglobin. The continuous wave fNIR system is connected to a flexible forehead sensor pad that contains four light sources (peak wavelengths at 730 nm and 850 nm) and 10 detectors. This configuration generates a total of 16 measurement locations (or voxels) per wavelength. With two wavelengths and dark current recordings for each of the 16 voxels, the system generates 48 measurements for each 2 Hz sampling period.

2.2.4. Voice Communication System

During the simulation, an ATC Subject Matter Expert (SME) provided ATC controller services to pilots upon request. The laboratory voice communication system provided a link between the pilots and the ATC controller through a Push-To-Talk (PTT) capability. The voice communication system equipment monitors and records the times and durations of pilot and ATC communications for subsequent analysis.

2.2.5. Software

The cockpit simulator runs on the Microsoft® Flight Simulator 2004 via the Project Magenta workstation control scheme for a Cessna 172 aircraft. Figure 2 shows the cockpit glass panel and the weather presentation display. All weather scenarios were superimposed on a moving geographical terrain map that included a symbol for the aircraft position. The simulator setup was for a single-engine aircraft as this is representative of common GA aircraft.

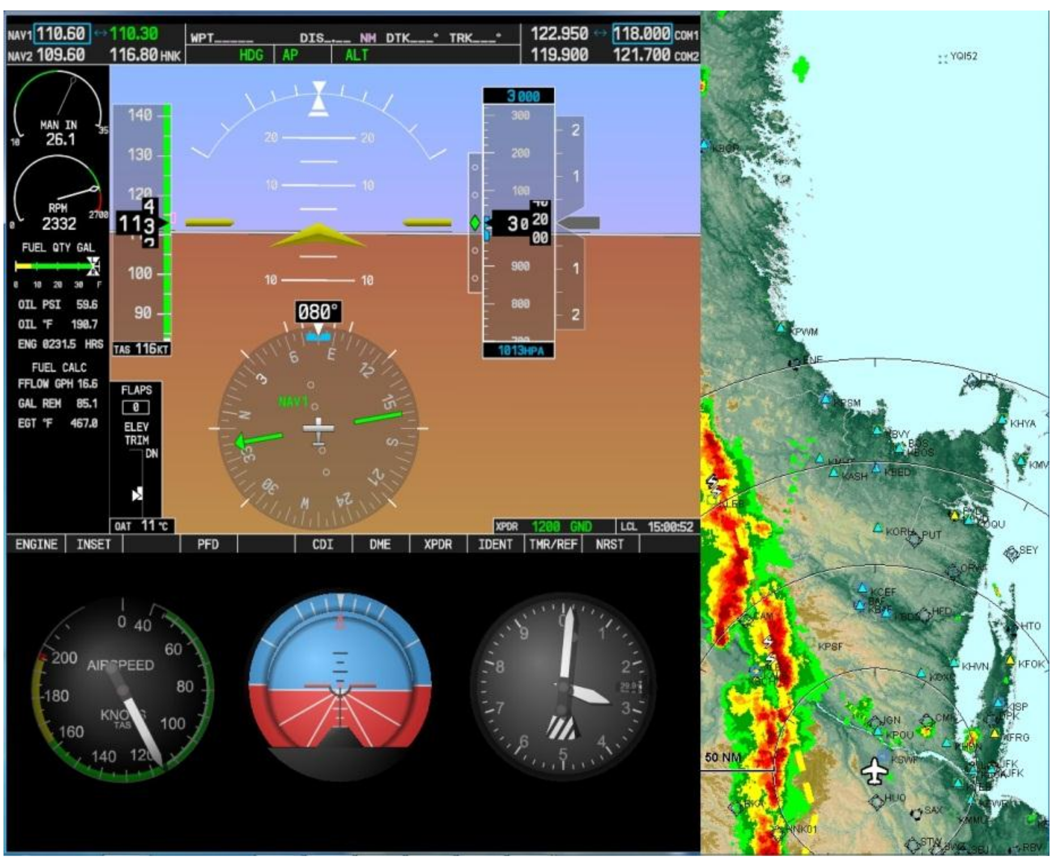


Figure 2. Cockpit glass and weather presentation display.

2.3. Materials

2.3.1. Weather Presentation Cheat Sheet

During the simulation, we affixed a small card, with a description and illustration of each weather data element, to the cockpit. This is equivalent to a pilot having the user manual for the weather presentation available during flight.

2.3.2. Weather Presentation Questionnaire

After completing the simulation flight scenario, the participants rated weather presentation features and usability on a weather presentation questionnaire (Appendix A). These ratings assess how participants perceived the presentation of weather information, flight efficiency, weather data interpretation, and weather situational awareness. In addition, participants could add additional comments that they deemed important for the simulation.

2.3.3. Weather Scenarios

During the two simulation flights, participants flew a Cessna 172 single-engine GA aircraft under Visual Meteorological Conditions (VMC) and Instrument Meteorological Conditions (IMC) from a starting point in Pennsylvania. The pilot managed flight controls and airspace constraints, as well as paid attention to the occurrence of hazardous aviation weather conditions. The scenarios navigated a route from the Milton (MIP) very high frequency (VHF) omnidirectional radio range (VOR) in Pennsylvania to Bradley International Airport in Connecticut. All scenarios started in mid-flight. There were no take-offs or landings. For both scenarios, pilots started in VMC but encountered IMC 5 minutes into their flight. At 25 minutes into the flight, the weather conditions changed and pilots were under VMC for the remainder of the flight.

2.4. Simulation Design

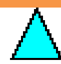
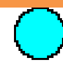
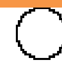
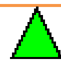
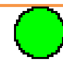
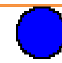
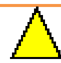
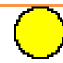
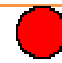
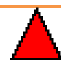
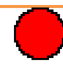
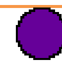
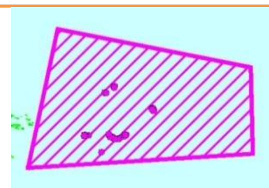
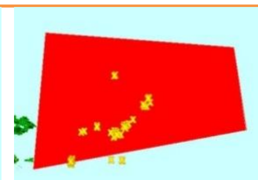
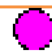
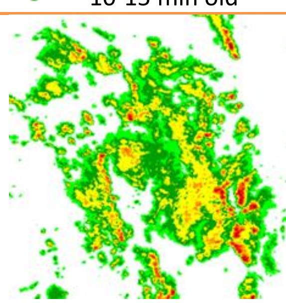
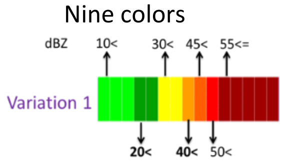
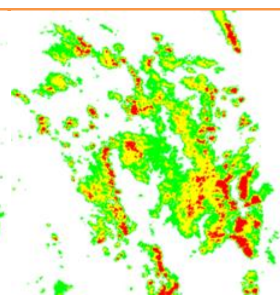
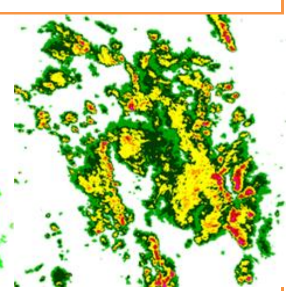
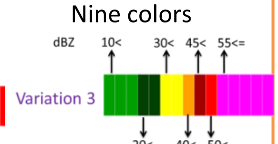
In this section, we will review the independent and dependent variables as well as the data collection and procedures needed to run the simulation.

2.4.1. Independent Variables

2.4.1.1. Weather presentation

During the simulation, each pilot group (Groups 1-3) used a different weather presentation alternative. Each of the data types in Table 3 presents different pieces of information to the pilot, with each presentation having its own set of symbols and colors. The data types that we used in the present simulation are **Precipitation, METAR, SIGMET area, and Lightning**. Pilots could view the weather presentation at three different zoom levels; Zoom 1 = 5 nmi, Zoom 2 = 20 nmi, and Zoom 3 = 50 nmi (also indicated on the weather display by range rings). In our weather presentation, the update rate was 5 minutes for all weather data elements. Figure 3 shows the weather presentations used by all groups (Groups 1-3) during the simulation.

Table 3. Primary Weather Data

Weather Data	Weather Parameters	Group 1	Group 2	Group 3
METAR	VFR			
	MVFR			
	IFR			
	LIFR			
SIGMET				
Lightning		 <5 min old  5-10 min old  10-15 min old		
Precipitation		 Nine colors 	 Five colors 	 Nine colors 

Note. METAR = Meteorological Report; SIGMET = Significant Meteorological Information; VFR = Visual Flight Rules; MVFR = Marginal Visual Flight Rules; IFR = Instrument Flight Rules; LIFR = Low Instrument Flight Rules.

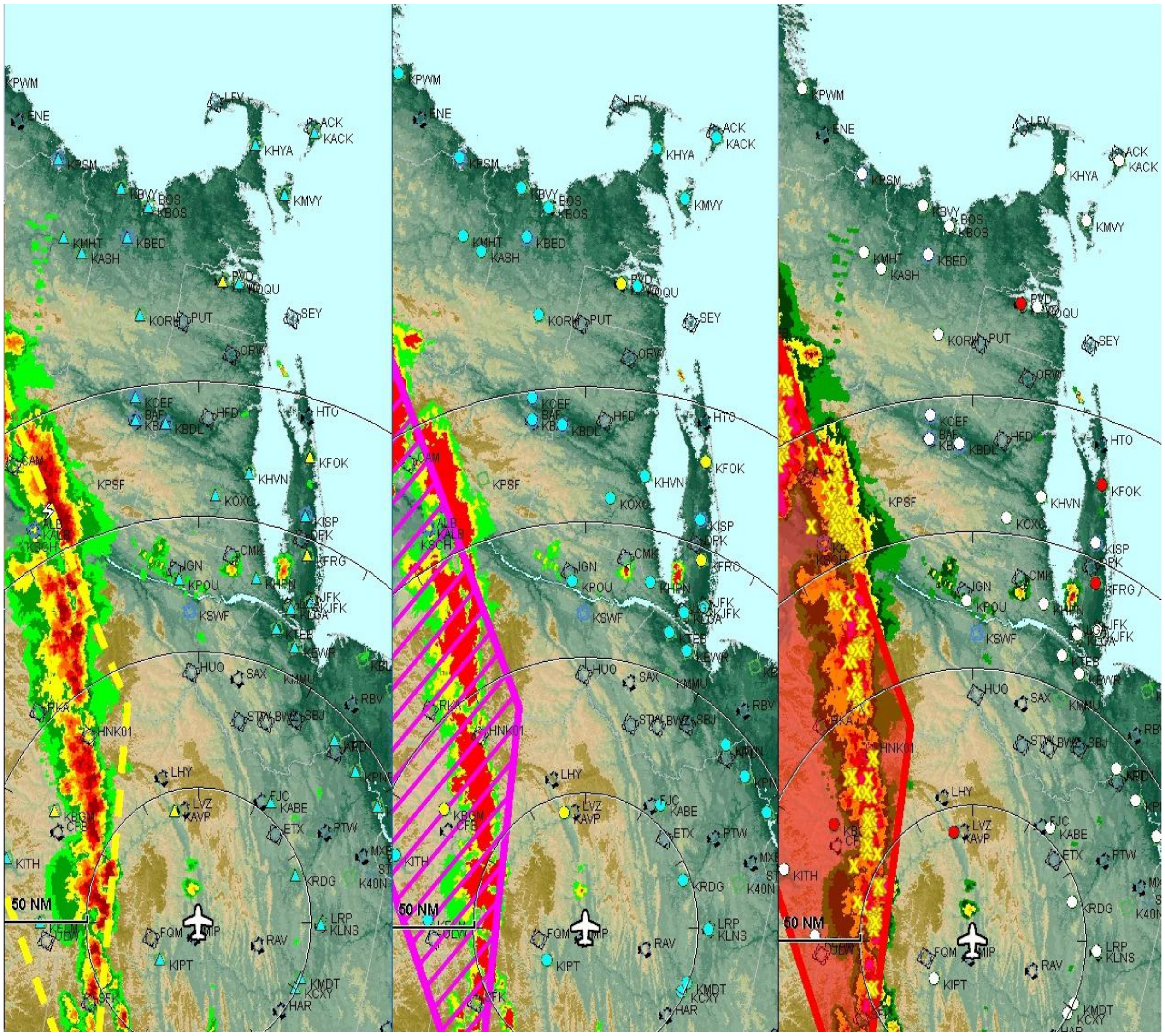


Figure 3. The three weather presentations used by Group 1 (left), Group 2 (middle), and Group 3 (right).

2.4.1.2. Precipitation variations

Precipitation based on radar information depicts the intensity of precipitation overlaid on the active map. This data updates every 5 minutes, on average. Each of the three presentations differs on the number of color codes for intensity. The Group 1 and Group 3 presentations display nine colors for precipitation intensities while the Group 2 presentation uses only five colors.

2.4.1.3. METAR variations

Our simulation also depicts higher level weather information. The first of these is the METAR, which refreshes every 5 minutes. However, the METAR information generally comes out approximately every hour, so while the display refreshes every 5 minutes the METAR information might be up to an hour old. As Table 4 shows, the METAR can be broken down into four weather parameters that are based on the flight category.

Table 4. Flight Categories

Category	Ceiling		Visibility
Low Instrument Flight Rules **LIFR***	below 500 feet AGL	And/Or	less than 1 mile
Instrument Flight Rules **IFR**	500 to below 1,000 feet AGL	And/Or	1 mile to less than 3 miles
Marginal Visual Flight Rules **MVFR**	1,000 to 3,000 feet AGL	And/Or	3 to 5 miles
Visual Flight Rules **VFR+**	greater than 3,000 feet AGL	And/Or	greater than 5 miles

Note. Table adapted from Federal Aviation Administration, & National Oceanic and Atmospheric Administration. (2010). Aviation weather services (AC 00-45). Oklahoma City, OK: FAA & NOAA.

2.4.1.4. SIGMET variations

The SIGMET (SIGnificant METeorological Information) advises of weather potentially hazardous to all aircraft. These advisories are divided into two categories: nonconvective and convective.

The first category is *Nonconvective SIGMETs.*

> Nonconvective SIGMETs advise of nonconvective weather that is potentially hazardous to all aircraft. SIGMETs are unscheduled products that are valid for 4 hours. However, conditions that are associated with hurricanes are valid for 6 hours. Unscheduled updates and corrections are issued as necessary. In the conterminous U.S., SIGMETs are issued when the following phenomena occur or are expected to occur: (a) Severe icing not associated with thunderstorms, (b) Severe or extreme turbulence or clear air turbulence (CAT) not associated with thunderstorms, (c) Dust storms or sandstorms lowering surface or inflight visibilities to below 3 miles, and (d) Volcanic ash. (FAA, 2012).

The second category is *Convective SIGMETs.*

> Convective SIGMETs (WST) are issued in the conterminous U.S. for any of the following: (a) Severe thunderstorm (due to Surface Winds greater than or equal to 50 knots, Hail at the surface greater than or equal to ¾ inches in diameter, and Tornadoes), (b) Embedded thunderstorms, (c) A line of thunderstorms, and (d) Thunderstorms producing precipitation greater than or equal to heavy precipitation affecting 40 percent or more of an area at least 3,000 square miles. Any convective SIGMET implies severe or greater turbulence, severe icing, and low-level wind shear. A convective SIGMET may be issued for any convective situation that the forecaster feels is hazardous to all categories of aircraft. (FAA, 2012).

The SIGMET information updates every 4 hours, unless a hurricane is present then it is updated every 6 hours. Convective SIGMETs are updated hourly. However, as stated previously, our presentation updates every 5 minutes regardless of new information. Each of the three presentations (Group 1, Group 2, and Group 3) depicts the SIGMET in different ways (as shown in Table 3). The Group 1 presentation uses a yellow dashed line, Group 2 shows a magenta solid outline (filled with magenta hash marks), and Group 3 shows a red solid outline (semitransparent red infill).

2.4.1.5. Lightning variations

Pilots need to be aware of lightning strikes in the area where they are flying. For the presentations, the lightning information is updated every 1-2 minutes. All three variations present lightning information in different ways. The Group 1 presentation shows lightning information by a "lightning bolt" symbol, Group 2 uses a magenta dot, and Group 3 uses a yellow "X."

2.4.2. Dependent Variables

To evaluate whether variations in weather presentation symbology affect pilot decision-making and flight behavior, we captured dependent variables in the following categories: Decision-making (weather deviations), Weather Avoidance (distance to weather), Weather Presentation Usage (zoom levels), Communication (pilot/ATC PTT), System Performance (altitude and heading), Eye Movements—fixations and saccades on Areas of Interests (AOIs)—and Workload (fNIR oxygenation changes). In Table 5, we provide a list of the dependent measures and a short description.

Table 5. Dependent Variables

#	VARIABLE	DESCRIPTION
1	Weather deviations	An important assessment in this study is whether presentation symbology affects pilot weather-deviation behavior. In the absence of weather, pilots can fly a straight-line path from start to destination. However, in both flight scenarios, pilots are presented with weather data that GA pilots must take into consideration when choosing their course. Therefore, an effective deviation measure is the lat/long difference between a straight-line flight from start to destination and the actual course flown by pilots. Any difference in the lat/long positions is due to the pilot's decision to deviate around unfavorable weather conditions. During the test scenarios, the aircraft's actual lat/long position is recorded and compared to the lat/long position for a straight-line course (once every second). The difference between these two lat/long positions provides a weather deviation distance.
2	Distance to weather	In addition to a pilot's weather-deviation behavior, we want to assess how close each pilot flies in relation to intense precipitation areas. For this assessment, we chose the "yellow"' precipitation intensities as they are the same and start at 30 dBZ in the Groups 1-3 weather presentations. In Scenario 2 and Scenario 3, there are two different precipitation areas that we will measure the distances to from the aircraft position. The first precipitation areas are the smaller precipitation "cells" located in front of the aircraft at scenario start-up. The locations of these smaller cells are shown in Figure 5 for Scenarios 2 and 3. The second precipitation area in each scenario is the "line" of precipitation to the West of the aircraft's starting position, also shown in Figure 5. We will report the mean distances to weather for Groups 1-3 from the aircraft position as the mean distance to the cells and the line. During each scenario run, we will measure the closest distance from the aircraft position to the cells and the line once a second. Current guidelines state hazardous weather should be avoided by at least 20 statute miles (FAA & NOAA, 1983).

(table continues)

Table 5. Dependent Variables (continued)

#	VARIABLE	DESCRIPTION
3	Weather presentation zoom	During each simulation run, the pilot can adjust the zoom level for the weather presentation. Three different zoom levels are available; Zoom 1 = 5 nautical miles (nmi), Zoom 2 = 20 nmi, and Zoom 3 = 50 nmi. The weather zoom distance will always be displayed on a bar (in nmi) in the lower right corner of the display, and as range rings around the aircraft position symbol.
4	Pilot/ATC communications	The number and duration of pilot/controller PTT communications. Communications can be indicators of workload, as well as providing detailed weather and flight inquires by pilots.
5	Altitude and heading changes (system performance measures)	Aircraft position, mean altitude, and mean heading during a scenario.
6	Eye-movement metrics	The number and duration of fixations, the number and distance for saccades, the pupil diameter, and the eye movement workload derived from the eyetracker POG recordings.
7	Visual areas of interest (AOI)	The number and durations of fixations and the number and saccade distances on the cockpit out-the-window, glass display, weather presentation, and cockpit console AOIs.
8	Workload	The oxygenation changes captured by the fNIR system.

Note. GA = General Aviation; lat/long = Latitude/Longitude; FAA = Federal Aviation Administration; NOAA = National Oceanic and Atmospheric Administration; PTT = Push-To-Talk; POG = Point of Gaze; fNIR = Functional Near-Infrared.

2.4.2.1. System performance measures

During the simulation flights, several parameters from the cockpit simulator system (Flight Simulator 2004 and the Magenta Instructor Station) were recorded to calculate a number of dependent measures that are associated with pilot flight behavior. We used the following dependent measures:

1. Altitude - The height of the airplane above mean sea level as displayed on the altimeter.
2. Heading - The direction where the airplane is pointed.
3. Playback timer and time compression setting - The time of each event in the overall flight scenario.

2.4.2.2. Video and audio recordings

All simulation runs were audio and video recorded. By recording participant runs on video, researchers have the ability to reexamine simulation flight events. In addition, we recorded the PTT communication between the pilot and the controller (ATC SME).

2.4.2.3. Eye-movement data

We recorded the POG data to derive the fixations and saccades on several predefined AOIs associated with the actual weather presentation, the cockpit out-the-window view, the glass cockpit display, and the cockpit console. This is an important assessment in light of previous findings that some pilot weather presentations cause suboptimal viewing behavior (Beringer & Ball, 2004). The dependent measures that we derived from the POG data were the number of fixations, fixation durations, number of saccades, saccade distance, pupil diameter, and eye-movement workload (EMW). The EMW is a common measure used in research of ATC controller visual scanning (Stein, 1992; Willems, Allen, & Stein, 1999). For the definition and calculation of fixations and saccades, see Ahlstrom and Friedman-Berg (2006).

2.4.2.4. Functional near- infrared spectroscopy

The pilots wore an fNIR forehead-sensor pad during simulation flights to measure changes in the relative ratios of deoxygenated hemoglobin and oxygenated hemoglobin during brain activity.

2.4.2.5. Zoom levels

The pilots had the ability to zoom in on weather areas or routes via a button next to the presentation panel. The pilot could choose from three different Zoom levels: Zoom 1 = 5 nmi, Zoom 2 = 20 nmi, and Zoom 3 = 50 nmi. For all simulation flights, we recorded each zoom activation and its associated duration.

2.5. Procedure

2.5.1. IRB Informed Consent Form and Biographical Information

Prior to the simulation, each pilot read and signed an Institutional Review Board (IRB) Informed Consent form, which describes the study, data confidentiality, and participant rights. Participants had the right to withdraw from the study at any time. After this, we asked the pilots about their flight experience including age, number of hours flown, number of instrument hours, and weather presentation experience.

2.5.2. Flight Briefing

A pilot SME provided a short briefing of what to expect during the simulation. During this briefing, each pilot was given the specifics of the *proposed routes*, airspace, surrounding VORs, and airports (the pilot had *planned* the routes during the previous day). Each pilot was briefed about the weather by the SME and was given a convective SIGMET to read. This convective SIGMET stated that a "significant squall-line thunderstorm associated with a fast moving cold front extending from Pittsburgh to Maine is moving southeast at 25 knots. This front has scattered, moderate to heavy thunderstorms and IFR areas below 200 AGL." Pilots were told that they were aware of the weather and had received a preflight weather briefing but had chosen to proceed with their flight plan from Milton VOR (MIP) to Bradley International Airport (KBDL) as depicted on a sectional map. Pilots were told to fly the scenario as if they were flying Visual Flight Rules (VFR) with two passengers aboard while receiving flight following from New York Center. For purposes of the study, the ATC at New York Center had a radar malfunction, making ATC unable to provide vectors around the weather. The pilots had to avoid the weather on their own using the cockpit weather presentation. Pilots were also informed that their NEXRAD weather presentation had no time stamp, so their precipitation information could be up to 10-17 minutes old (just like the real world). The SME instructed the pilots that they had to use the autopilot while flying and that they were not bound to their original *proposed* route. However, pilots had to stay in the air at all times and could not turn around or land. They were also instructed to obey the Federal Aviation Regulation (FAR) to the best of their ability, and if encountering IMC, they were instructed to contact ATC to request an IFR flight plan. Finally, the SME stressed that although the pilot would encounter weather along the route, there would always be several safe alternatives for deviations around the weather. Pilots would never risk being "boxed in" by adverse weather conditions.

2.5.3. Simulator Briefing

In addition to the regular briefing about the flight procedures and the scenarios, the SME informed pilots about the constraints of the simulation system. First, pilots were told that flying the Micro-Jet simulator was *a little bit* different from flying a Cessna 172. However, this difference would be unimportant because pilots would be using the autopilot during flight and there would be no take-offs or landings. Second, pilots were instructed that the out-the-window view would not always correspond fully with the weather depiction on the cockpit weather presentation. Again, this discrepancy would have no major importance because pilots would encounter IMC early into their flight. Pilots were instructed to always rely on their cockpit weather presentation. If the cockpit weather presentation showed a precipitation cell 10 miles ahead of the aircraft, which did not fully correspond with the out-the-window view, pilots were instructed to always rely on the weather presentation and act accordingly. As an initial orientation of the cockpit weather display, pilots were briefed on the weather data elements by reviewing the weather presentation cheat sheet.

At the end of this briefing, the SME demonstrated the basic aircraft controls on the cockpit console. The SME instructed the pilot how to change radio frequencies and how to perform autopilot operations. Pilots were also instructed not to touch or move a sensor for the eyetracker system—located on the cockpit windshield—and not to move the fNIR headband and eyetracker headset, once placed on the pilot and calibrated by the researchers. After this briefing, the SME and the pilot relocated from the briefing room to the Micro-Jet cockpit simulator.

2.5.4. Simulation Scenarios

2.5.4.1. Practice scenario – Scenario 1

During this initial training session, participants flew a practice scenario designed to allow participant familiarization with the Cessna 172 simulator and the weather display. The pilots took off from Atlantic City Airport (ACY) and flew in the vicinity of NJ and PA to get used to the cockpit controls and autopilot mode. During the practice run, the pilot SME was seated next to the participant to answer questions and guide participants as they operated the flight controls. The SME also demonstrated the weather presentation and how to use the three different zoom levels. As part of the weather presentation training, pilots were instructed to fly a course that depicted the larger weather scenario on the cockpit weather display. Pilots were asked to point out the four weather data elements on the weather presentation and to explore the different views from each zoom level. This practical weather presentation training continued until the pilot SME was assured that pilots fully understood and could demonstrate the use of the weather presentation and the zoom levels. No data collection occurred during the practice scenario.

2.5.4.2. Test scenarios – Scenarios 2 and 3

After completion of the practice scenario, pilots were asked to fly a route from the MIP in Pennsylvania to KBDL in Connecticut (see Figure 4). The pilots were aware that they could not reach their destination within the time allotted due to the speed restrictions of the Cessna 172. The flight was therefore broken into two separate flight scenarios, Scenario 2 and Scenario 3. The presentation order of these two test scenarios was counterbalanced within each group. While half the pilots in a group flew Scenario 2 first and then Scenario 3, the remaining pilots in the group flew the two scenarios in the opposite order. Each flight scenario lasted for 25 minutes.

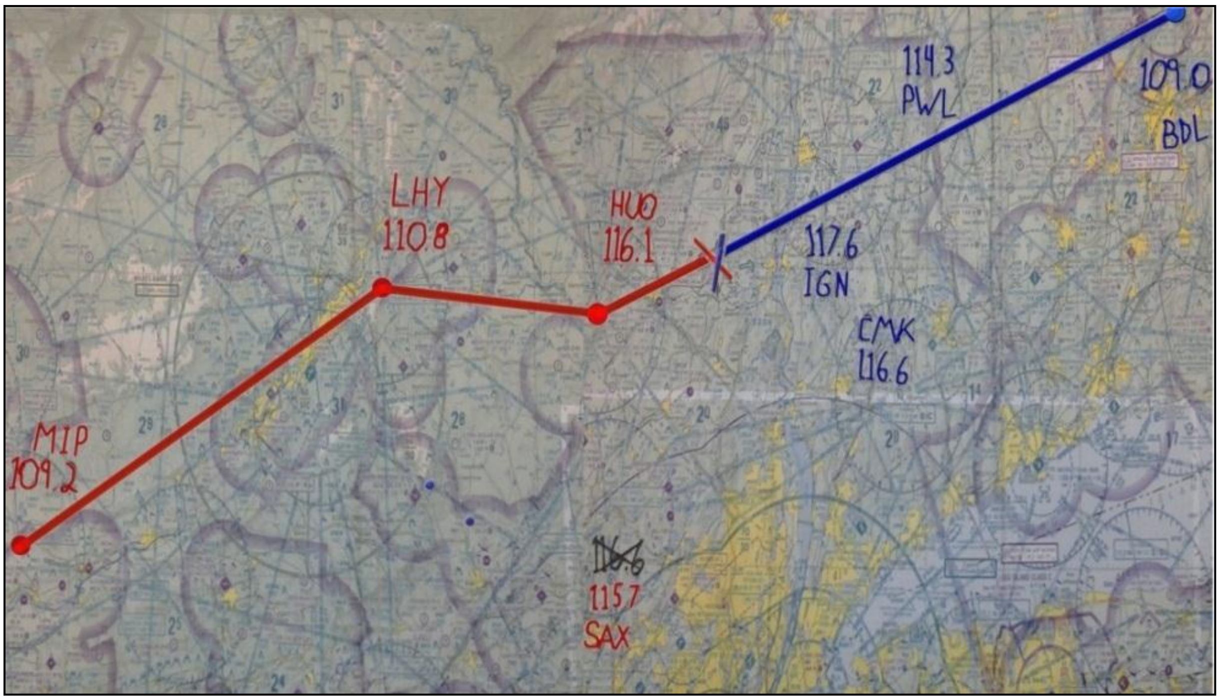


Figure 4. The routes for Test Scenario 2 (red) and Scenario 3 (blue).

2.5.4.2.1. Scenario 2

Scenario 2 was the first leg of the route; the pilots flew from MIP over the Lake Henry VOR (LHY) and up to the Huguenot VOR starting at an altitude of 2500 ft (after 5 minutes, the altitude increased due to an IFR clearance). It is depicted in red in Figure 4, and it is illustrated on the weather presentation in Figure 5. During this scenario, the pilots encountered precipitation cells while between MIP and LHY. The pilots were unable to make the flight all the way to the destination KBDL within the allotted time.

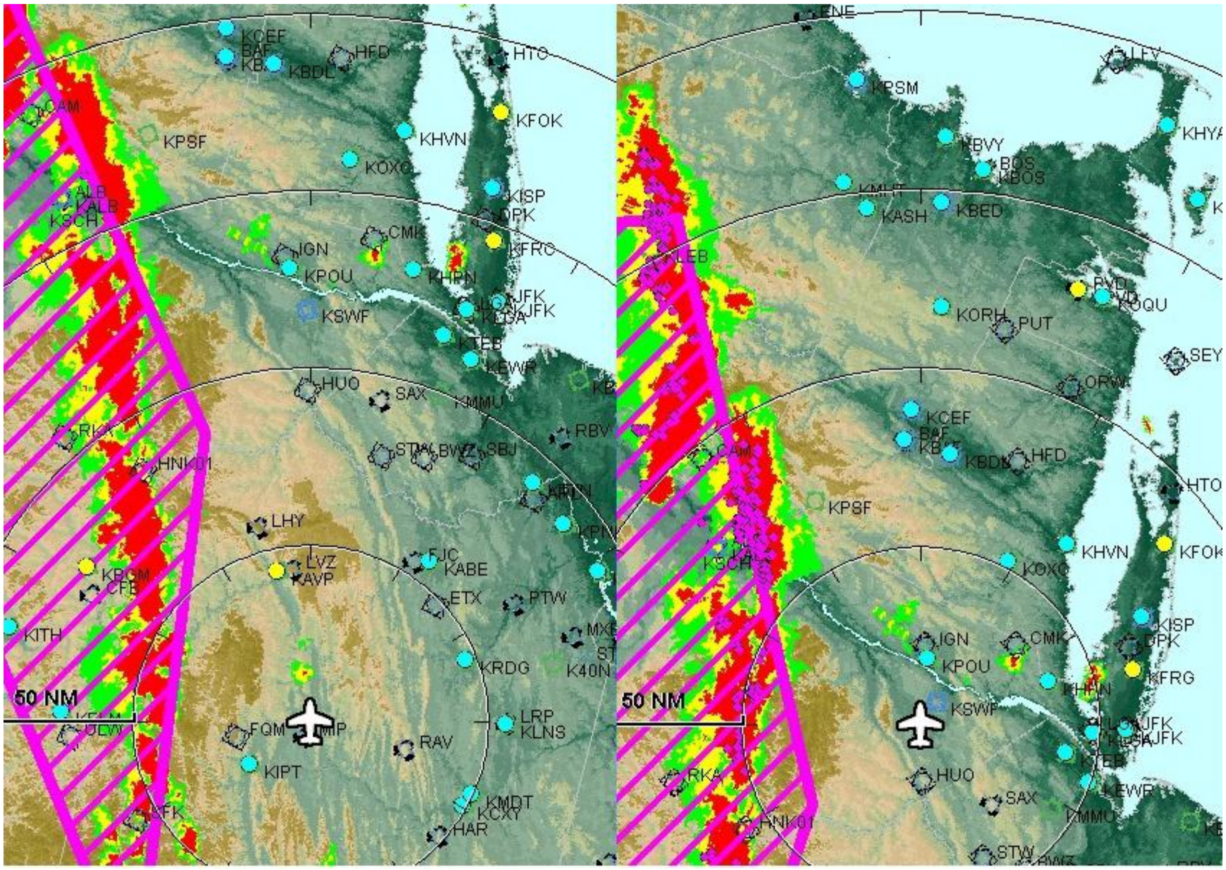


Figure 5. The aircraft position at scenario start-up and the Bradley International Airport (KBDL) destination for Scenario 2 (left) and Scenario 3 (right).

2.5.4.2.2. Scenario 3

Scenario 3 was the second leg of the route; the pilots flew from the Huguenot (HUO) VOR up to KBDL starting at an altitude of 5500 ft (after 5 minutes, the altitude changed due to an IFR clearance). This route was illustrated in blue in Figure 4, and it is illustrated on the weather presentation in Figure 5. In this scenario, pilots encountered a small line of precipitation cells while between HUO and Dutchess County Airport (KPOU) in Poughkeepsie, NY. Like Scenario 2, the pilots were unable to make the flight all the way to the destination KBDL within the allotted time.

2.5.5. Data Collection Procedures

Before the start of each flight scenario, the pilots were fitted with the fNIR and the eye-tracking equipment, which they wore during their simulation flights. Next, the researchers calibrated the Eyetracker by means of a standard nine-dot matrix presented on the cockpit glass display. Once the Eyetracker was calibrated and the pilot was ready to fly, the simulation scenario began. After the flight scenario ended, the participants completed the Weather Presentation Questionnaire (see Appendix A). After the questionnaire, the research team had a final debriefing with the pilots. The pilots were asked about the simulation in general and to comment on the usability of the weather presentation elements, including the background map with range rings. Finally, pilots were asked about any enhancements in procedure or cockpit presentation that would be useful for future simulations.

2.5.6. Data Analysis

The core tenets of NextGen are operational and safety improvements for all aspects of the future NAS. Assessing improvements amounts to finding effects and measuring the sizes of those effects. In some cases, empirical data can show an operationally important effect on operator performance. In other cases, empirical data can highlight the absence of an effect on operator or system performance. To be meaningful, this NextGen effect analysis requires credible and consistent methods. Traditionally, researchers have used the Null Hypothesis Significance Testing (NHST) framework. The goal of this framework is to unveil "statistically significant" effects by rejecting a null hypothesis of no effect. Unfortunately, the NHST testing framework is asymmetric; researchers can only reject the null but never accept evidence *for* the null. Using the NHST framework, researchers can only infer the probability (e.g., p = .021) of getting an outcome, or more extreme values, given that the null hypothesis is true, or $p(\text{data}\,|\,\textit{null})$. Regrettably, this implies that significance tests overstate the evidence *against* the null. We need to assess the *null* hypothesis (no effect) against an *alternative* hypothesis (effect), and not consider the null in isolation. In addition to being asymmetric, the NHST framework is also inconsistent. If the null is false, significance tests are consistent and the probability of correctly rejecting the null increases with an increasing number of observations. If the null is true, however, the inability to infer evidence *for* the null implies that NHST significance tests will mistakenly reject the null and overstate the evidence *against* the null (despite an increasing number of observations). Asymmetric and inconsistent analysis methods strike us as very troublesome, and they provide a feeble foundation for NextGen research.

In this study, we perform Bayesian model comparison on outcome variables for the three pilot groups. Because this study is specifically tailored for an assessment of the presence of effects (i.e., differences in outcome variables between the three groups), we use Bayesian model comparison. In the model comparison approach, we use a statistic called the Bayes factor (*BF*). The Bayes factor is the overall likelihood of the data for one model relative to the overall likelihood of the data for the other model. Bayesian parameter estimation, on the other hand, provides more informative results with complete distributions of credible values for group means, effect sizes, standard deviations, and data normality. However, parameter estimation is more useful in situations where we are assessing the operational impact of effects. That is, when the specific goal of the study is to use parameter estimates for implementing or changing current practices in relation to some operationally important factor.

We are comparing two models: the null and the alternative. The null model states that the difference of the means is zero, whereas the alternative model states that the difference of the means is not equal to zero. Using Bayesian statistics, we can compute the probability of a hypothesis conditionally on observed data, or $p(\textit{hypothesis}\,|\,\text{data})$. More important, unlike the NHST framework, Bayesian tests allow the researcher to state evidence *for* the null. For our model comparisons, we use the Jeffrey-Zellner-Siow (JZS) Bayes Factor (*BF*) *t*-test developed by Rouder, Speckman, Sun, Morey, and Iverson (2009). In doing so, we first compute the traditional *t*-statistic that, together with the number of observations, are used in the formula to compute the JZS *BF*. For all tests, we use the objective JZS prior (Cauchy distribution on effect size) with the scale factor, r, set to 1. To see how the conventional *t*-statistic, the number of observations, and the degrees of freedom are used in the JZS Bayesian model comparison *t*-test, see the formula in Rouder et al. (2009). For simplicity, we will refer to this JZS Bayes factor as *BF* throughout the result section. In using Bayesian model comparison for the *null* versus the *alternative* model, we have the formula,

$$\underbrace{\frac{p(\text{Model}_{null}|data)}{p(\text{Model}_{alternative}|data)}}_{\text{Posterior odds}} = \underbrace{\frac{p(\text{Model}_{null})}{p(\text{Model}_{alternative})}}_{\text{Prior odds}} \, x \, \underbrace{\frac{p(data|\text{Model}_{null})}{p(data|\text{Model}_{alternative})}}_{\text{Bayes factor } (BF)}$$

where the Posterior odds = Prior odds x *BF*. The *BF* is directly interpretable as an odds ratio in the form *null/alternative*. Therefore, a *BF* greater than 1 favors the null, and a *BF* less than 1 favors the alternative. A *BF* of exactly 1 prefers neither the null nor the alternative. As another example, when the *BF* is 1.5, this indicates that the data are 1.5 times more likely to have occurred under the Model_{null} than under the $\text{Model}_{alternative}$. In addition to allowing inferences about the presence of effects, the *BF* can be expressed in terms of categories of evidential strength (see Table 6); these categories were as originally proposed by Jeffreys (1961).

Table 6. Classification Scheme for the Bayes Factor

Bayes factor		BF	Interpretation
>		100	Extreme evidence for the null (no effect)
30	-	100	Very strong evidence for the null
10	-	30	Strong evidence for the null
3	-	10	Substantial evidence for the null
1	-	3	Anecdotal evidence for the null
	1		No evidence
1/3	-	1	Anecdotal evidence for the alternative (effect)
1/10	-	1/3	Substantial evidence for the alternative
1/30	-	1/10	Strong evidence for the alternative
1/100	-	1/30	Very strong evidence for the alternative
<		1/100	Extreme evidence for the alternative

Note. Table adapted from "Why Psychologists Must Change The Way They Analyze Their Data: The Case of Psi: Comment on Bem," by Wagenmakers, Wetzels, Borsboom, and van der Maas (2011). Unpublished manuscript.

Unlike NHST "significance" tests, in Bayesian statistics, there is no drawback or penalty in computing and reporting multiple tests (Morey & Rouder, 2011). Also, because we do not compute *p* values (in the NHST sense), there is no need to make corrections for multiple comparisons (Dienes, 2011; Kruschke, 2010).

We present results for group comparisons on outcome variables. Each data graph presents the mean result for each of the three pilot groups. We compare the groups to see whether the *BF* supports the null or the alternative model. If a comparison favors the null model, it shows evidence that the weather presentation symbology does not have an effect on that particular dependent variable. On the other hand, if a group comparison favors the alternative model, it shows evidence that the weather presentation symbology affects that particular variable. For each comparison, we provide the number of subjects in each group (e.g., $n = 8$), the conventional t-statistic (e.g., $t = 1.12$), and the *BF* (e.g., $BF = 1.95$). In addition, we describe the *BF* for each comparison in terms of its categorical strength (e.g., $BF = 2.8$ provides "anecdotal" evidence for the null).

3. RESULTS

In this section, we present the simulation results in the following categories: System performance (aircraft and instrument panel data), Communication (pilot/ATC PTT), Decision-making (deviations), Weather presentation usage (visual scanning behavior, zoom usage), Weather avoidance (distance to hazardous weather), and Workload (fNIR oxygenation changes).

3.1. Weather Deviations

Figure 6 shows the mean weather deviation distance for Groups 1-3 during Scenario 2 (first leg of route). The "error bars" in Figure 6 represent one standard error (*SE*) of the mean (we used one *SE* of the mean for all subsequent graphs). The mean deviation distances for Group 1 ($n = 8$) and Group 2 ($n = 9$) provided "substantial" evidence for the alternative hypothesis, $t = 3.34$, $BF = 0.10$. The mean deviation distances for Groups 1 and 3 ($n = 8$) were similar, providing anecdotal evidence for the null hypothesis, $t = .82$, $BF = 2.28$. The mean deviation distances for Groups 2 and 3 provided anecdotal evidence for the alternative, $t = 2.08$, $BF = 0.64$.

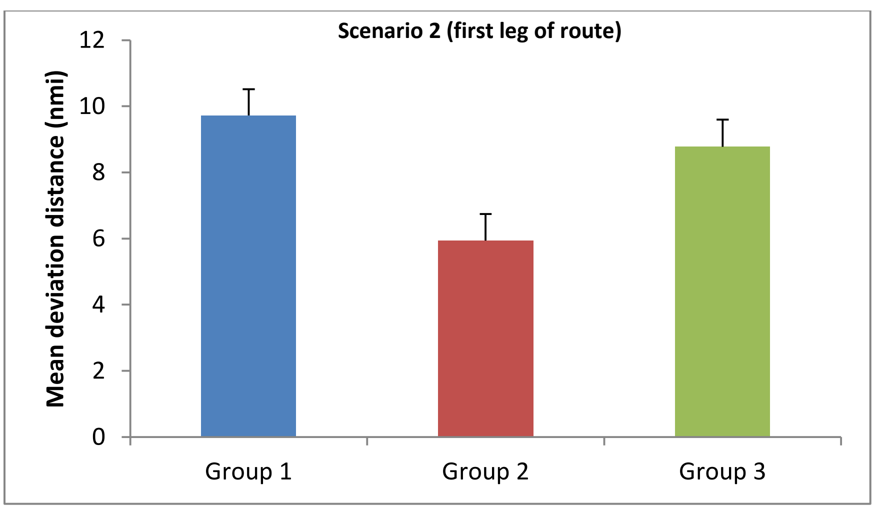


Figure 6. Mean deviation distances for Scenario 2.

Figure 7 shows the mean deviation distance for Groups 1-3 during Scenario 3 (second leg of route). Contrary to the result for Scenario 2, the mean deviation distances were very similar and all group comparisons provided anecdotal evidence for the null: Groups 1 and 2, $t = .46$, $BF = 2.8$; Groups 1 and 3, $t = .17$, $BF = 2.94$; and Groups 2 and 3, $t = .31$, $BF = 2.92$.

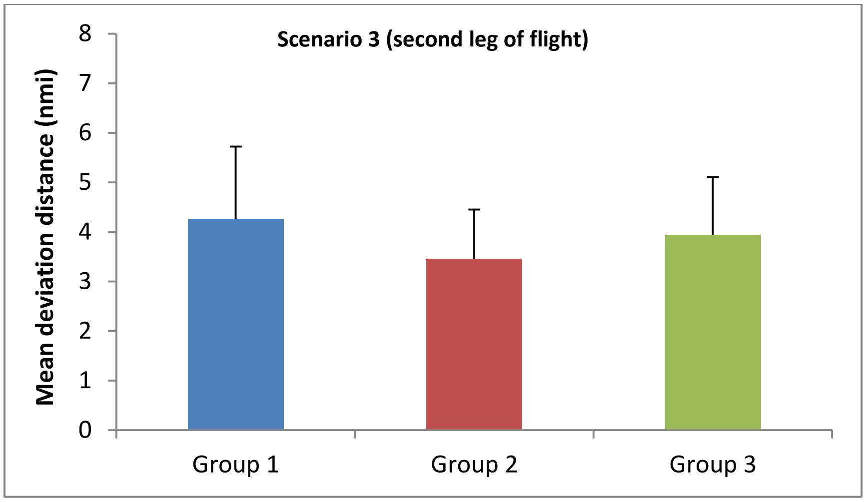


Figure 7. Mean deviation distances for Scenario 3.

3.2. Distance to Weather

Figure 8 shows that pilots had similar behavior in how close they flew to the precipitation cells during Scenario 2. The mean distances to cells for all three groups were roughly 14 nautical miles. All group comparisons of the distances to cells provided anecdotal evidence for the null: Group 1 ($n = 8$) and Group 2 ($n = 9$), $t = .24$, $BF = 2.97$; Groups 1 and 3 ($n = 8$), $t = .57$, $BF = 2.63$; and Groups 2 and 3, $t = .88$, $BF = 2.24$.

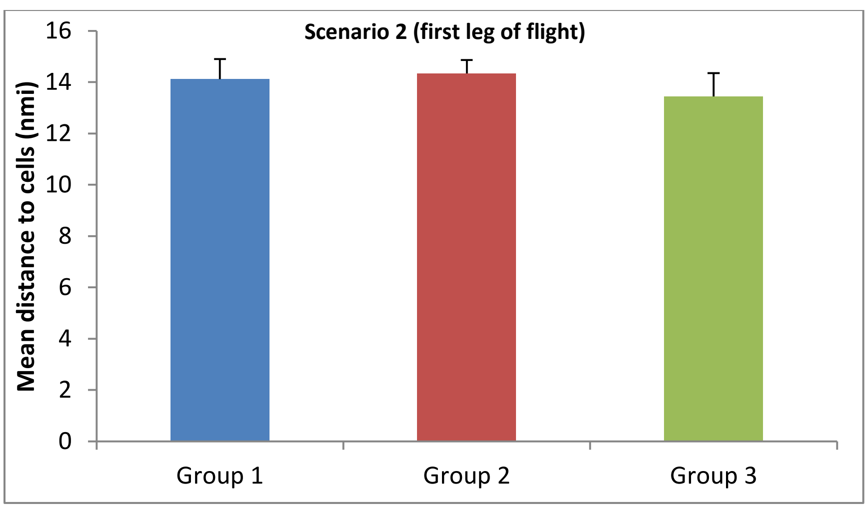


Figure 8. Mean distance to cells during Scenario 2.

The outcome for Scenario 3 (see Figure 9) was similar to the outcome for Scenario 2. The mean distances to cells for all three groups were roughly 15 nautical miles. All group comparisons of the distances to cells provided anecdotal evidence for the null: Groups 1 and 2, $t = .62$, $BF = 2.6$; Groups 1 and 3, $t = .33$, $BF = 2.84$; and Groups 2 and 3, $t = .25$, $BF = 3.0$.

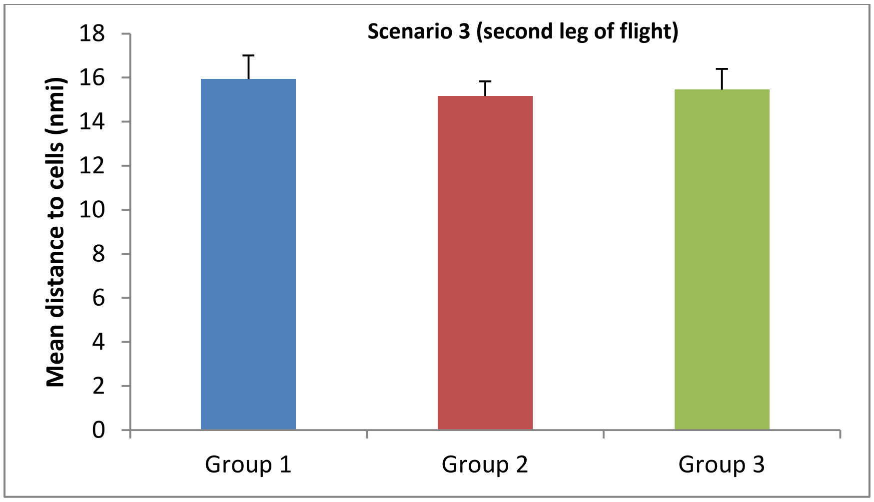


Figure 9. Mean distance to cells during Scenario 3.

The pilots in all groups (Groups 1-3) kept a similar distance from the precipitation *line*. Figure 10 shows the mean distances to the line for Scenario 2. All distances were roughly 40 nautical miles and group comparisons provided anecdotal evidence for the null: Groups 1 and 2, $t = .22$, $BF = 2.98$; Groups 1 and 3, $t = .67$, $BF = 2.50$; and Groups 2 and 3, $t = 1.08$, $BF = 1.95$.

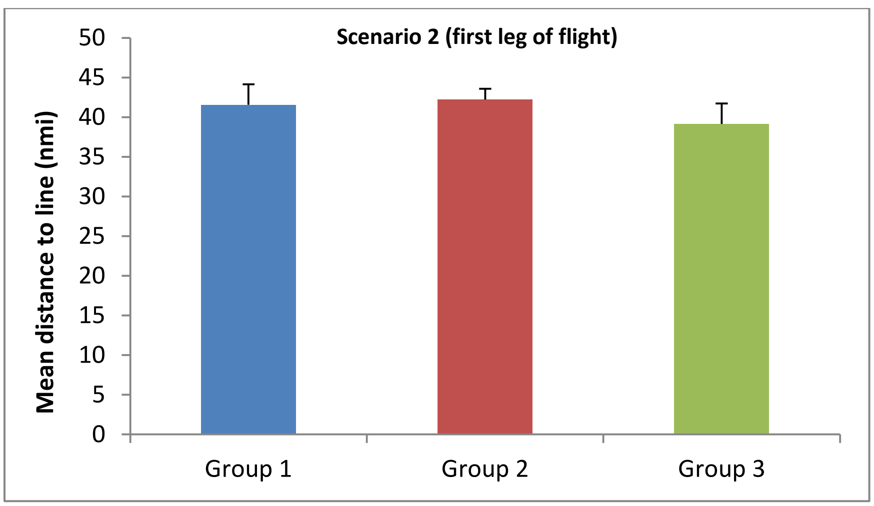


Figure 10. Mean distance to line during Scenario 2.

The result for Scenario 3 (see Figure 11) was similar to the result for Scenario 2. All distances were roughly 40 nautical miles and group comparisons provided anecdotal evidence for the null: Groups 1 and 2, t = .17, BF = 3.0; Groups 1 and 3, t = .82, BF = 2.29; and Groups 2 and 3, t = 1.40, BF = 1.44.

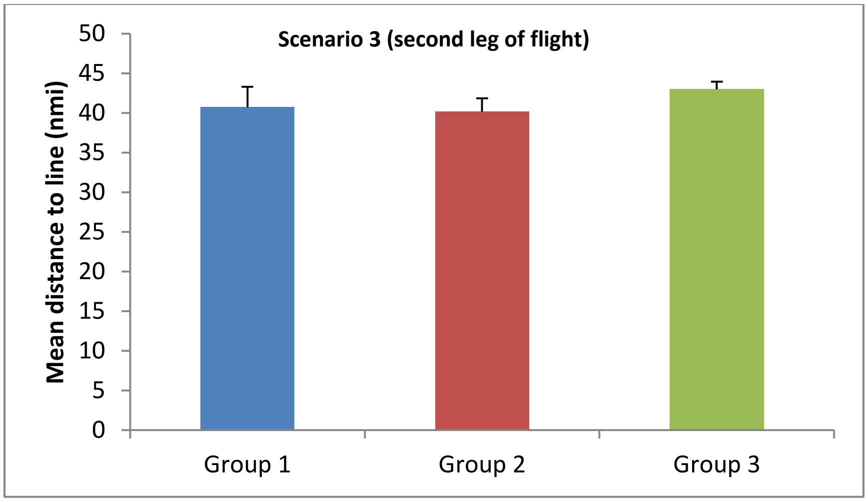


Figure 11. Mean distance to line during Scenario 3.

3.3. Weather Presentation Zoom

Figure 12 and Figure 13 show the mean number of zoom activations and zoom durations for Scenarios 2 and 3, respectively. The zoom usage was very similar among the three groups. Zoom 2 was used most frequently and was displayed for longer durations compared to Zoom 1 and Zoom 3. Due to the small activation numbers for Zoom 1 and Zoom 3, we performed group comparisons for Zoom 2 only.

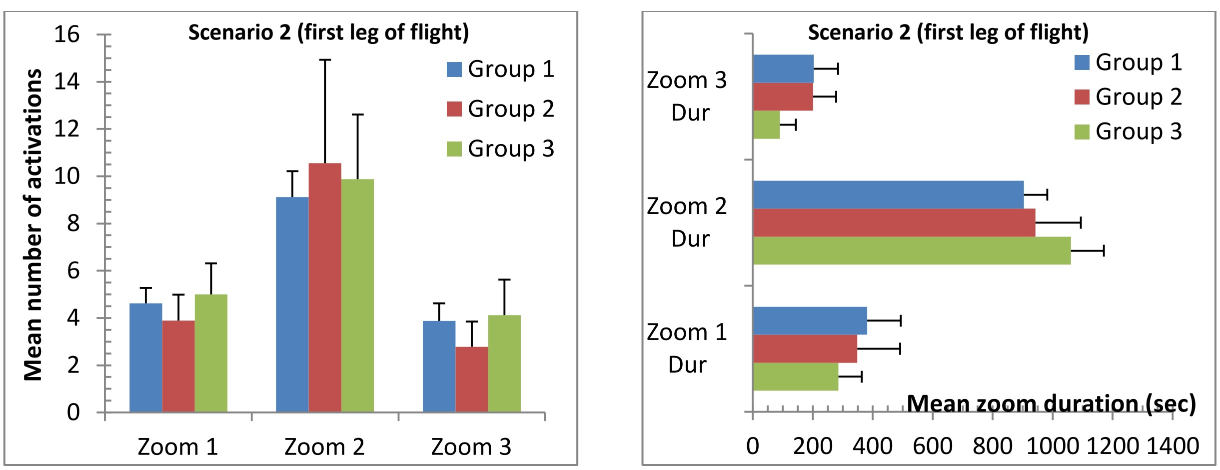


Figure 12. Mean number of weather presentation zoom activations (left) and zoom durations (right) for Scenario 2. Zoom 1 = 5 nmi, Zoom 2 = 20 nmi, and Zoom 3 = 50 nmi.

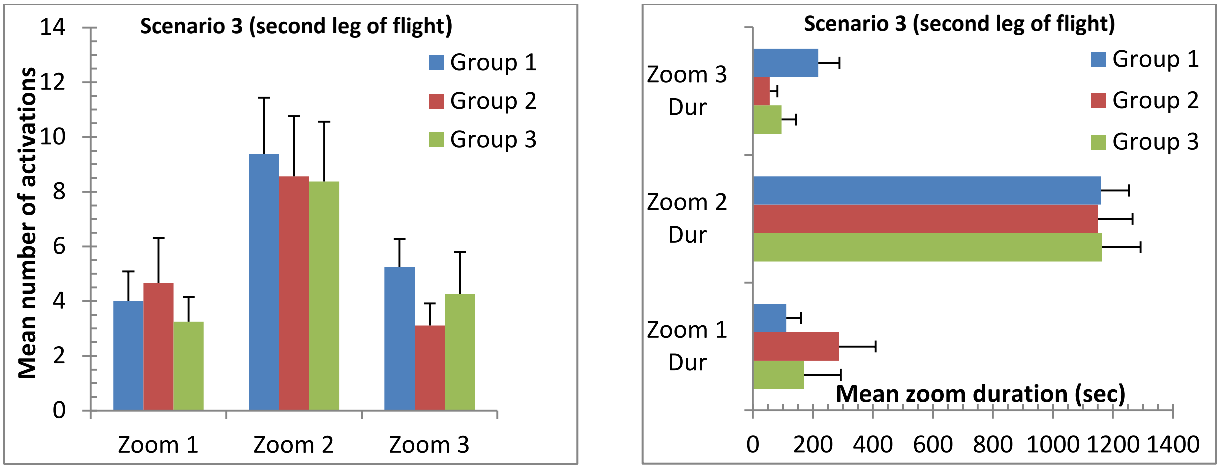


Figure 13. Mean number of weather presentation zoom activations (left) and zoom durations (right) for Scenario 3. Zoom 1 = 5 nmi, Zoom 2 = 20 nmi, and Zoom 3 = 50 nmi.

For Scenario 2, all group comparisons of the mean number of activations provided anecdotal evidence for the null: Groups 1 and 2, $t = 0.30$, $BF = 2.93$; Groups 1 and 3, $t = .25$, $BF = 2.90$; and Groups 2 and 3, $t = .12$, $BF = 3.0$. Similarly, all group comparisons of the mean zoom durations provided anecdotal evidence for the null: Groups 1 and 2, $t = 0.21$, $BF = 2.98$; Groups 1 and 3, $t = 1.17$, $BF = 1.76$; and Groups 2 and 3, $t = .62$, $BF = 2.60$.

For Scenario 3, all group comparisons of the mean number of activations provided anecdotal evidence for the null: Groups 1 and 2, $t = 0.26$, $BF = 2.95$; Groups 1 and 3, $t = .33$, $BF = 2.84$; and Groups 2 and 3, $t= .05$, $BF = 3.0$. In addition, all group comparisons of the mean zoom durations provided anecdotal evidence for the null: Groups 1 and 2, $t = .06$, $BF = 3.0$; Groups 1 and 3, $t = .01$, $BF = 2.97$; and Groups 2 and 3, $t = .07$, $BF = 3.0$.

3.4. Air Traffic Control Communications

During the test scenarios, pilots communicated with New York Center ATC. It was always the pilot who initiated the contact with ATC. In Figures 14, 15, 16, and 17, we summarize the number of contacts (communications) per pilot and we present the computed duration for each communication. Figure 14 shows the mean number of ATC contacts during Scenario 2. Group 1 had a slightly larger number of contacts, but all group comparisons provided anecdotal evidence for the null: Group 1 ($n = 7$) and Group 2 ($n = 9$), $t = 1.38$, $BF = 1.44$; Group 1 and Group 3 ($n = 8$), $t = .98$, $BF = 2.0$; and Group 2 and Group 3, $t = .45$, $BF = 2.8$.

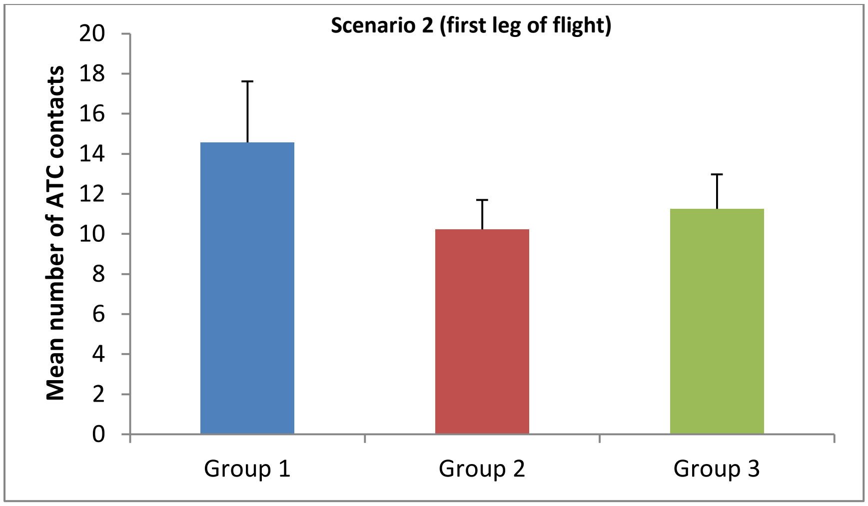


Figure 14. Mean number of ATC communications during Scenario 2.

The result for Scenario 3 (see Figure 15) was very similar to the result for Scenario 2. Group 1 had a slightly higher number of ATC contacts, but all group comparisons provided anecdotal evidence for the null: Group 1 and Group 2, $t = .51$, $BF = 2.74$; Group 1 and Group 3, $t = .78$, $BF = 2.36$; and Group 2 and Group 3, $t = .16$, $BF = 3.0$.

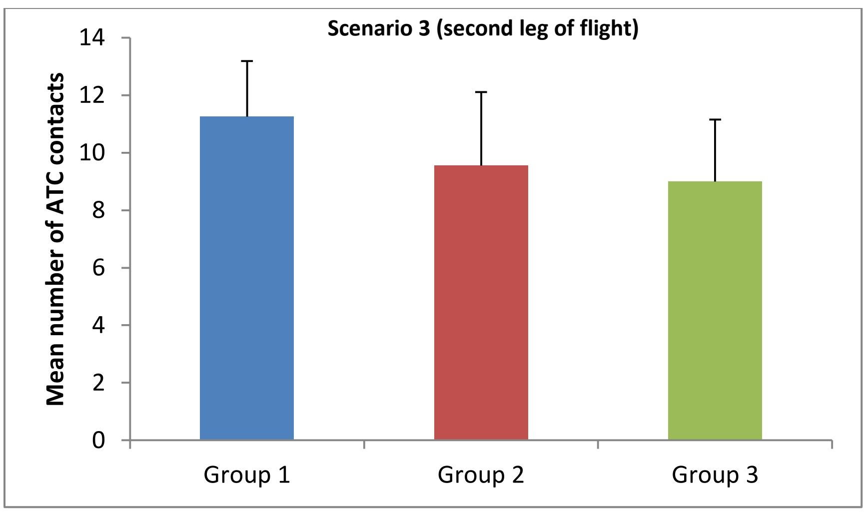


Figure 15. Mean number of ATC communications during Scenario 3.

The duration of each ATC contact was very similar among the three groups. Figure 16 shows the mean pilot/ATC contact duration for Scenario 2. All group comparisons provided anecdotal evidence for the null: Group 1 and Group 2, t = .12, BF = 2.94; Group 1 and Group 3, t = .48, BF = 2.65; and Group 2 and Group 3, t = .73, BF = 2.46.

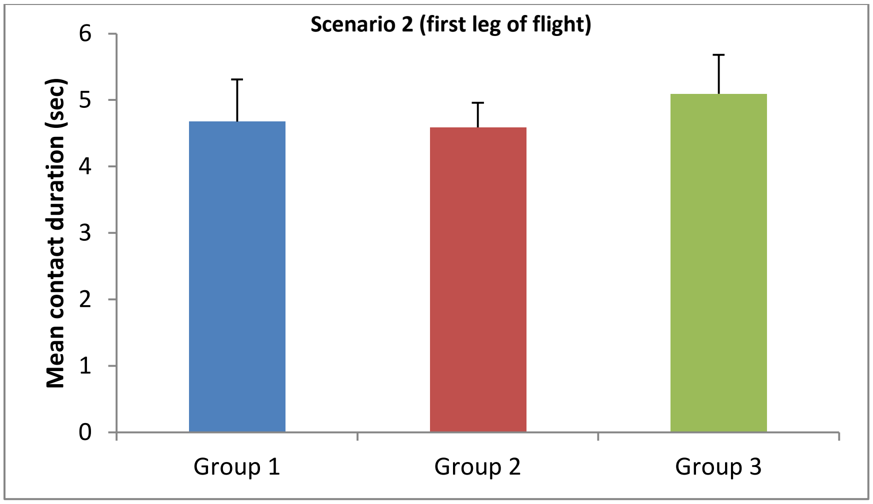


Figure 16. Mean ATC contact duration for Scenario 2.

Figure 17 shows the mean contact duration for Scenario 3. The result was very similar to the mean contact durations for Scenario 2. All group comparisons provided anecdotal evidence for the null: Group 1 and Group 2, t = .36, BF = 2.88; Group 1 and Group 3, t = .58, BF = 2.61; and Group 2 and Group 3, t = .07, BF = 3.0.

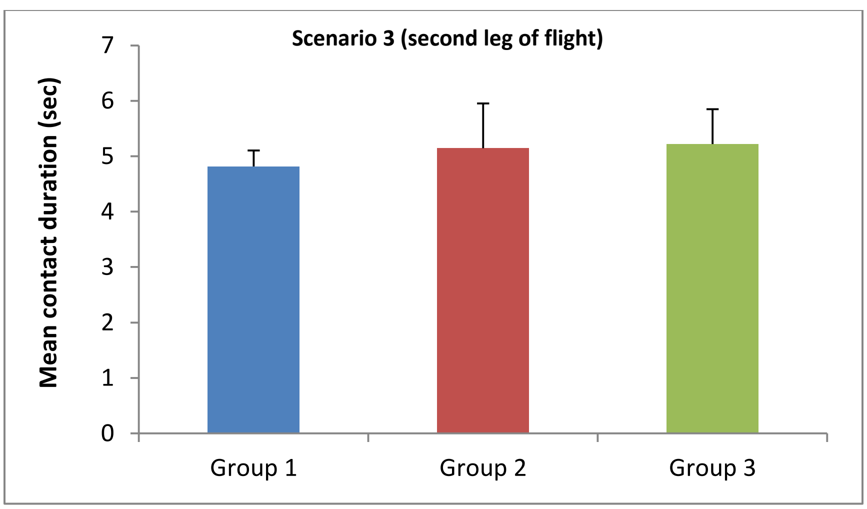


Figure 17. Mean ATC contact duration for Scenario 3.

3.5. Altitude and Heading Changes

During Scenario 2 and Scenario 3, pilots deviated from their "planned" course to avoid precipitation areas. Apart from the deviation distances from a "straight-line" course (from scenario start to scenario destination), we were also interested in analyzing pilot altitude and heading changes. Figure 18 shows the mean altitude for all groups (Groups 1-3) during Scenario 2. The mean altitude was almost identical for all three groups and all group comparisons provided anecdotal evidence for the null: Group 1 (n = 8) and Group 2 (n = 9), t = .20, BF = 3.0; Group 1 and Group 3 (n = 8), t = .001, BF = 2.97; and Group 2 and Group 3, t = .07, BF = 3.0.

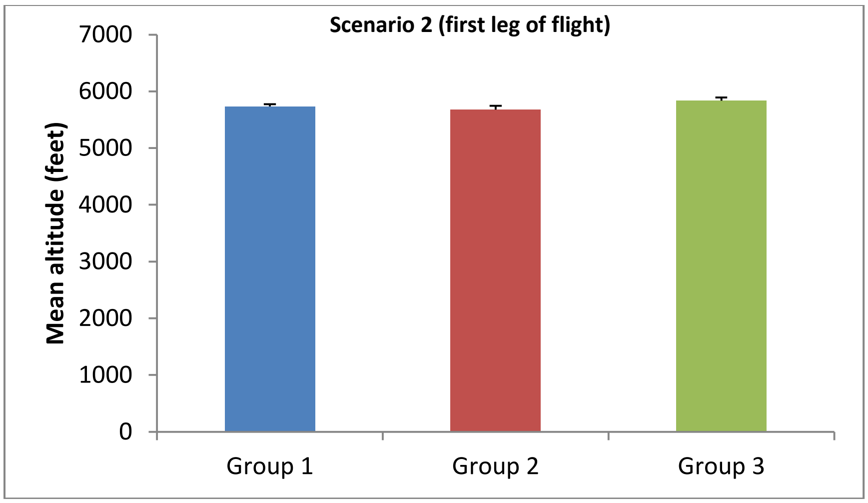


Figure 18. Mean altitude for Scenario 2.

Figure 19 shows the mean heading for Groups 1-3 during Scenario 2. The mean heading was almost identical for all three groups and all group comparisons provided anecdotal evidence for the null: Group 1 and Group 2, t = .07, BF = 3.0; Group 1 and Group 3, t = .002, BF = 2.97; and Group 2 and Group 3, t = .07, BF = 3.0.

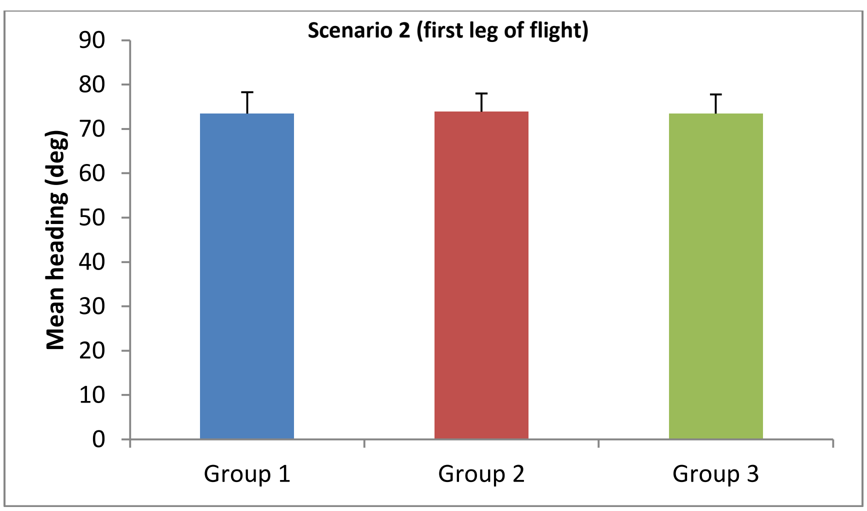


Figure 19. Mean heading for Scenario 2.

Figure 20 shows the mean altitude for Groups 1-3 during Scenario 3. The mean altitude was similar for all three groups and all group comparisons provided anecdotal evidence for the null: Group 1 and Group 2, $t = .77$, $BF = 2.42$; Group 1 and Group 3, $t = .54$, $BF = 2.66$; and Group 2 and Group 3, $t = .31$, $BF = 2.92$.

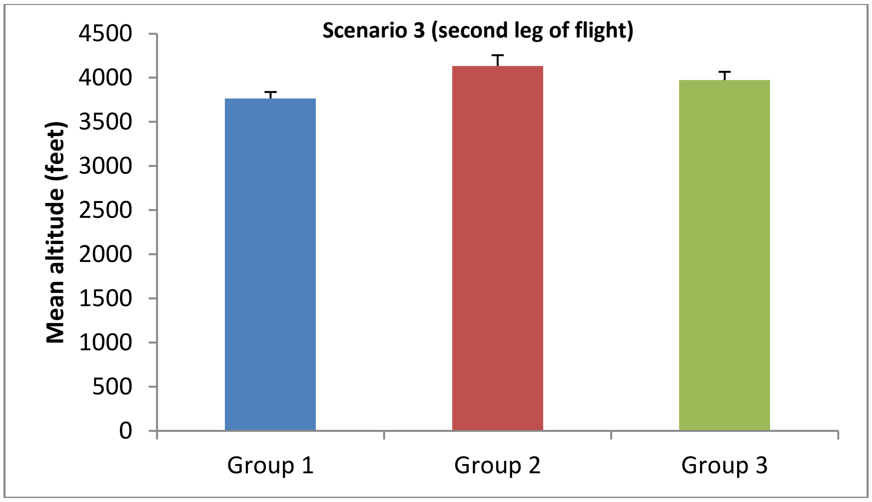


Figure 20. Mean altitude for Scenario 3.

Figure 21 shows the mean heading during Scenario 3. The mean heading was slightly larger for Group 1 compared to Group 2 and Group 3. Group comparisons between Group 1 and Group 2, $t = 1.67$, $BF = 1.08$, and Group 2 and Group 3, $t = 1.04$, $BF = 2.01$, provided anecdotal evidence for the null. However, the comparison between Group 1 and Group 3, $t = 2.08$, $BF = .66$, provided anecdotal evidence for the alternative.

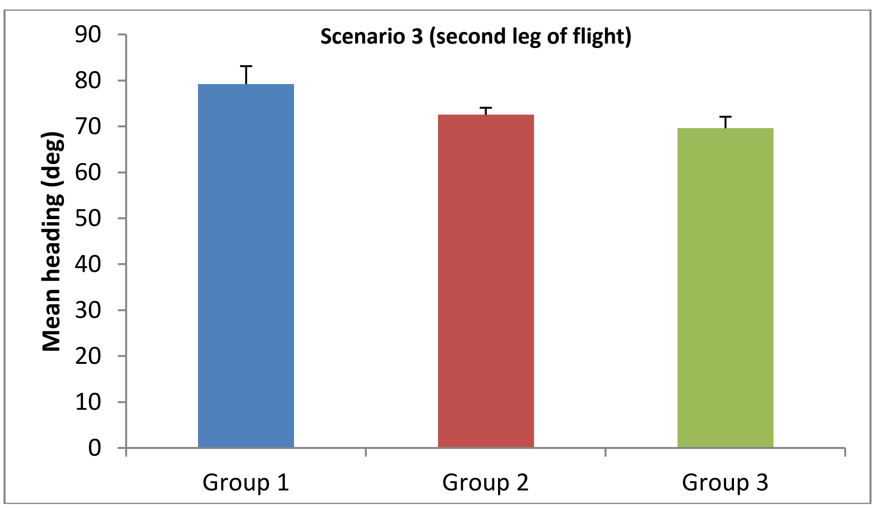


Figure 21. Mean heading for Scenario 3.

3.6. Eye-Movement Metrics

In this section, we analyze and present eye-movement data. Because our eye-tracking system cannot be used with participants wearing eyeglasses, we had only 15 participants that could participate in the eye-movement recordings (15 out of 25). Furthermore, due to hardware failures, we were unable to collect eye-movement data during all simulation runs. Therefore, in the eye-movement result presentation, we indicate the number of pilots (*n*) for each Group that was included in each analysis.

As objective measures of cognitive and visual workload, researchers have used different 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. Previous research has found that blink rate, blink duration, and saccade duration all decrease; whereas pupil diameter, the number of saccades, and the frequency of long fixations all increase with increased cognitive workload (Ahlstrom & Friedman-Berg, 2006). In addition to serving as proxies for cognitive and visual workload, we use these parameters to analyze pilot scanning behavior.

3.6.1. Fixations

Figure 22 shows the mean number for all group (Groups 1-3) fixations during Scenario 2. The mean number of fixations for Group 3 was higher than the number of fixations for Group 1 (n = 3) and Group 2 (n = 6). A comparison between Group 1 and Group 2 showed anecdotal evidence for the alternative, t = 1.83, *BF* = .86. Comparing Group 1 and Group 3 (n = 6) provided anecdotal evidence for the null, t = 1.09, *BF* = 1.56. The comparison of Group 2 versus Group 3 provided anecdotal evidence for the alternative, t = 2.68, *BF* = .32.

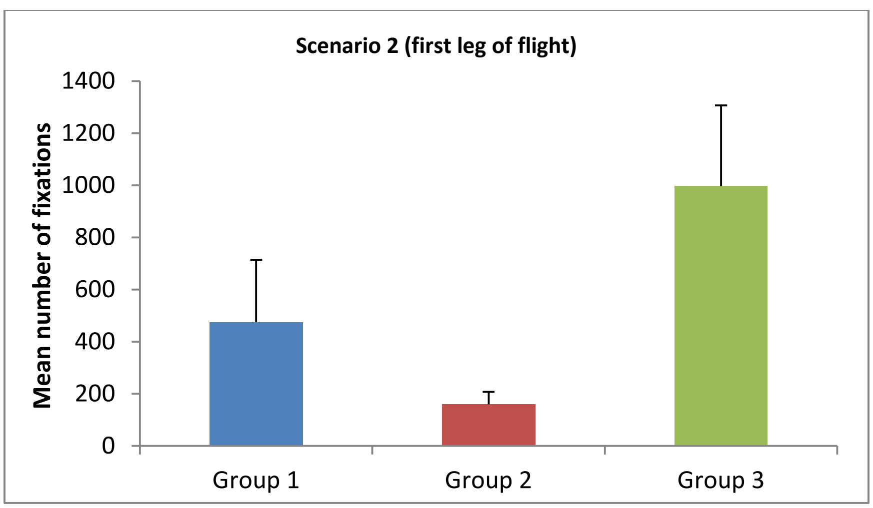


Figure 22. Mean number of fixations for all groups (Groups 1-3) during Scenario 2.

Figure 23 shows the mean number for all group (Groups 1-3) fixations during Scenario 3. Again, Group 3 had the highest mean number of fixations followed by Group 2 and Group 3. However, all group comparisons provided anecdotal evidence for the null: Group 1 and Group 2, t = .65, BF = 1.92; Group 1 and Group 3, t = 1.47, BF = 1.17; and Group 2 and Group 3, t = .52, BF = 2.24.

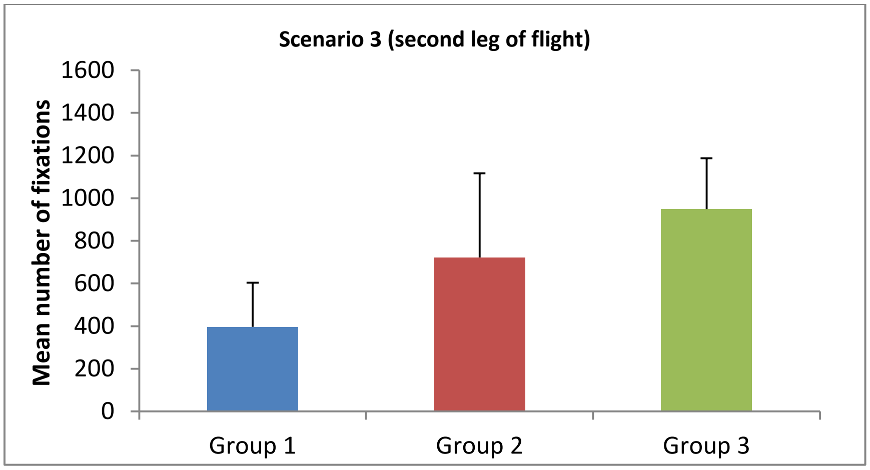


Figure 23. Mean number of fixations for all groups (Groups 1-3) during Scenario 3.

3.6.2. Areas of Interest

In addition, we analyzed fixations on several predefined AOIs. The AOIs corresponded to the area covering the cockpit out-the-window view, the cockpit glass display, the weather presentation, and the cockpit console.

3.6.2.1. Cockpit out-the-window AOI

Figure 24 shows the mean number of fixations on the AOI for the cockpit out-the-window for Scenario 2. The comparison between Group 1 (n = 2) and Group 2 (n = 4) provided anecdotal evidence for the alternative, t = 1.94, BF = .82. The comparison between Group 1 and Group 3 (n =5) showed anecdotal evidence for the null, t = 1.04, BF = 1.5, as did the comparison between Group 2 and Group 3, t = 1.26, BF = 1.41.

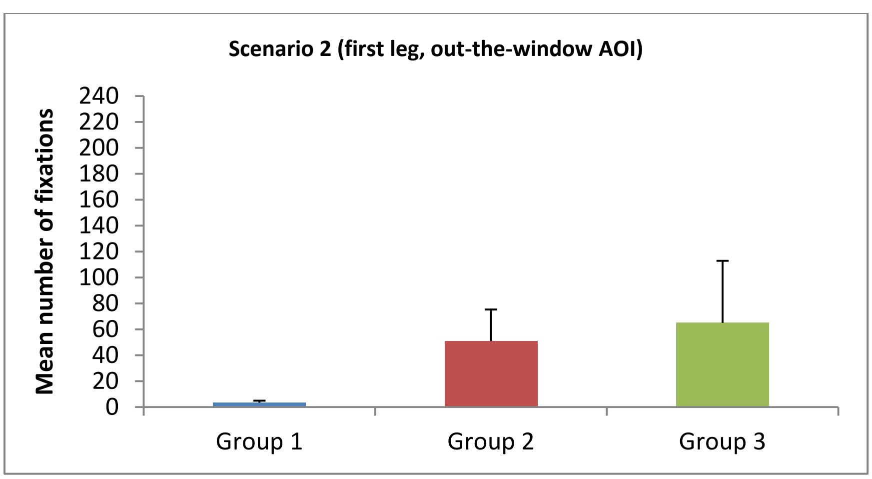


Figure 24. Mean number of fixations on the cockpit out-the-window AOI during Scenario 2.

Figure 25 shows the mean number of fixations of the cockpit out-the-window AOI for Scenario 3. All group comparisons provided anecdotal evidence for the null: Group 1 and Group 2, $t = 1.16$, $BF = 1.39$; Group 1 and Group 3, $t = .76$, $BF = 1.74$; and Group 2 and Group 3, $t = .26$, $BF = 2.45$.

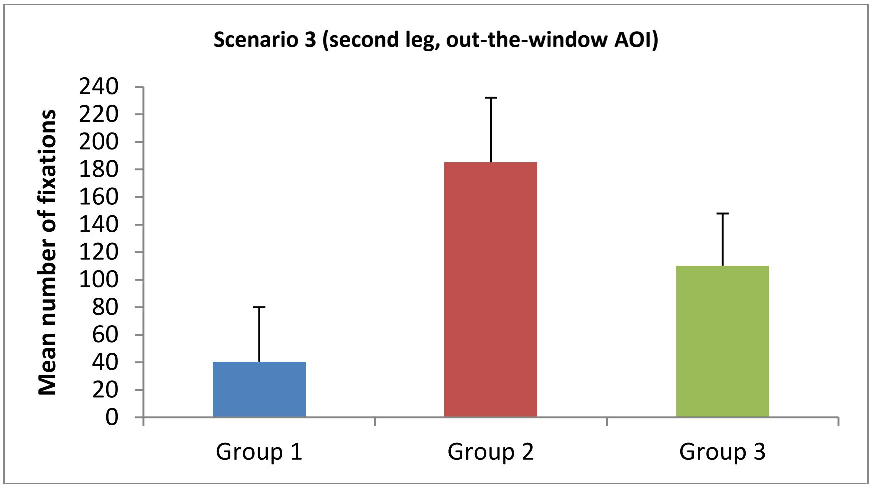


Figure 25. Mean number of fixations on the cockpit out-the-window AOI during Scenario 3.

Figure 26 shows, the mean fixation durations of the cockpit out-the-window AOI for Scenario 2. All comparisons showed anecdotal evidence for the null: Group 1 and Group 2 ($n = 5$), $t = .19$, $BF = 2.08$; Group 1 and Group 3, $t = .77$, $BF = 1.73$; and Group 2 and Group 3, $t = .84$, $BF = 1.94$.

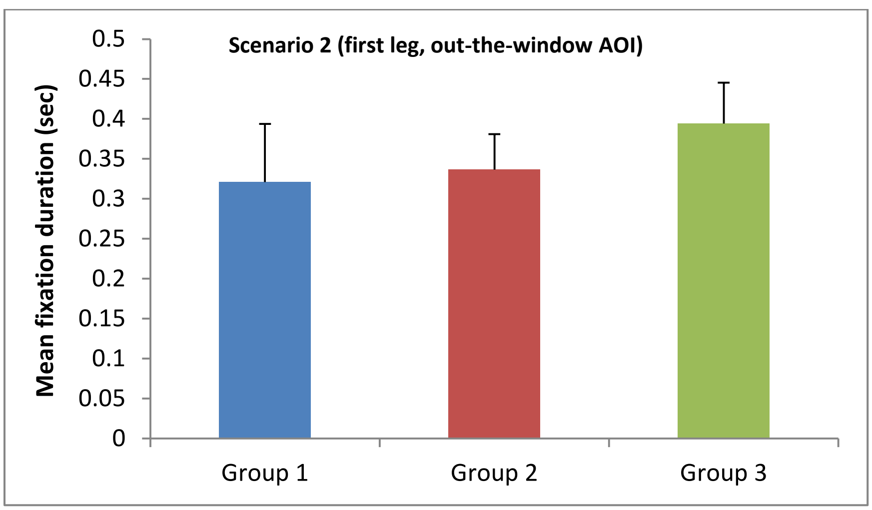


Figure 26. Mean fixation durations on the cockpit out-the-window AOI during Scenario 2.

Figure 27 shows the mean fixation durations on the cockpit out-the-window AOI for Scenario 3. Comparisons between Group 1 and Group 2 (n = 4), t = 1.83, BF = .88, and Group 1 and Group 3, t = 2.12, BF = .7, showed anecdotal evidence for the alternative. The comparison between Group 2 and Group 3 showed anecdotal evidence for the null, t = 1.0, BF = 1.68.

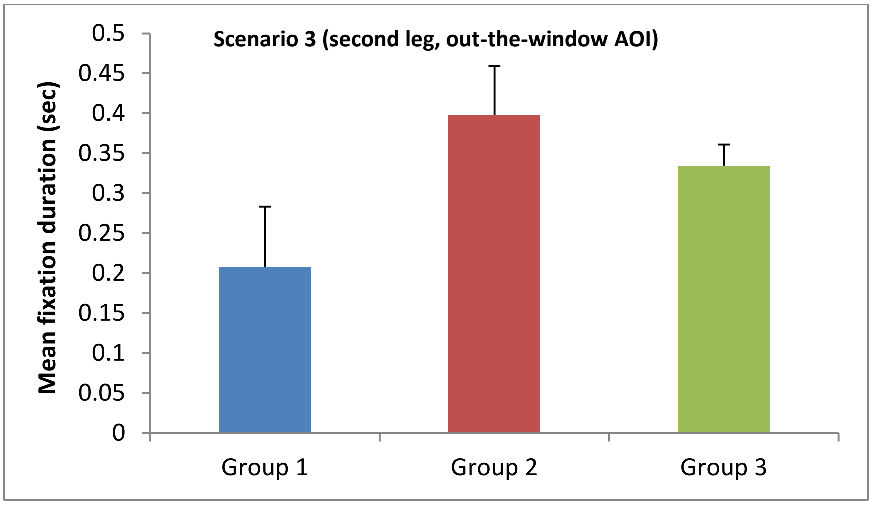


Figure 27. Mean fixation durations on the cockpit out-the-window AOI during Scenario 3.

3.6.2.2. Cockpit glass display AOI

Figure 28 shows the mean number of fixations on the cockpit glass display AOI for Scenario 2. Group 3 (n = 5) had the largest number of fixations followed by Group 1 (n = 3) and Group 2 (n = 5). Comparisons between Group 1 and Group 2 provided anecdotal evidence for the null, t = 1.2, BF = 1.43, as did the comparison between Group 1 and Group 3, t = 1.4, BF = 1.23. The comparison between Group 2 and Group 3, however, provided strong evidence for the alternative, t = 4.35, BF = .06.

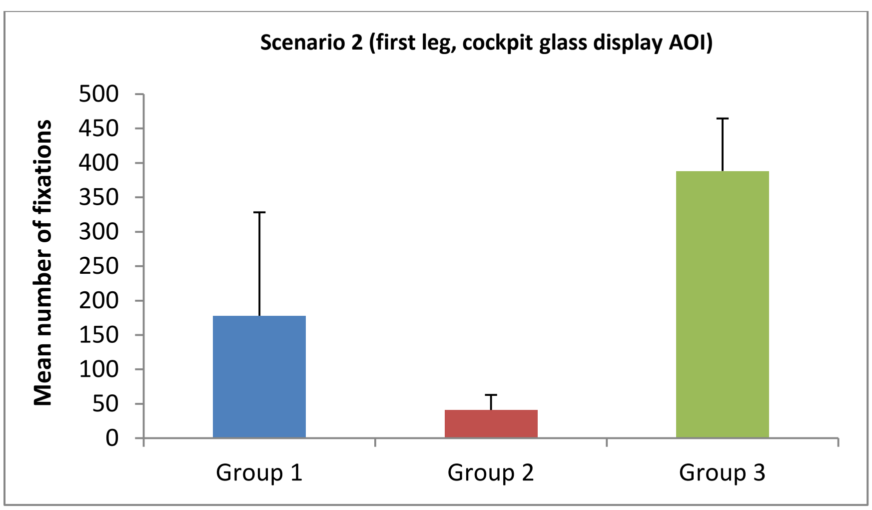


Figure 28. Mean number of fixations on the cockpit glass display AOI during Scenario 2.

Figure 29 shows the mean number of fixations on the cockpit glass display AOI for Scenario 3. Although Group 3 had a higher number of fixations (due to intragroup variability) compared to Group 1 and Group 2, all comparisons provided anecdotal evidence for the null: Group 1 (n = 2) and Group 2 (n = 4), t = .07, BF = 2.06; Group 1 and Group 3 (n = 6), t = .64, BF = 1.86; and Group 2 and Group 3, t = .51, BF = 2.25.

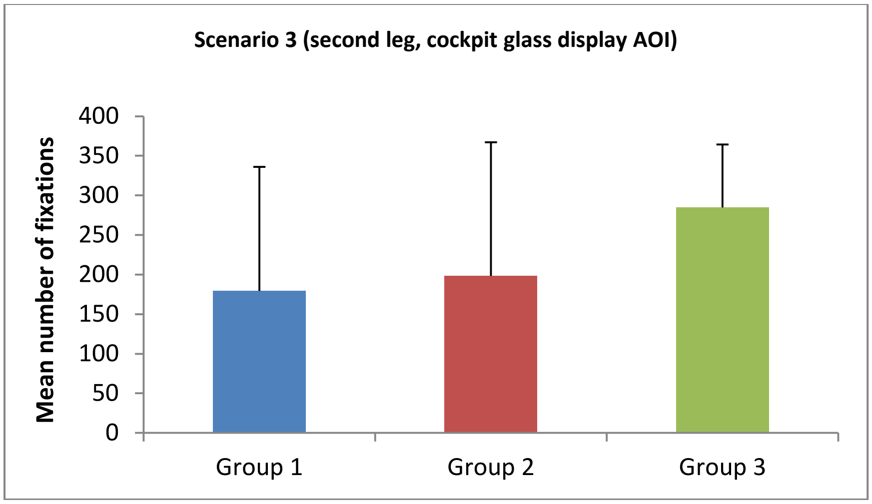


Figure 29. Mean number of fixations on the cockpit glass display AOI during Scenario 3.

Figure 30 shows mean fixation durations for the cockpit glass display AOI for Scenario 2. Again, Group 3 (n = 5) had longer fixation durations compared to Group 1 (n = 3) and Group 2 (n = 5). A comparison between Group 1 and Group 2 revealed anecdotal evidence for the null, t = .66, BF = 1.97, as did the comparison between Group 2 and Group 3, t = .47, BF = 2.31. The comparison between Group 1 and Group 3 yield anecdotal evidence for the alternative, t = 2.4, BF = .5.

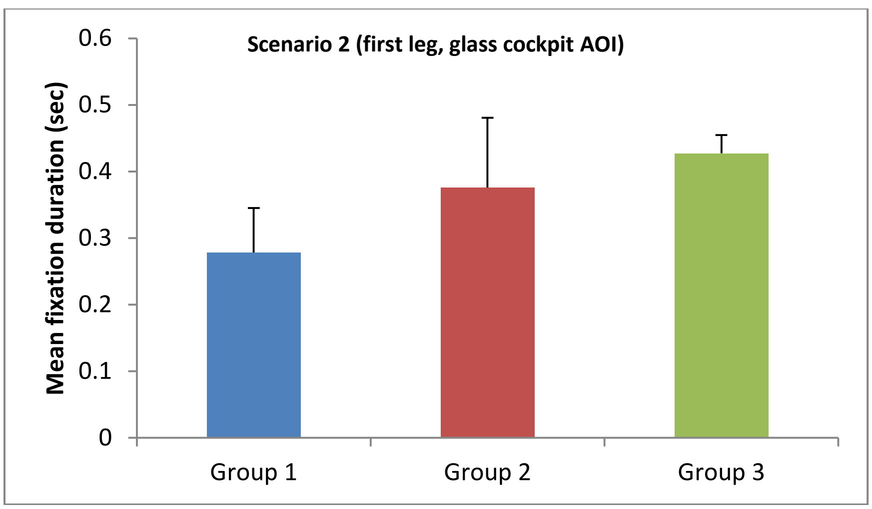


Figure 30. Mean fixation duration for the cockpit glass display AOI during Scenario 2.

Figure 31 shows mean fixation durations for the cockpit glass display AOI for Scenario 3. A comparison between Group 1 (n = 2) and Group 2 (n = 4) revealed anecdotal evidence for the null, t = 1.36, BF = 1.34, as did the comparison between Group 2 and Group 3 (n = 6), t = 1.31, BF = 1.4. The comparison between Group 1 and Group 3 yielded anecdotal evidence for the alternative, t = 2.15, BF = .65.

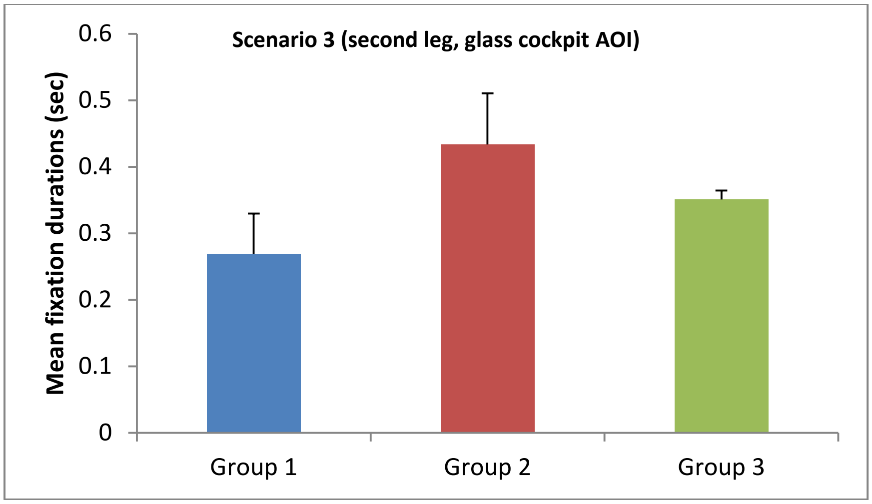


Figure 31. Mean fixation duration for the cockpit glass display AOI during Scenario 3.

3.6.2.3. Weather presentation AOI

Because of the importance of the weather presentation AOI, we analyzed the number and durations for fixations and the number and distances for saccades. Figure 32 shows the mean number of fixations on the weather presentation AOI for Scenario 2. Group comparisons between Group 1 (n = 2) and Group 2 (n = 5) provided substantial evidence for the alternative, t = 4.6, BF = .12. The comparison Group 1 and Group 3 (n = 5) provided anecdotal evidence for the null, t = .71, BF = 1.78. In addition, finally, the comparison between Group 2 and Group 3 provided substantial evidence for the alternative, t = 3.2, BF = .20.

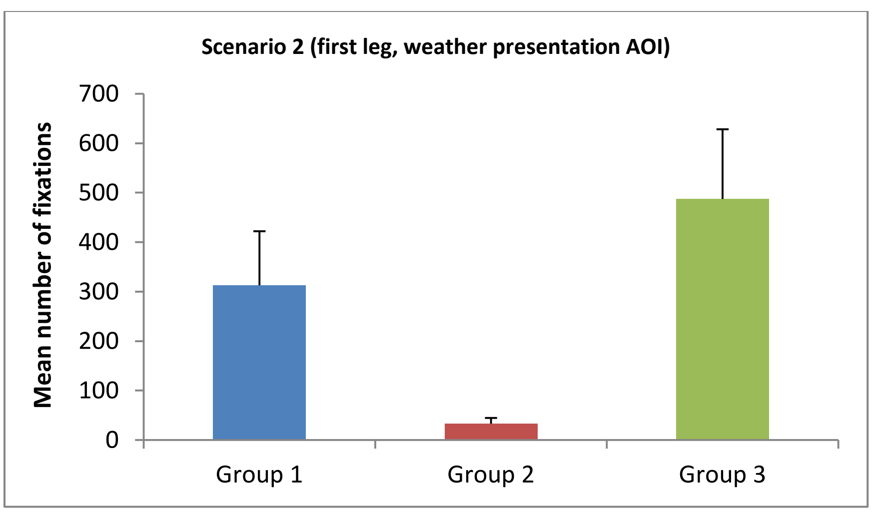


Figure 32. Mean number of fixations on the weather presentation AOI during Scenario 2.

Figure 33 shows the mean number of fixations on the weather presentation AOI for Scenario 3. Due to intragroup variability, the estimates of mean differences were uncertain, and all group comparisons provided anecdotal evidence for the null: Group 1 and Group 2, $t = .70$, $BF = 1.75$; Group1 and Group 3, $t = .02$, $BF = 2.14$; and Group 2 and Group 3, $t = .88$, $BF = 1.88$.

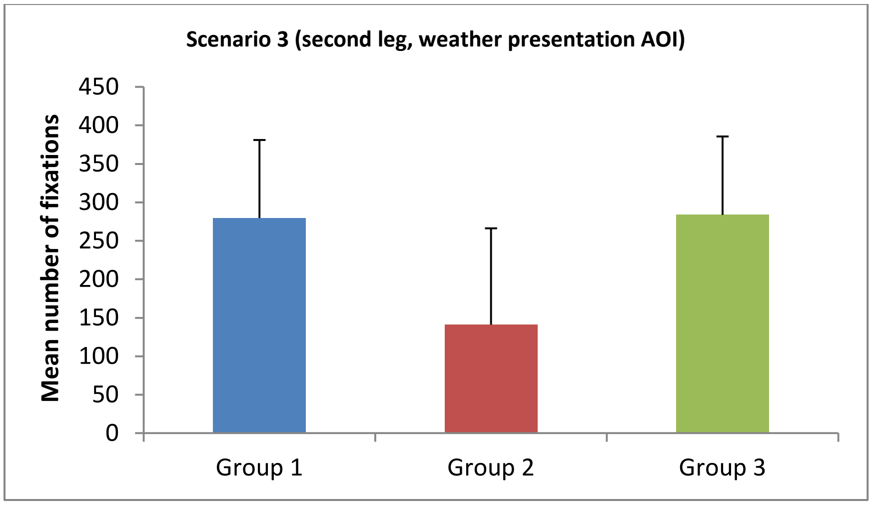


Figure 33. Mean number of fixations on the weather presentation AOI during Scenario 3.

Figure 34 shows the fixation durations for the weather presentation AOI during Scenario 2. The mean fixation durations were similar across groups and all comparisons provided evidence for the null: Group 1 ($n = 2$) and Group 2 ($n = 5$), $t = .41$, $BF = 1.99$; Group 1 and Group 3 ($n = 5$), $t = .17$, $BF = 2.09$; and Group 2 and Group 3, $t = .91$, $BF = 1.88$.

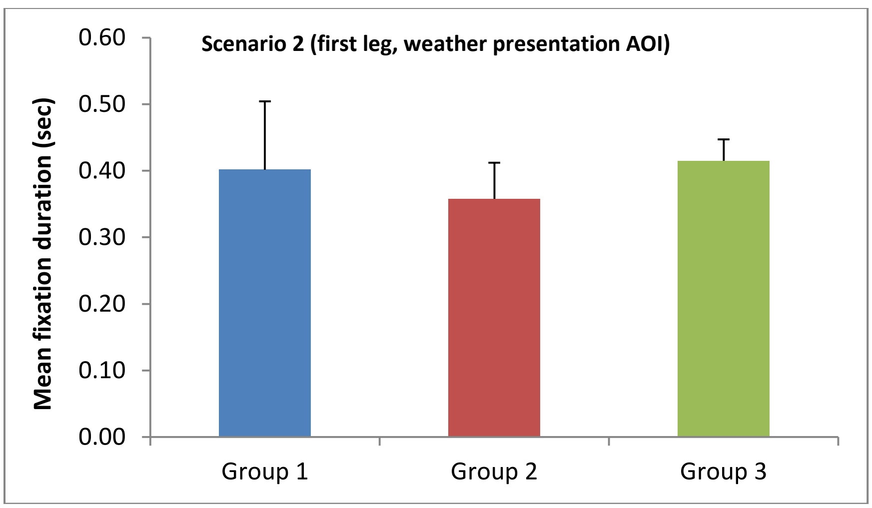


Figure 34. Mean fixation durations for the weather presentation AOI during Scenario 2.

Figure 35 shows the mean fixation durations for the weather presentation AOI during Scenario 3. Group 3 had longer fixation durations compared to Group 1 and Group 2, but all comparisons provided evidence for the null: Group 1 (n = 2) and Group 2 (n = 4), t = .05, BF = 2.07; Group 1 and Group 3 (n = 6), t = 1.39, BF = 1.2; and Group 2 and Group 3, t = 1.43, BF = 1.25.

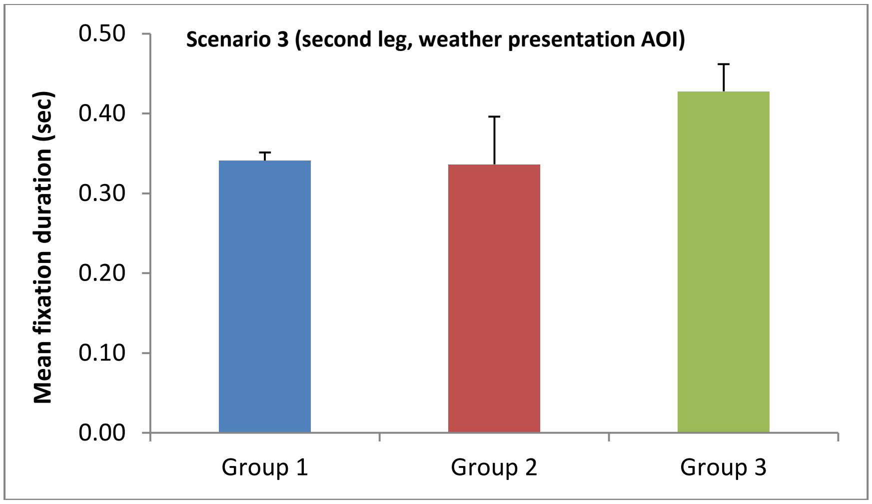


Figure 35. Mean fixation durations for the weather presentation AOI during Scenario 3.

Figure 36 shows the mean number of saccades for the weather presentation AOI during Scenario 2. Group 3 (n = 5) had a larger number of saccades on the weather AOI compared to Group 1 (n = 3) and Group 2 (n = 6). A comparison between Group 1 and Group 2 revealed anecdotal evidence for the alternative, t = 2.04, BF = .70. The comparison between Group 1 and Group 3 showed anecdotal evidence for the null, t = 1.32, BF = 1.32. Finally, the comparison between Group 2 and Group 3 revealed substantial evidence for the alternative, t = 3.53, BF = .12.

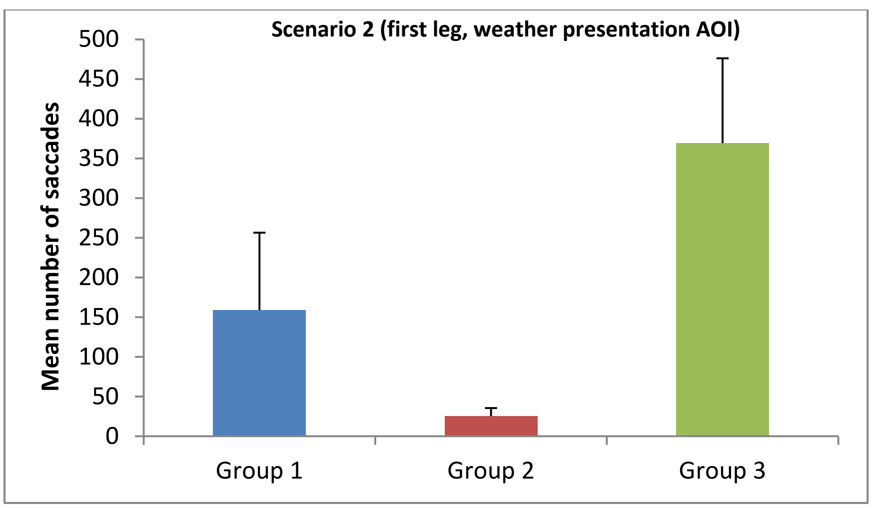


Figure 36. Mean number of saccades for the weather presentation AOI during Scenario 2.

Figure 37 shows the mean number of saccades for the weather presentation AOI during Scenario 3. Due to a large intragroup variability, all group comparisons revealed anecdotal evidence for the null: Group 1 (n = 3) and Group 2 (n = 6), t = .05, BF = 2.14; Group 1 and Group 3 (n = 6), t = .50, BF = 2.14; and Group 2 and Group 3, t = .36, BF = 2.23.

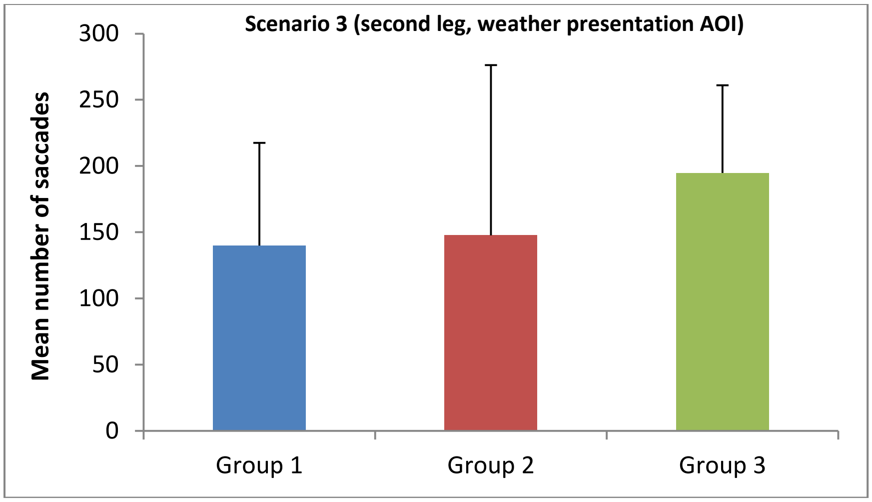


Figure 37. Mean number of saccades for the weather presentation AOI during Scenario 3.

Figure 38 shows the mean saccade distances for the weather presentation AOI during Scenario 2. All group comparisons revealed anecdotal evidence for the null: Group 1 (n = 3) and Group 2 (n = 6), t = .80, BF = 1.87; Group 1 and Group 3 (n = 5), t = 1.21, BF = 1.42; and Group 2 and Group 3, t = .61, BF = 2.25.

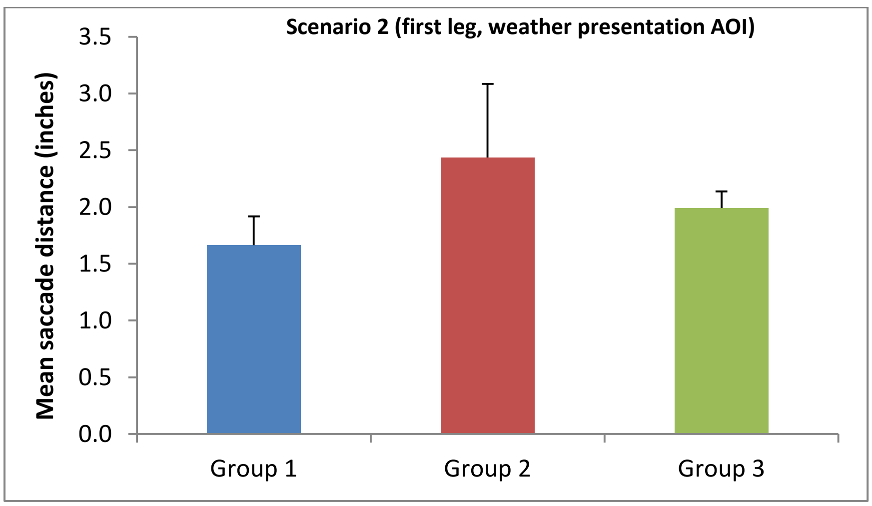


Figure 38. Mean saccade distance for the weather presentation AOI during Scenario 2.

Figure 39 shows the mean saccade distances for the weather presentation AOI during Scenario 3. Group comparisons between Group 1 (n = 3) and Group 2 (n = 3), t = 1.58, BF = 1.06; and between Group 2 and Group 3 (n = 6), t = .59, BF = 2.06, revealed anecdotal evidence for the null. The comparison between Group 1 and Group 3, however, revealed anecdotal evidence for the alternative, t = 1.96, BF = .76.

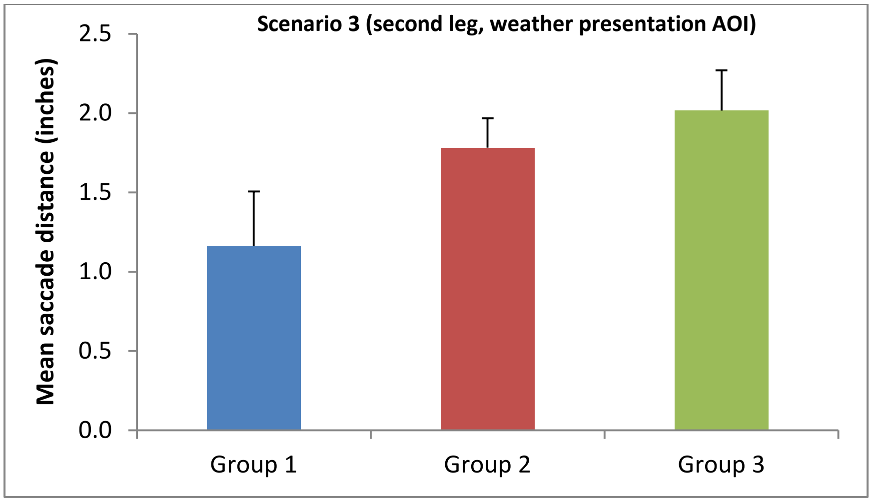


Figure 39. Mean saccade distance for the weather presentation AOI during Scenario 3.

3.6.2.4. Cockpit console AOI

Figure 40 shows the mean number of fixations on the cockpit console AOI during Scenario 2. For the cockpit console AOI, Group 3 has the largest number of fixations. Comparisons between Group 1 (n = 3) and 2 (n = 6) showed anecdotal evidence for the null, t = .92, BF = 1.72. When comparing Group 1 and Group 3 (n = 5), we found anecdotal evidence for the alternative, t = 1.87, BF = .84. The comparison between Group 2 and Group 3 yielded substantial evidence for the alternative, t = 3.28, BF = .16.

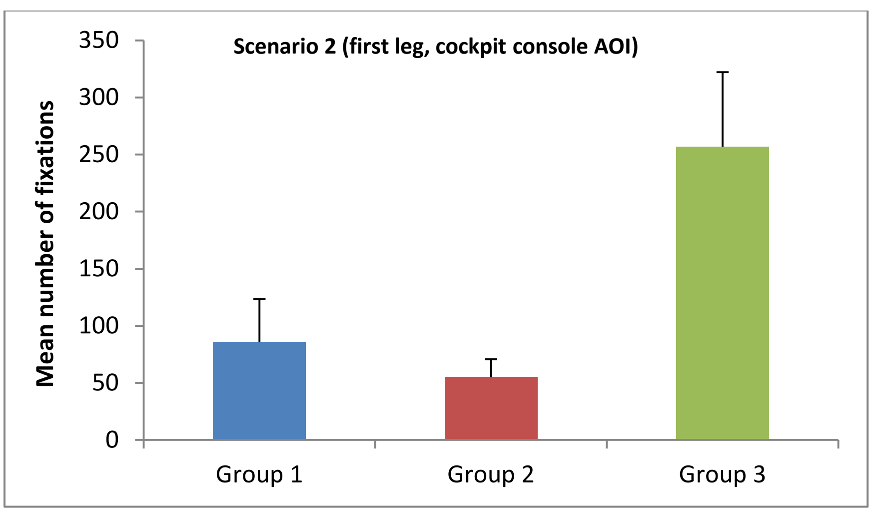


Figure 40. Mean number of fixations for the cockpit console AOI during Scenario 2.

Figure 41 shows the mean number of fixations on the cockpit console AOI during Scenario 3. Similar to Scenario 2, Group 3 had the largest number of fixations. Comparisons between Group 1 (n = 2) and Group 2 (n = 4) showed anecdotal evidence for the null, t = .54, BF = 1.87. When comparing Group 1 and Group 3 (n = 5), we found anecdotal evidence for the alternative, t = 2.44, BF = .54. In addition, finally, the comparison between Group 2 and Group 3 yielded anecdotal evidence for the null, t = 1.15, BF = 1.54.

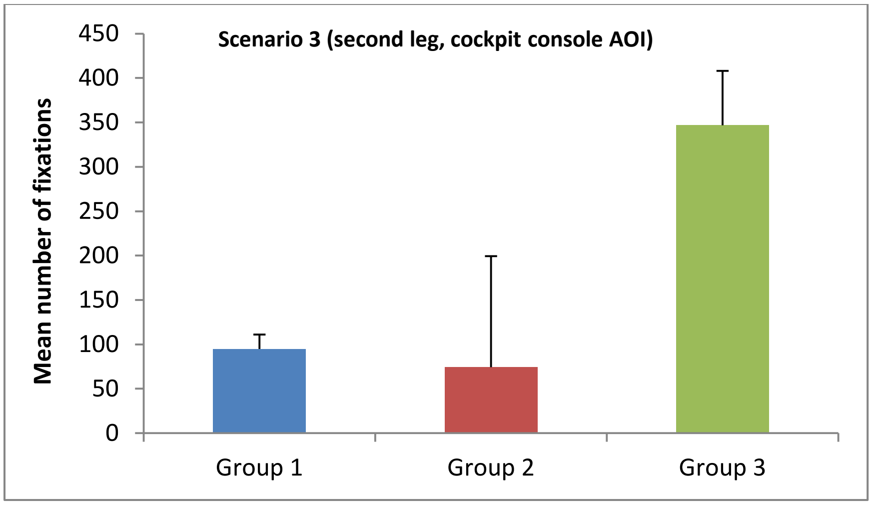


Figure 41. Mean number of fixations for the cockpit console AOI during Scenario 3.

Figure 42 shows the mean fixation durations for the cockpit console AOI during Scenario 2. A comparison between Group 1 (n = 3) and Group 2 (n = 6) revealed anecdotal evidence for the null, t = .86, BF = 1.8, as did the comparison between Group 2 and Group 3 (n = 5), t = .65, BF = 2.21. The comparison between Group 1 and Group 3 showed anecdotal evidence for the alternative, t = 1.73, BF = .94.

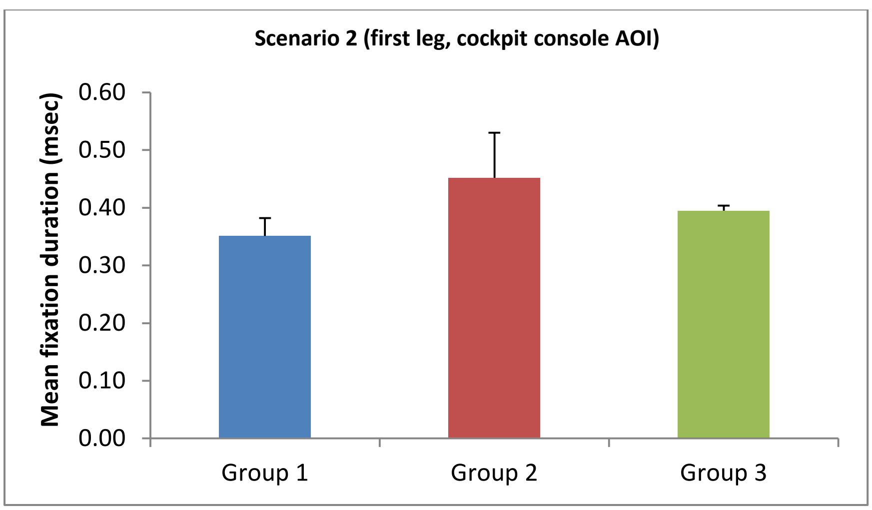


Figure 42. Mean fixation durations for the cockpit console AOI during Scenario 2.

Figure 43 shows the mean fixation durations for the cockpit console AOI during Scenario 3. The result was similar to the outcome for Scenario 2. A comparison between Group 1 (n = 2) and Group 2 (n = 4) revealed anecdotal evidence for the null, t = .77, BF = 1.7, as did the comparison between Group 2 and Group 3 (n = 5), t = .48, BF = 2.23. The comparison between Group 1 and Group 3 showed anecdotal evidence for the alternative, t = 2.36, BF = .58.

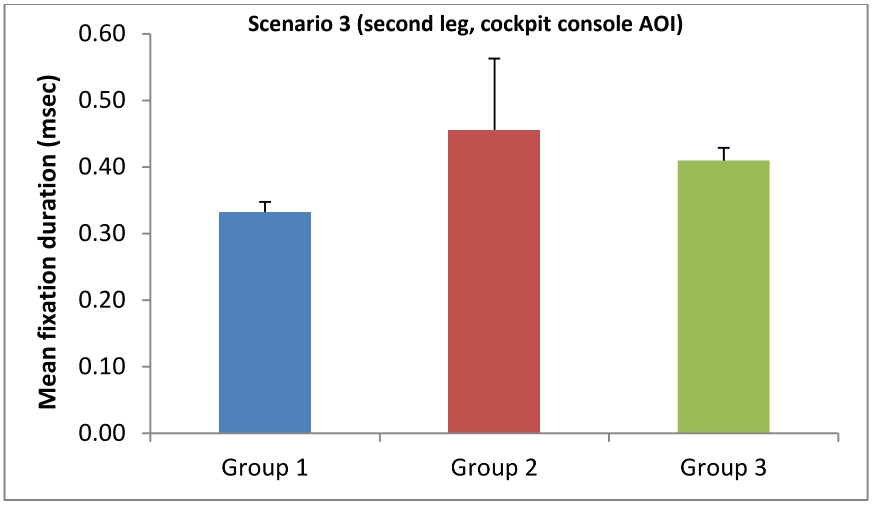


Figure 43. Mean fixation durations for the cockpit console AOI during Scenario 3.

3.6.3. Pupil Diameter

Figure 44 shows the mean pupil diameter for Groups 1-3 during Scenario 2. All group means are very similar and all comparisons provided anecdotal evidence for the null: Group 1 (n = 3) and Group 2 (n = 6), t = .43, BF = 2.18; Group 1 and Group 3 (n = 6), t = 1.17, BF = 1.48; and Group 2 and Group 3, t = .31, BF = 2.58.

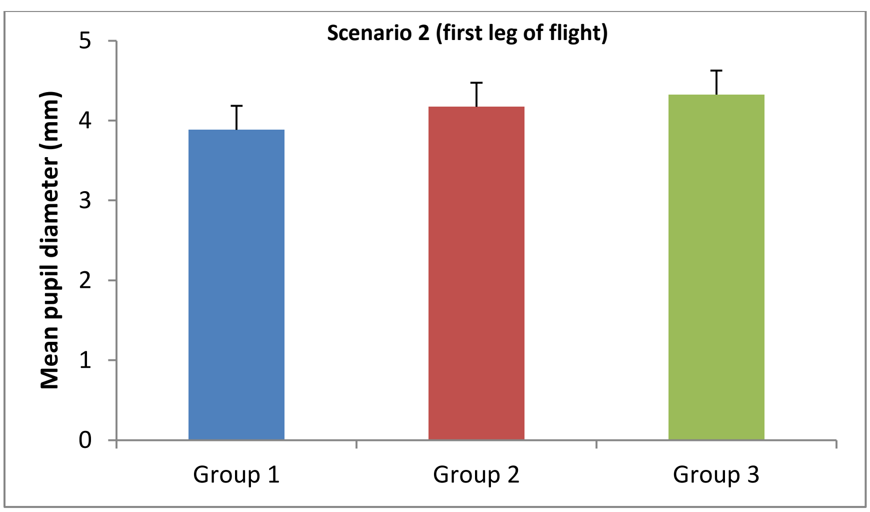


Figure 44. Mean pupil diameter for the first 5 minutes of Scenario 2.

Figure 45 shows the mean pupil diameter for Groups 1-3 during Scenario 3. Again, all means were very similar and all comparisons provided anecdotal evidence for the null: Group 1 and Group 2, $t = .48$, $BF = 2.15$; Group 1 and Group 3, $t = 1.41$, $BF = 1.23$; and Group 2 and Group 3, $t = 1.28$, $BF = 1.48$.

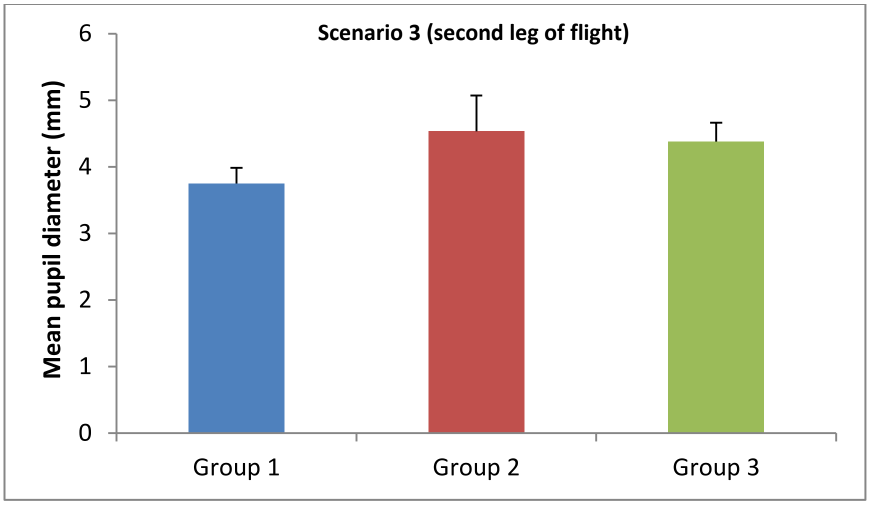


Figure 45. Mean pupil diameter for the last 20 minutes of Scenario 3.

3.6.4. Eye-Motion Workload

EMW is the average degrees per second that the eyes move around during the course of a scenario. Figure 46 shows the mean EMW for Groups 1-3 during Scenario 2. A group comparison between Group 1 ($n = 3$) and Group 2 ($n = 6$) provided no evidence for the null nor the alternative, $t = 1.63$, $BF = 1.0$. A comparison between Group 1 and Group 3 ($n = 6$) provided anecdotal evidence for the null, $t = 1.27$, $BF = 1.37$. A comparison between Group 2 and Group 3, finally, provided anecdotal evidence for the alternative, $t = 2.44$, $BF = .43$.

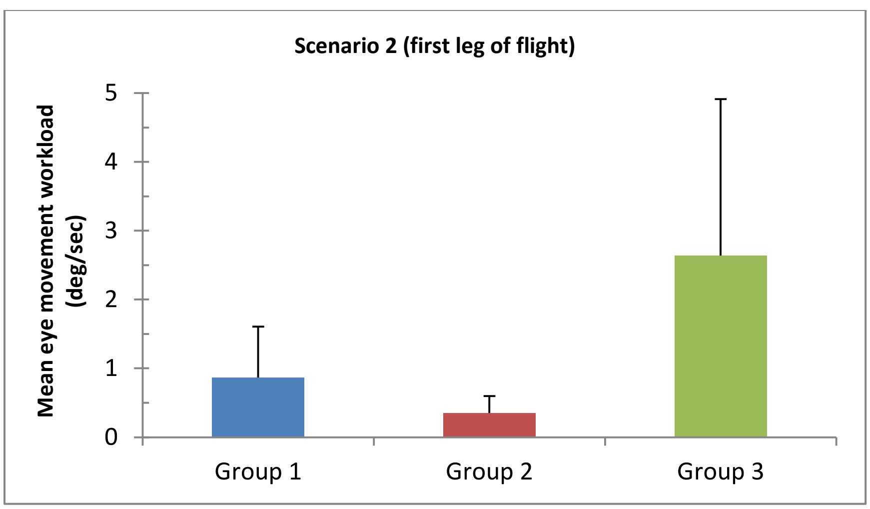


Figure 46. Mean eye-movement workload for Groups 1-3 during Scenario 2.

Figure 47 shows the mean EMW for Groups 1-3 during Scenario 3. Similar to the result for Scenario 2, Group 3 had a higher mean EMW workload. However, due to a larger intragroup variability, all comparisons yielded anecdotal evidence for the null: Group 1 and Group 2, $t = .81$, $BF = 1.77$; Group 1 and Group 3, $t = 1.47$, $BF = 1.17$; and Group 2 and Group 3, $t = .63$, BF = 2.14.

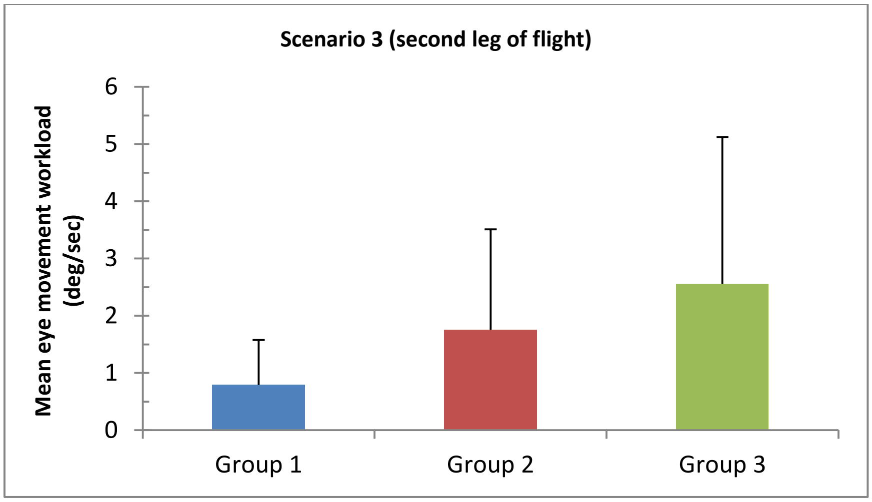


Figure 47. Mean eye-movement workload for Groups 1-3 during Scenario 3.

3.7. Functional Near-Infrared Analysis

In this section, we present results from the fNIR recordings. During Scenario 2 and Scenario 3, pilots inadvertently entered IMC 5 minutes into the scenario. We explored whether there were any differences in pilot oxygenation between the first VFR portion (0-1 min) and the initial IFR portion (5-6 min) of the flight. Figure 48 shows the mean oxygenation for Groups 1-3 during Scenario 2.

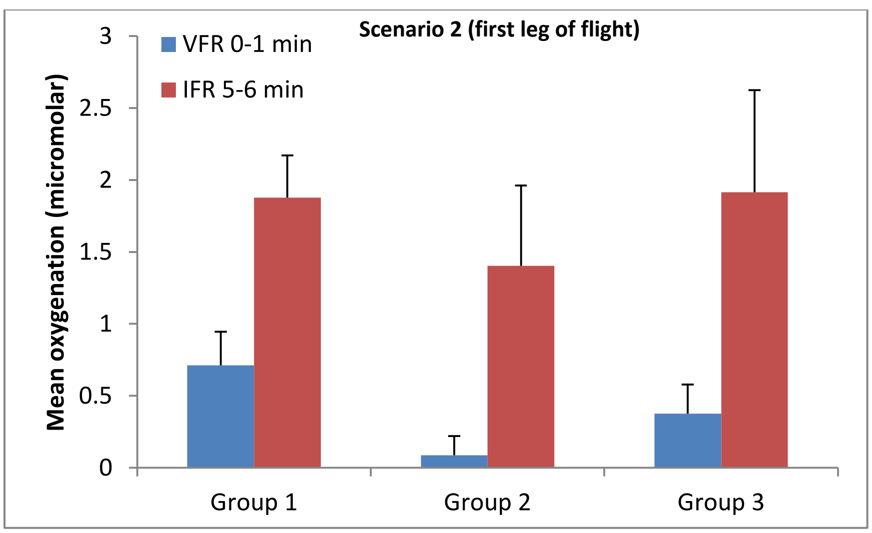


Figure 48. Number oxygenation for the initial VFR (0-1 min) and IFR (5-6 min) segments of Scenario 2.

All groups showed a larger mean oxygenation during the IFR portion of the flight compared to the VFR portion of the flight. For Group 1 (n = 8), the VFR vs. IFR comparison provided strong evidence for the alternative, t = 4.65, BF = .04. For Group 2 (n = 7), the VFR vs. IFR comparison provided anecdotal evidence for the alternative, t = 2.71, BF = .39. Finally, for Group 3 (n = 7), the comparison also provided anecdotal evidence for the alternative, t = 2.77, BF = .37.

A comparison of the oxygenation between Group 1 and Group 2 for the VFR portion (0-1 min) of the flight provided anecdotal evidence for the alternative, t = 2.22, BF = .54. The comparison between Group 1 and Group 3 provided anecdotal evidence for the null, t = 1.06, BF = 1.88, as did the comparison between Group 2 and Group 3, t = 1.2, BF = 1.65.

Comparing the oxygenation between groups for the initial IFR portion (5-6 min) of the flight, we found that all comparisons provided anecdotal evidence for the null: Group 1 and Group 2, t = .77, BF = 2.3; Group 1 and Group 3, t = .05, BF = 2.89; and Group 2 and Group 3, t = .56, BF = 2.5.

Figure 49 shows the mean oxygenation for Groups 1-3 during Scenario 3. As we can see by the SE bars, the Scenario 3 oxygenation showed much more intragroup variability compared to the Scenario 2 result. For Group 1, the VFR vs. IFR comparison provided substantial evidence for the null, t = .31, BF = 3.54. For Group 2, the VFR vs. IFR comparison provided anecdotal evidence for the null, t = 1.16, BF = 1.97. Finally, for Group 3, the comparison also provided anecdotal evidence for the null, t = 1.52, BF = 1.47.

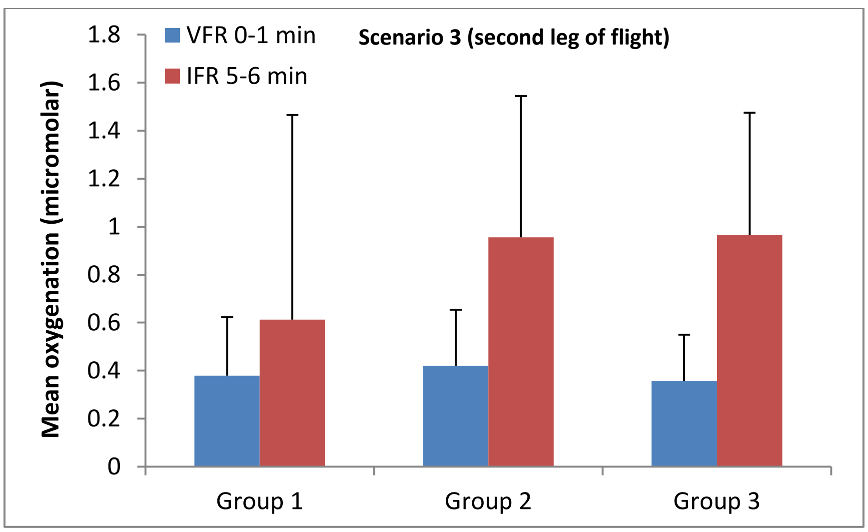


Figure 49. Mean oxygenation for the initial VFR (0-1 min) and IFR (5-6 min) segments of Scenario 3.

All VFR group comparisons provided anecdotal evidence for the null: Group 1 and Group 2, t = .12, BF = 2.73; Group 1 and Group 3, t = .06, BF = 2.82; and Group 2 and Group 3, t = .22, BF = 2.7. Similarly, the IFR group comparisons also provide anecdotal evidence for the null: Group 1 and Group 2, t = .31, BF = 2.65; Group 1 and Group 3, t = .36, BF = 2.69; and Group 2 and Group 3, t = .01, BF = 2.75.

In addition, we explored whether there were any differences in pilot oxygenation between the initial VFR (0-5 min) portion and the initial IFR (5-10 min) portion of the flight. Figure 50 shows the mean oxygenation for Groups 1-3 during Scenario 2. All groups displayed a larger mean oxygenation during the IFR portion of the flight compared to the VFR portion of the flight. For Group 1 (n = 8), the VFR vs. IFR comparison provided anecdotal evidence for the alternative, t =

2.46, *BF* = .51. For Group 2 (*n* = 7), the VFR vs. IFR comparison provided substantial evidence for the alternative, *t* = 3.48, *BF* = .18. Finally, for Group 3 (*n* = 7), the comparison provided anecdotal evidence for the alternative, *t* = 2.16, *BF* = .73.

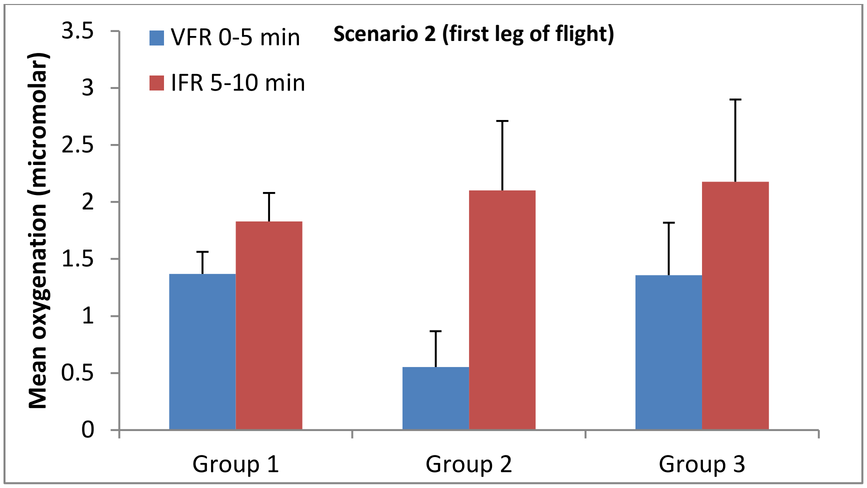


Figure 50. Mean oxygenation for VFR (0-5 min) and IFR (5-10 min) segments of Scenario 2.

A comparison of the oxygenation between Group 1 and Group 2 for the VFR portion (0-5 min) of the flight provided anecdotal evidence for the alternative, *t* = 2.26, *BF* = .51. The comparison between Group 1 and Group 3 provided anecdotal evidence for the null, *t* = .02, *BF* = 2.89, as did the comparison between Group 2 and Group 3, *t* = 1.43, *BF* = 1.34.

Comparing the oxygenation between groups for the IFR portion (5-10 min) of the flight, we found that all comparisons provided anecdotal evidence for the null: Group 1 and Group 2, *t* = .43, *BF* = 2.69; Group 1 and Group 3, *t* = .47, *BF* = 2.65; and Group 2 and Group 3, *t* = .07, *BF* = 2.82.

Figure 51 shows the mean oxygenation for Scenario 3. The results for Scenario 3 are similar to the results for Scenario 2, concerning pilot oxygenation between the initial VFR (0-5 min) portion and the initial IFR (5-10 min) portion of the flight. However, only Group 1 and Group 3 displayed a larger mean oxygenation during the IFR portion of the flight compared to the VFR portion of the flight. For Group 1, the VFR vs. IFR comparison provided anecdotal evidence for the alternative, *t* = 1.99, *BF* = .88. For Group 2, the VFR vs. IFR comparison provided anecdotal evidence for the null, *t* = 1.21, *BF* = 1.89. Finally, for Group 3, the comparison provided anecdotal evidence for the alternative, *t* = 2.72, *BF* = .39.

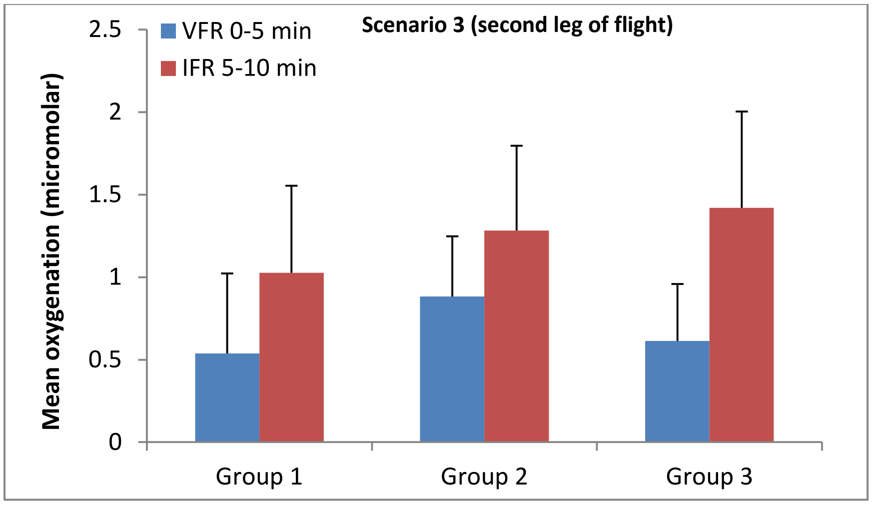


Figure 51. Mean oxygenation for VFR (0-5 min) and IFR (5-10 min) segments of Scenario 3.

A comparison of the oxygenation between Group 1 and Group 2 for the VFR portion (0-5 min) of Scenario 3 provided anecdotal evidence for the null, $t = .55$, $BF = 2.44$. The comparison between Group 1 and Group 3 also provided anecdotal evidence for the null, $t = .13$, $BF = 2.81$, as did the comparison between Group 2 and Group 3, $t = .55$, $BF = 2.44$.

Comparing the oxygenation between groups for the IFR portion (5-10 min) for Scenario 3, we find that all comparisons provided anecdotal evidence for the null: Group 1 and Group 2, $t = .34$, $BF = 2.63$; Group 1 and Group 3, $t = .52$, $BF = 2.54$; and Group 2 and Group 3, $t = .18$, $BF = 2.71$.

As part of the fNIR analysis, we assessed the mean oxygenation for the entire VFR and IFR portion of Scenario 2 and Scenario 3. The VFR portion consisted of the first 5 minutes and the last 5 minutes of the scenario. The IFR portion was the 15 minutes following pilots' inadvertent IMC encounter at 5 minutes into the scenario.

Figure 52 shows the mean oxygenation for the VFR and IFR portions of Scenario 2. Two of the groups displayed a higher mean oxygenation during the IFR portion of the flight compared to the VFR portion of the flight. For Group 1, the VFR vs. IFR comparison provided anecdotal evidence for the null, $t = .39$, $BF = 2.77$. For Group 2, the VFR vs. IFR comparison provided anecdotal evidence for the alternative, $t = 2.38$, $BF = .40$. Finally, for Group 3, the comparison provided substantial evidence for the alternative, $t = 3.0$, $BF = .29$.

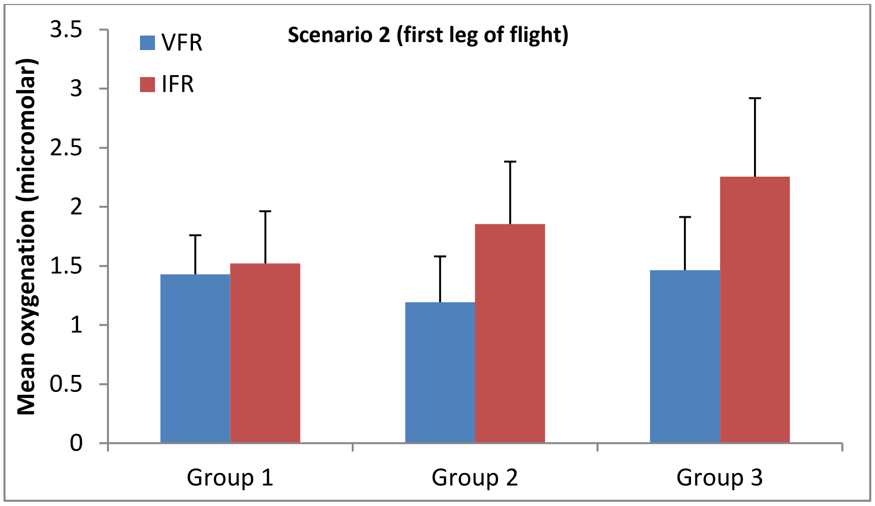


Figure 52. Mean oxygenation for the entire VFR (0-5 min + 20-25 min) and IFR (5-20 min) segments of Scenario 2.

All group comparisons for the VFR portion of the Scenario 2 flights provided anecdotal evidence for the null: Group 1 and Group 2, t = .46, BF = 2.66; Group 1 and Group 3, t = .06, BF = 2.89; and Group 2 and Group 3, t = .45, BF = 2.61. Similarly, the group comparisons for the IFR portion of the flight also provided anecdotal evidence for the null: Group 1 and Group 2, t = .48, BF = 2.64; Group 1 and Group 3, t = .94, BF = 2.05; and Group 2 and Group 3, t = .47, BF = 2.59.

Figure 53 shows the mean oxygenation for the VFR and IFR portions of Scenario 3. All groups displayed a larger mean oxygenation during the IFR portion of the flight compared to the VFR portion of the flight. For Group 1, the VFR vs. IFR comparison provided substantial evidence for the alternative, t = 3.47, BF = .17. For Group 2, the VFR vs. IFR comparison provided anecdotal evidence for the alternative, t = 2.48, BF = .51. Finally, for Group 3, the comparison provided substantial evidence for the alternative, t = 3.86, BF = .11.

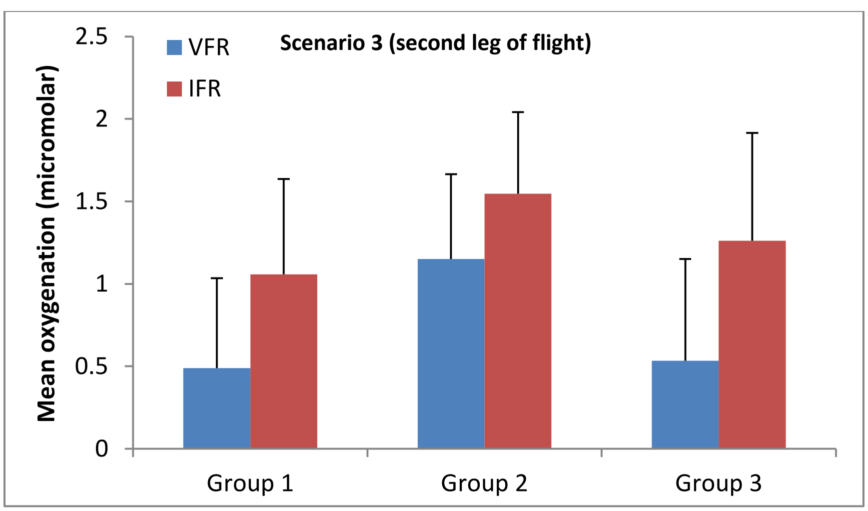


Figure 53. Mean oxygenation for the entire VFR (0-5 min + 20-25 min) and IFR (5-20 min) segments of Scenario 3.

All group comparisons for the VFR portion of the Scenario 3 flights provided anecdotal evidence for the null: Group 1 and Group 2, $t = .87$, $BF = 2.06$; Group 1 and Group 3, $t = .05$, $BF = 2.82$; and Group 2 and Group 3, $t = .79$, $BF = 2.17$. Similarly, the group comparisons for the IFR portion of the Scenario 3 flights also provided anecdotal evidence for the null: Group 1 and Group 2, $t = .63$, $BF = 2.36$; Group 1 and Group 3, $t = .24$, $BF = 2.76$; and Group 2 and Group 3, $t = .35$, $BF = 2.62$.

From the current oxygenation analysis, we have evidence that pilot oxygenation is higher during IFR than during VFR portions of flight. As pilots transitioned from VFR to IFR, the oxygenation levels increased. In addition, we explored whether oxygenation levels decreased during the transition from IFR to VFR. At 20 minutes into the scenario, the IFR portion of the flight transitioned into VFR. Figure 54 shows the mean oxygenation for the last minute of the IFR portion (min 19-20) and the first minute of the VFR portion (min 20-21) for Scenario 2. For Group 1, the oxygenation difference between the transition from IFR to VFR provided substantial evidence for the null, $t = .09$, $BF = 3.9$. For Group 2, the transition from IFR to VFR provided anecdotal evidence for the null, $t = 1.49$, $BF = 1.51$. For Group 3, however, the oxygenation difference between the transition from IFR to VFR provided anecdotal evidence for the alternative, $t = 2.68$, $BF = .41$.

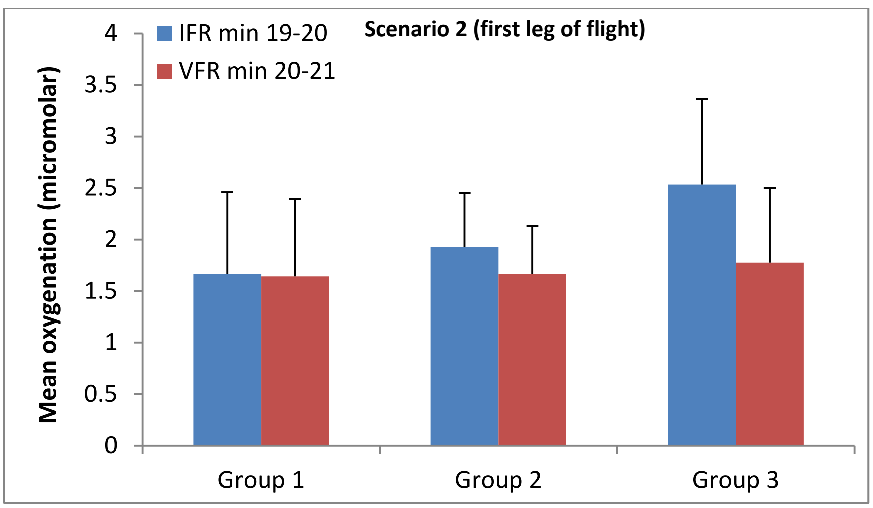


Figure 54. Mean oxygenation for the IFR (19-20 min) to VFR (20-21 min) transition during Scenario 2.

All group comparisons of the IFR portion of flight provided anecdotal evidence for the null: Group 1 and Group 2, $t = .26$, $BF = 2.82$; Group 1 and Group 3, $t = .75$, $BF = 2.32$; and Group 2 and Group 3, $t = .61$, $BF = 2.45$. Similarly, group comparisons of the oxygenation during the VFR portion also provided anecdotal evidence for the null: Group 1 and Group 2, $t = .02$, $BF = 2.89$; Group 1 and Group 3, $t = .12$, $BF = 2.88$; and Group 2 and Group 3, $t = .12$, $BF = 2.81$.

Figure 55 shows the mean oxygenation for the last minute of the IFR portion (min 19-20) and the first minute of the VFR portion (min 20-21) for Scenario 3. Again, as we can see by the SE bars, the Scenario 3 oxygenation showed much more intragroup variability compared to the Scenario 2 result. For all groups, the oxygenation change during the IFR to VFR transition provided anecdotal evidence for the null: Group 1, $t = .87$, $BF = 2.63$; Group 2, $t = 1.67$, $BF = 1.22$; and Group 3, $t = 1.64$, $BF = 1.29$.

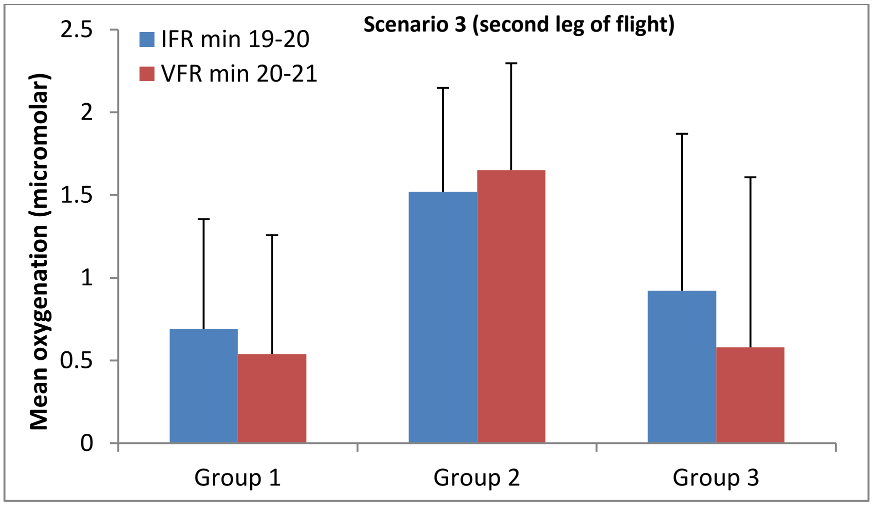


Figure 55. Mean oxygenation for the IFR (19-20 min) to VFR (20-21 min) transition during Scenario 3.

In a similar way, all group comparisons of the oxygenation for both the IFR portion and the VFR portion provided anecdotal evidence for the null. For the IFR portion of flight, Group 1 and Group 2, t = .90, BF = 2.02; Group 1 and Group 3, t = .21, BF = 2.78; and Group 2 and Group 3, t = .53, BF = 2.46. For the VFR portion of flight, Group 1 and Group 2, t = 1.13, BF = 1.71; Group 1 and Group 3, t = .03, BF = 2.83; and Group 2 and Group 3, t = .89, BF = 2.04.

3.8. Weather Questionnaire Results

After completing the flight runs, we asked each pilot to rate his experience with the weather presentation (see Appendix A). We focused our questions on the overall ease of use and if the weather presentation affected pilot decision-making during weather avoidance. In this section, we report descriptive statistics to determine whether important differences exist in perceived usability (i.e., ease of information retrieval), workload, and weather situational awareness between the three weather presentations.

3.8.1. Weather Avoidance

Across groups, we found that 18 out of 25 pilots reported that it was easy to distinguish convective SIGMET areas on their presentation. These same 18 pilots believed that their weather presentation was very helpful in determining the locations of severe precipitation areas and that they used it very frequently throughout their flights. Out of the remaining seven pilots, three pilots in Group 1 and one pilot each in Group 2 and Group 3 stated that it was somewhat helpful. Only one pilot, in Group 1, stated that the presentation was not helpful. That same pilot also commented that the Group 1 presentation was not helpful in determining severe precipitation locations. All other pilots reported that their precipitation classification symbols were either helpful or very helpful. In a related issue, we asked pilots how helpful the weather presentation was in determining what areas of weather to avoid. Twenty-three out of 24 of the pilots commented that the information was very helpful for weather avoidance, and the remaining pilot still rated it as helpful. This suggests that, regardless of condition, the presentation gave pilots beneficial information for determining and avoiding unfavorable weather conditions.

3.8.2. Weather Presentation Usage and Trust

Although the pilots were told in their briefing to use the weather presentation to avoid weather, not all pilots used the presentation all the time throughout the experiment. In terms of frequency of use, we found that all pilots did use the weather presentation during their flight. One pilot out of 25 pilots interviewed commented that he used the weather presentation *only sometimes* to navigate through or around adverse weather. This pilot was using the Group 1 presentation. All other pilots reported that they used their presentation either frequently or very frequently.

Situational awareness is paramount to the PDM process, so we examined pilot usability and trust in the information being presented. We asked the pilots to compare the benefits of their respective presentation information against other sources of weather information they have used. Twenty-two of the 25 pilots rated their presentation as very good compared to other sources they have used. All pilots, regardless of group, rated their presentation as either good or very good. None of the pilots found the information unhelpful. Pilots were asked to give a trust rating from 1 to 5 that the information they were being given was correct. Twenty-one pilots gave high trust ratings (4 or 5). The remaining three pilots gave ratings of 3. The pilots trusted the Group 2 presentation the most (8 out of 25), then Group 1 (7 out of 25), and, finally, Group 3 (6 out of 25). In addition to comparing the Groups to other presentations, we asked the pilots how much their respective presentations decreased their workload while flying. On a scale ranging from 1 (*very little*) to 5 (*very much*), 13 pilots rated their presentation as either 4 or 5. The remaining pilots rated their group as either 2 or 3. No pilots gave a rating of 1 to any of the three weather presentations. This suggests that, subjectively, use of a weather presentation decreased workload for the pilots, regardless of the weather presentation group. Ability to determine the position of the aircraft was also rated by the pilots. Twenty-three out of 24 pilots reported that it was very easy to determine the aircraft position. The remaining pilot felt that there was no difference, neither good nor bad.

3.9. Pilot Debriefing Comments

At the conclusion of each participant's second flight scenario, we debriefed the pilot. We asked pilots about the overall functionality and usability of their weather presentation display. The pilots provided criticisms and some suggestions for future simulation improvements. The debriefing session was broken up into questions about the weather presentation of the SIGMET area, terrain map, precipitation display, lightning symbols, METAR symbols, zoom levels, and range rings. We summarize participant responses to items pertaining to flight realism, cockpit simulator function, and weather presentation effectiveness, as these responses can help us improve the simulation conditions for future studies.

3.9.1. SIGMET Symbols

Out of the 25 pilots interviewed, one pilot in Group 1, three pilots in Group 2, and two pilots in Group 3 all commented that the SIGMET was overwhelming and that it obscured information. None of the pilots felt that the SIGMET was unnecessary, but three pilots commented that it was helpful. Pilots remarked that a time stamp as well as cross-hatched lines, instead of solid colors, would be helpful improvements to the design.

3.9.2. Terrain Map

All group weather presentations included a physical terrain map overlay (as illustrated previously in Figure 3). Over half of the pilots interviewed felt that the terrain was a good and helpful map to have under weather presentations. In particular, six pilots felt that it was a great tool for knowing Airports and VORS and for going direct, especially in small planes.

3.9.3. Precipitation Display

Our pilots felt that the precipitation display conveyed useful information. However, many pilots felt that 3 to 5 levels of precipitation was more than adequate and anything more would be excessive. Two pilots thought that the precipitation colors should be "the same as on TV"—meaning the color palette should match the schema of the Storm Center Doppler radar on the news. When asked what color of weather they would normally fly through, only three pilots commented that they would avoid anything above dark green on the precipitation scale, and only one pilot commented that the display was helpful for gauging storm cell growth.

3.9.4. Lightning Symbols

Pilots had very little problem with the lightning display, except for several pilots in the Group 2 condition commented that the lightning was obscured by the magenta SIGMET. The pilots also suggested that a time stamp or some way to know the age of the lightning strike would be helpful.

3.9.5. METAR Symbols

We asked the pilots for their opinion on the use of the METAR data on their situational awareness while flying. We found mixed attitudes on whether METAR information is better given in a graphical or textual form. Even with this uncertainty, the pilots suggested that the inclusion of METAR symbology was helpful in performing deviations and landings, identifying VFR airports, and spotting VORs. They suggested that adding speed vectors to the data for the weather cells and having some sort of indicator for whether the METARs are visibility based, or low-cloud ceiling based, would be helpful to future pilots using the system.

3.9.6. Weather Presentation Zoom and Range Rings

Pilots used a zoom feature while flying their scenarios. The zoom levels given were 5 nmi, 20 nmi, and 50 nmi out. We collected feedback on the use of the zoom feature as well as the use of range rings for determining distances while in flight. All of the pilots interviewed acknowledged that the zoom functionality was very good for helping with situational awareness. The consensus was that the pilots used the smaller levels for detail work and used the larger level for an overview of the scenario: "I used the 20- and 5-mile zoom a lot for distance keeping and then would zoom out to the 50 miles for the big picture." The pilots used these zoom levels in conjunction with built-in range rings. We asked the pilots if this was a useful feature for them. They replied that the rings were very helpful for judging distances and helped in determining storm cell scale as well. However, they suggested that having more range rings, or the ability to customize this scale preference, would be better.

4. DISCUSSION

In this study, we assessed GA pilot behavior during weather avoidance operations. Twenty-five IFR-rated pilots were randomly allocated to one of three simulation groups. The only difference between groups was the symbology of the weather presentation. During two weather scenarios, we measured pilot performance and behavior by dependent measures that included ATC communication, decision-making, weather avoidance, and cognitive workload. On the tacit assumption that matched pilot groups, on average, will perform similarly when confronted with the same operational tasks, we analyzed our simulation outcomes according to two model hypotheses. The first hypothesis, the *null*, states that there is no difference in performance or behavior between groups. The second hypothesis, the *alternative*, states that there are differences in performance or behavior between groups.

The magnitude for most of our simulation effects is anecdotal and, in most cases, the effects favor the null over the alternative. For this simulation, the null is a viable model because some dependent measures are likely unaffected by variations in weather symbology. One example is our distance to weather measure where all pilots essentially performed the same and kept a safe distance from the 30 dBZ intensities, despite variations in precipitation symbology. We find that, on average, pilots stayed 14 miles away from the smaller cells and 40 miles away from the line storm as measured from the aircraft position to the closest area of 30 dBZ intensities. This illustrates desirable decision-making on part of the pilots. It also shows that the weather presentations for Groups 1-3 provided the necessary information to stay safely away from hazardous precipitation. All three weather presentations are seemingly equivalent because they provide information about cell intensity and location and all outcome comparisons support the null. Even for studies that systematically manipulated precipitation symbology and measured pilot distance to weather, the outcomes were similar. For example, a study by Beringer and Ball (2004) measured the closest approach distance to convective cells for GA groups using different NEXRAD resolutions (2 km, 4 km, and 8 km resolution and a baseline). Beringer and Ball computed an F-test, $F(3, 28 = 2.109)$, but did not find significant effects of NEXRAD resolution on pilot distance to weather. In Bayes Factor terms, the t-value for this test provides anecdotal evidence for the null. Therefore, we conclude from these results that precipitation symbology alone does not introduce a credible variability in pilots' weather-distance decisions.

We found null support for weather zoom usage and for pilot/ATC communications. For radio communications, the number and durations are similar across pilot groups. For the zoom usage, all groups displayed the same use of the weather presentation zoom, and they displayed each zoom level for similar durations. Most of the pilots used the level 2 zoom with a 20 nmi distance between successive range rings. Much less time was spent using the 5 nmi (Zoom 1) and the 50 nmi (Zoom 3) presentations. Evidently, the 20 nmi presentation provided a balance between sufficient weather detail for avoidance and provided enough look-ahead distance for near-time planning. The feedback from pilots was positive, and pilots stated that the zoom functionality was very good for maintaining weather situational awareness. The consensus from pilots was that the smaller zoom levels were good for details and the largest zoom level provided an overview of the entire weather scenario. Pilots also suggested a user-defined zoom to make the presentation even more useful.

Additional results that support the null are pilot heading and altitude changes. Overall, the three groups kept the same average scenario heading and altitude. One exception is for Group 1 and Group 3 in Scenario 3. For altitude, this general result is somewhat expected because pilots did not have access to any echo top or echo bottom information and both scenarios started at an altitude. The average heading changes are correlated with the overall deviation behavior, which we will discuss in more detail below. For other weather scenarios and aircraft types, the heading, and altitude changes would likely play a more important role for assessing pilot weather-avoidance behavior.

Throughout our results, we found effects that reflect different behaviors between groups. Although most of these effects are anecdotal in support of the alternative, they are nevertheless operationally important due to the potential impact on GA pilots. These effects include weather deviations, fixations, EMW, and cognitive workload. In contrast to the anecdotal effects, for most of these outcomes we find credible effects (lower intragroup variability) that range from substantial to strong. From our pilot deviation measure, we found that pilot groups chose different deviation routes from start to destination. Evidently, based on the Group 1 and Group 3 weather presentations, Group 1 and Group 3 pilots chose a very different route from start to destination compared to routes that Group 2 pilots chose.

We found credible effects for a differential scanning pattern among the pilot groups. During flight, pilots sampled visual information from the out-the-window view, the cockpit glass display, the weather presentation, and the cockpit console. The movement of the eyes from one location to the next is called a saccade, and the prolonged foveal attention on an object or on detail is called a fixation. In combination, the array of saccades and fixations create a visual scanning pattern. The visual scan is an active process of exploring regions for goal-relevant information. If we perform a scan pattern analysis, it can reveal how pilots allocate their visual attention during flight. In this study, we used POG data to define saccades and fixations and to assess how pilots scanned the AOIs during flight. Overall, Group 3 had more fixations and saccades than Group 1 and Group 2. Specifically, we found Group 3 had credible effects with a different scan pattern, including an increased number of fixations and saccades on the cockpit glass display AOI and the weather presentation AOI. Judging from the weather presentations of Groups 1-3 (as shown in Figure 3), it is clear that the various combinations of colors and symbols create very different images. Although the weather data are the same, the perceptual impact from the combinations of elements, symbols, and colors is clearly different. We know from previous research on integrated cockpit displays (Arend, 2003) that simultaneous presentations of terrain maps and weather information can pose salience issues, especially in the area of color-coding. In addition, we know that different color combinations can affect the perceived urgency of weather data (Ahlstrom & Arend, 2005). These factors also influence user scan patterns and visual workload. Most important, variations in colors and weather symbology seem to affect pilot behavior and decision-making.

In addition to visual workload, we used fNIR, an optical imaging technology that measures neural activity and hemodynamic response in the prefrontal cortex, to assess the cognitive load as pilots navigated scenarios under both VFR and IFR conditions. For this study, we focused on analyzing the oxygenated hemoglobin. Typically, the fNIR signal from neural activation is a decrease of deoxygenated hemoglobin accompanied by an increase of oxygenated hemoglobin. Our findings show that oxygenation levels were very similar across groups, predominantly, showing anecdotal support for the null. One exception is a group difference in oxygenation levels during the VFR portion of flight, which provides anecdotal evidence for the alternative. This means that pilots were under similar cognitive loads during flight, regardless of the weather presentation (Groups 1-3) used. In contrast, we find results favoring the alternative with consistently higher oxygenation levels for the IFR portions of flight compared to the VFR portions of flight. During VFR, the pilot navigates by visual sampling from the out-the-window view. During IFR, when visibility is zero, the pilot navigates by means of cockpit instruments under ATC control. Clearly, these two flight rule modes impose very different planning and decision states on part of the pilot. However, future research needs to unveil what information seeking, planning, and decision-making cause the increased cognitive load as evidenced by the increased oxygenation. These are important factors that need to be addressed in future assessments of weather symbology presentations.

5. CONCLUSION

In the present study, we found credible effects on pilot behavior, workload, and decision-making from the use of weather presentations. The only manipulation during the simulation was the weather presentation symbology; the weather data were the same. This raises the question if other scenario and weather factors, not investigated in this study, also affect pilot behavior during GA weather avoidance. For example, we deliberately constrained our weather presentation to four different weather data elements. This meant that pilots could learn the weather presentation quickly, and it negated the need for exhaustive pilot training and practice. In addition, we took great care to limit clutter in the display, yet we provided sufficient symbology for weather avoidance flights to allow for a symbology comparison. Commercial MET products allow the user to select and display a great number of weather data elements. From a pilot's perspective, this is advantageous; however, it can also cause serious legibility and salience issues. Besides the three zoom levels, we also omitted all pilot interactions with the weather presentation. This was not done because weather presentation interfaces are not important; it was done because interface design is a confounding factor that tells us nothing about the effects of weather symbology. Another important factor is pilot weather informational needs and usage during different phases of flight (FAA, 2010). In the present simulation, pilots flew a Cessna 172 simulator. With GA appropriate weather scenarios and an average speed of 130 knots (with no take-offs or landings), the scenarios posed particular time stress, decision-making, and flight constraints on pilots. With a faster aircraft, a different flight scenario, and more weather encounters, there would be a different time-constraint for pilot decision-making. Therefore, it is very likely that behavioral and workload effects will differ and that some effects will be more pronounced.

6. RECOMMENDATIONS

This simulation was specifically tailored to answer the primary question "Are there effects on pilot decision-making and behavior from weather presentation symbology?" Because this study revealed effects on pilot behavior, we need to answer the next question "Are these effects of weather symbology operationally important?" This entails follow-up studies that specifically focus on key parameters to assess their operational impact on pilot behavior; therefore, we make the following recommendations for the WTIC Program Office.

1. Conduct follow-up studies to assess the operational importance of variations in pilot decision-making and behavior from the use of weather presentations. The present study shows that pilot weather-avoidance behavior differs as a function of weather presentation symbology. The follow-up study needs to asses if these differences in flying behavior are operationally important. That is, are certain weather symbology presentations causing pilots to make suboptimal decisions?
2. Conduct follow-up studies to assess how weather presentation symbology affects pilots with different levels of experience. The participants in this study were experienced GA pilots. Are the effects of weather presentation symbology the same for pilots with much less experience?
3. Conduct follow-up studies to assess pilot workload under operationally relevant flight scenarios. The present study found differences in workload (i.e., fNIR) between pilot

groups. However, the flight scenarios were specifically tailored to assess the presence of workload effects. Furthermore, the weather presentation consisted of four weather elements that were displayed at all times. There is a need to assess how other flight conditions, weather presentation designs, and time constraints affect pilot workload. For example, how is pilot workload affected by weather symbology when pilots are flying without the use of auto pilot? How is pilot workload affected when the user needs to select or deselect weather presentation elements from an interface?

4. Conduct follow-up studies to assess the operational impact of latencies and time stamps for weather presentations. In the present study, pilots did not have a time stamp on the weather presentation but were told that the precipitation information could be up to 10-17 minutes old. Depending on how the time stamps for weather information are generated, they could potentially be erroneous or misleading. What is the operational impact on pilot behavior from the presence of time stamps? Are pilots aware of and do they adjust their decision-making based on the age of weather presentation time stamps?
5. Conduct follow-up studies to evaluate the impact on pilot visual scanning from weather symbology presentations. This study shows that visual scan patterns among pilots differ as a function of weather presentation symbology. Are these differences due to suboptimal color/symbology variations that reduce the efficiency by which pilots retrieve information? Are these effects magnified with an increasing number of weather elements in the presentation? If so, what combination of symbols and colors reduces the scan pattern variability among pilots?

References

Ahlstrom, U., & Arend, L. (2005). Color usability on air traffic control displays. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (pp. 93-97). Santa Monica, CA: Human Factors and Ergonomics Society.

Ahlstrom, U., & Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, *36*(7), 623-636.

Applied Science Laboratories Inc. (1991). 4100H *Eye-tracking system with head mounted optics specifications* [Computer hardware]. Bedford, MA: ASL, Inc.

Arend, L. (2003). Graphics issues of an aviation integrated hazard displays. In *Proceedings of the 12th International Symposium on Aviation Psychology* (pp. 60-64). Moffett Field CA: International Symposium on Aviation Psychology.

Beringer, D., & Ball, J. (2004). *The effects of NEXRAD graphical data resolution and direct weather viewing on pilots' judgments of weather severity and their willingness to continue flight* (DOT/FAA/AM-04/5). Oklahoma City, OK: Civil Aerospace Medical Institute Federal Aviation Administration.

Chamberlain, J. P., & Latorella, K. A. (2002). *Convective weather detection by general aviation pilots with conventional and data-linked graphical weather information sources* (NASA-2002-01-1521). Hampton, VA: NASA.

Comerford, D. A. (2004). *Recommendations for a cockpit display that integrates weather information with traffic information* (NASA/TM-2004-212830).

Dienes, Z. (2011). Bayesian versus orthodox statistics: which side are you on? *Perspectives on Psychological Science*, *6*(3), 274-290.

Elgin, P. D., & Thomas, R. P. (2004). *An integrated decision-making model for categorizing weather products and decision aids.* Hampton, VA: NASA.

Federal Aviation Administration. (2010). *Weather technology in the cockpit program capabilities report* (DTFAWA-09-C-00088). Atmospheric Technology Services Company, LLC, Norman, OK.

Federal Aviation Administration. (2012). *FAR/AIM (Federal aviation regulations/aeronautical information manual).* New York: Author.

Federal Aviation Administration, & National Oceanic and Atmospheric Administration. (1983). *Thunderstorms* (DOT/FAA/AC00-24B). Oklahoma City, OK: FAA.

Federal Aviation Administration, & National Oceanic and Atmospheric Administration. (2010). *Aviation weather services* (AC 00-45). Oklahoma City, OK: FAA and NOAA.

Grasse, T., Schilke, C., & Schifele, J. (2008). Symbology evaluation for strategic weather information on the flight deck. *IEEE.*

Hegarty, M., Canham, M. S., & Fabrikant, S. I. (2010). Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 37-53.

Izzetoglu, K., Bunce, S. C., Shewokis, P. A., & Ayaz, H. (2010). *Conformance monitoring and controller workload part task* (DTFA01-00-C-00068). Philadelphia, PA: Drexel University Press.

Izzetoglu, M., Bunce, S. C., Izzetoglu, K., Onaral, B., & Pourrezaei, K. (2007, July/August). Functional brain imaging using near-infrared technology: Assessing cognitive activity in real-life situations. *IEEE Engineering in Medicine and Biology Magazine*, *26*(4), 38-46.

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14(7)*, 293-300. doi:101016/jtics201005001

McAdaragh, R. M. (2002). *Toward a concept of operations for aviation weather information implementation in the evolving national airspace system* (NASA/TM-2002-212141). Hampton, VA: NASA.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406-419.

Radio Technical Commission for Aeronautics, Inc. (2004). *Minimum aviation system performance standards (MASPS) for flight information service-broadcast (FIS-B) data link* (DO-267A). Washington, DC: Author.

Rouder, J., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225-237.

Stein, E. S. (1992). *Air traffic control visual scanning* (DOT/FAA/CT-TN92/16). Atlantic City International Airport, NJ: DOT/FAA Technical Center.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426-432.

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.

Yuchnovicz, D. E., Novacek, P. F., Burgess, M. A., Heck, M. L., & Stokes, A. F. (2001). *Use of data-linked weather information display and effects on pilot navigation decision making in a piloted simulation study* (NASA/CR-2001-211047). Hampton, VA: NASA.

Acronyms

AGL	Above Ground Level
AOI	Area of Interest
ATC	Air Traffic Control
ATIS	Automatic Terminal Information Service
BF	Bayes Factor
CAT	Clear Air Turbulence
DUAT	Direct User Access Terminal
EMW	Eye Movement Workload
FAA	Federal Aviation Administration
FAR	Federal Aviation Regulation
fNIR	Functional Near Infrared
FSS	Flight Service Station
GA	General Aviation
HUO	Huguenot VOR
IFR	Instrument Flight Rules
IMC	Instrument Meteorological Conditions
IRB	Institutional Review Board
JZS	Jeffrey-Zellner-Siow
KBDL	Bradley International Airport
LHY	Lake Henry VOR
LIFR	Low Instrument Flight Rules
MET	Meteorological
METAR	Meteorological Report
MIP	Milton VOR
MVFR	Marginal Visual Flight Rules
NAS	National Airspace System
NDI	Northern Digital Inc.
NEXRAD	Next Generation Radar
NextGen	Next Generation Air Transportation System
NHST	Null Hypothesis Significance Testing
NOAA	National Oceanic and Atmospheric Administration
PDM	Pilot Decision Making
POG	Point of Gaze
PTT	Push-To-Talk
RTCA	Radio Technical Commission for Aeronautics
SIGMET	Significant Meteorological Advisory
SME	Subject Matter Expert

VFR	Visual Flight Rules
VHF	Very High Frequency
VMC	Visual Meteorological Conditions
VOR	Omnidirectional Radio Range
WJHTC	William J. Hughes Technical Center
WTIC	Weather Technology in the Cockpit

Author Biography

Ulf Ahlstrom is an Engineering Research Psychologist with the Federal Aviation Administration (FAA) William J. Hughes Technical Center (WJHTC), Human Factors Branch. He received his Ph.D. in Psychology from Uppsala University, Sweden, in 1994. Ulf has been supporting the FAA since 1997 (first as a contractor, then as a federal employee) following a 2-year, post-doctoral research fellowship at the Vanderbilt University Vision Research Center. His current research interests are in broad areas of air traffic control, user-interface design, operator workload, and weather information displays.

Matt Dworsky is a Human Factors Specialist with TASC, Inc. at the Federal Aviation Administration (FAA) William J. Hughes Technical Center (WJHTC), Human Factors Branch. In 2011, he received his M.S. in Human Factors Psychology from George Mason University, Fairfax, VA. His research interests are in weather information displays, flight deck operations, electronic flight bag (EFB) research, and interface design.

Appendix: Weather Presentation Questionnaire

Weather Presentation Questionnaire

1. Using the weather presentation, how easy was it to see the convective SIGMET areas?

Very Hard					Very Easy
1	2	3	4	5	

2. Using the weather presentation, how easy was it to determine the location of severe precipitation areas?

Very Hard					Very Easy
1	2	3	4	5	

3. How much did the weather presentation assist you in determining what areas of weather to avoid?

Not at all					Very Much
1	2	3	4	5	

4. How would you rate the benefits of the weather presentation information to other sources of weather information (ATIS, Flight Watch, etc.)?

Very Poor					Very Good
1	2	3	4	5	

5. During the simulation, how frequently did you use the weather presentation to help you navigate through or around adverse weather conditions?

Very Rarely					Very Frequently
1	2	3	4	5	

6. How much do you think the weather presentation decreased your workload?

Not at all					Very Much
1	2	3	4	5	

7. How much did you trust the weather presentation to give you correct information?

Not at all					Very Much
1	2	3	4	5	

8. How easy was it to determine the position of the aircraft based on the presentation?

Very Hard					Very Easy
1	2	3	4	5	

Thank you very much for participating in our study, we appreciate your help.